



Claude Ménard
Mary M. Shirley
Editors

Handbook of New Institutional Economics

Second Edition

*Open access made possible through
a generous donation in honor of
the Ronald Coase Institute*

OPEN ACCESS



Springer

Handbook of New Institutional Economics

Claude Ménard • Mary M. Shirley
Editors

Handbook of New Institutional Economics

Second Edition

With 34 Figures and 31 Tables



Springer

Editors

Claude Ménard
Sorbonne Economics Centre
Pantheon-Sorbonne University
Paris, France

Mary M. Shirley 
The Ronald Coase Institute
Philomont, VA, USA



ISBN 978-3-031-50809-7

ISBN 978-3-031-50810-3 (eBook)

<https://doi.org/10.1007/978-3-031-50810-3>

This work was supported by The Ronald Coase Institute.

© The Editor(s) (if applicable) and The Author(s) 2005, 2025

This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG. The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

When the first edition of the *Handbook of New Institutional Economics* was published in 2005, NIE was a promising, emerging field. In the decades since the term “new institutional economics” was first coined by Oliver Williamson in 1975, new institutional economics had been growing rapidly, winning many new adherents. Nevertheless, in 2005 there were still many who ignored institutions in their research, especially among mainstream economists. Ronald Coase’s prediction in 1988 that eventually “all economics will have become what we now call ‘the new institutional economics’” was far from reality (Coase, 1998, pp. 73–74). As this second edition of the *Handbook* heads to publication, NIE’s success both inside and outside of economics is strong, thanks to its usefulness “for understanding real world phenomena through its trademark combination of theoretical concepts rooted in a concrete vision of reality with an emphasis on empirical testing. . .” (Ménard & Shirley, 2022, p. x).

In the 20 years since the first edition was published, there has been an outpouring of new research. This second edition has been extensively revised and expanded to reflect the many new discoveries, topics, and methods characterizing new institutional economics today. Eighteen of the 41 chapters are entirely new, and 23 have been thoroughly updated, with nine classic chapters unchanged from the first edition. For example, NIE’s increasing emphasis on social norms and culture—what Douglass North termed “informal institutions”—is reflected in this new edition as is its expanding diversity of methodologies, including those made possible with advances in computing capacity and machine learning.

In revising and updating the *Handbook* we have preserved its strengths, most importantly, its portrayal of NIE’s diversity. As we have stated elsewhere: “New institutional economics (NIE) encompasses a diverse set of ideas unified around a shared belief in the central role of institutions, transaction costs, property rights, and contracts. From this strong central trunk, NIE developed branches focusing on the fundamental institutional environment, the governance of transactions and organizations, and the role of communal organizations and collective action” (Ménard & Shirley, 2022, p. x). As with the first edition, the goal of the second edition of the *Handbook* is to provide interested readers with an essential overview of this diverse and dynamic field.

We are delighted that thanks to a generous donation in honor of the Ronald Coase Institute, which has been supporting institutional research for 25 years, the digital version of this new edition is open access, making it available for free to researchers worldwide.

Paris, France
Philomont, VA, USA
December 2024

Claude Ménard
Mary M. Shirley

References

- Coase, R. H. (1998). New institutional economics. *American Economic Review*, 88(2), 72–74.
- Ménard, C., & Shirley, M. M. (2022). *Advanced introduction to new institutional economics*. Northampton (US) and Cheltenham (UK): Edward Elgar Pub.

Acknowledgments

We are grateful for The Nobel Foundation 1991 for permission to publish the Prize Lecture “The Institutional Structure of Production” by Ronald H. Coase, with changes introduced by Professor Coase. We also thank the Nobel Foundation and the heirs and estates of deceased authors for permission to publish their chapters as open access.

With a few exceptions, each chapter in this book went through a very extensive process of review and updating. Besides a systematic review by the two editors, the following reviewers did a remarkable job and we are very grateful for their assistance:

Douglass W. Allen, Samuel Bazzi, Luigi Butera, Bruno Deffains, Liesbeth Dries, Thrainn Eggertson, Sebastian Galiani, Scott Gehlbach, Miriam Golden, Zhangkai Huang, Philip Keefer, Cheryl Long, Gaetano Martino, Victor Menaldo, Thomas Merrill, Anthony Niblett, John Nye, Dominic Parker, Jens Prüfer, Emmanuel Raynaud, Davis Skarbek, Henry Smith, Giancarlo Spagnolo, Enrico Spolaore.

Contents

1 Introduction: Handbook of New Institutional Economics	1
Claude Ménard and Mary M. Shirley	
Part I The Domain of New Institutional Economics	23
2 Institutions and the Performance of Economies Over Time	25
Douglass C. North	
3 The Institutional Structure of Production	37
Ronald H. Coase	
4 Transaction Cost Economics	47
Oliver E. Williamson	
5 Doing Institutional Analysis Digging Deeper than Markets and Hierarchies	73
Elinor Ostrom	
Part II Political Institutions	103
6 Electoral Institutions and Political Competition: Coordination, Persuasion, and Mobilization	105
Gary W. Cox	
7 Presidential, Parliamentary, and Hybrid Forms of Government	129
John M. Carey and Juan Pablo Micozzi	
8 The Performance and Stability of Federalism: An Institutional Perspective	159
Barry R. Weingast	
9 Local Politics in Nations and Empires	185
Roger B. Myerson	

10	Authoritarian Institutions	209
	Georgy Egorov and Konstantin Sonin	
11	The Emergence and Failure of Democracy: Explaining Paths of Political Regime Change	237
	Daniel Treisman	
Part III	Legal Institutions of a Market Economy	265
12	Revisiting the Many Legal Institutions that Support Contractual Commitments in a Globalized World	267
	Gillian K. Hadfield and Alexander Bernier	
13	Property Law	293
	Carol M. Rose	
14	Property Rights, the State, and Politics	319
	Lee J. Alston and Bernardo Mueller	
15	State Regulation of Open-Access, Common-Pool Resources	339
	Gary D. Libecap	
16	Legal Institutions and the Development of Financial Systems Revisited	359
	Thorsten Beck and Ross Levine	
17	Licit and Illicit Responses to Regulation: Revisited	391
	Lee Benham	
Part IV	Organizations and Governance	415
18	Vertical Integration	417
	Paul L. Joskow	
19	The Make-or-Buy Decision Revisited	447
	Peter G. Klein and João Fernando Rossi Mazzoni	
20	Hybrids and Other Challenges in Organizational Economics	475
	Claude Ménard	
21	The Persistent Institutions of Corporate Governance	505
	Mark J. Roe	
22	Firms and the Creation of New Markets	535
	Erin Anderson and Hubert Gatignon	
23	Governance and Regulation of Platforms	565
	Martin Peitz	

Part V Contracts	595
24 The Empirics of Relational Contracts: The Good, the Bad, and the Ugly	597
Rocco Macchiavello	
25 Governing Relationships: Theory, Evidence, and Opportunities	621
Ricard Gil and Giorgio Zanarone	
26 Contracting with Public Authorities: Recent Developments	651
Jean Beuve and Stéphane Saussier	
27 Agricultural Contracts and Transaction Costs	673
Douglas W. Allen and Dean Lueck	
Part VI Institutional Change	699
28 The Economics of Institutional Change: Achievements and Challenges	701
Jean-Philippe Platteau	
29 Institutions and Development: Persistent Puzzles	729
Mary M. Shirley	
30 Institutional and Non-institutional Explanations of Economic Differences	757
Stanley L. Engerman and Kenneth L. Sokoloff	
31 Institutional Change	785
Desiree Desierto and Mark Koyama	
32 Commitment, Coercion, and Markets: The Nature and Dynamics of Institutions Supporting Exchange	811
Avner Greif	
Part VII Culture and Informal Institutions	869
33 Institutions Before States: Aggressive Egalitarianism and Private Property	871
Samuel Bowles and Jung-Kyoo Choi	
34 Culture Versus Institutions in the Great Enrichment	897
Joel Mokyr	
35 Formal and Informal Institutions Explaining China's Modern Economic Development: Their Historical Roots and Recent Evolution	927
Mary M. Shirley and Lixin Colin Xu	

36 Religion and New Institutional Economics	953
Jared Rubin	
37 Trust and the New Institutional Economics	975
Philip Keefer	
38 Corruption, Norms, and the Law	1001
Alberto Simpser	
Part VIII Methodology	1027
39 Combining Machine Learning and Econometrics to Examine the Historical Roots of Institutions and Cultures	1029
Peter Grajzl and Peter Murrell	
40 New Developments in Case Study Methods for Institutional Economics	1059
Andrew Bennett	
41 Analyzing the Deep Determinants of Institutions: Methodological Choices and Challenges	1089
Robbert Maseland	
42 Using Field Experiments to Understand the Impact of Institutions on Economic Growth	1117
Omar Al-Ubaydli, Faith Fatchen, and John List	
Index	1145

About the Editors



Claude Ménard is Professor of Economics (Emeritus) at the Centre d'Economie de la Sorbonne (University of Paris Pantheon-Sorbonne). He has published extensively in peer-reviewed journals and authored several books, most recently *A Research Agenda for New Institutional Economics* (with Mary Shirley; Edward Elgar 2018), *Network Infrastructures: Technology Meets Institutions* (Oxford University Press, 2021; with Rolf Kunneke and John Groenewegen), and *Advanced Introduction to New Institutional Economics* (Edward Elgar, 2022; with Mary Shirley). He is a founder and past President of the International Society for New Institutional Economics (now Society for Institutional & Organizational Economics) and a founder and member of the board of the Ronald Coase Institute.



Mary M. Shirley is a founder and President of the Ronald Coase Institute. Prior to that she worked for over 20 years in World Bank research. She has published numerous scholarly articles and books, including *Advanced Introduction to New Institutional Economics* (with Claude Ménard, 2022), *A Research Agenda for New Institutional Economics* (with Claude Ménard, 2018), and *Institutions and Development* (2009). She is a founder and past President of the International Society for New Institutional Economics (now Society for Institutional and Organizational Economics) and co-editor (with Mike Sykuta) of *New Institutional Economics*, a Social Science Research Network e-journal.

Contributors

Douglas W. Allen Department of Economics, Simon Fraser University, Burnaby, BC, Canada

Lee J. Alston Department of Economics, Indiana University, Bloomington, IN, USA

National Bureau of Economic Research, Cambridge, MA, USA

Omar Al-Ubaydli Bahrain Center for Strategic, International and Energy Studies, Studies and Research Department, Riffa, Bahrain

George Mason University, Fairfax, VA, USA

Erin Anderson INSEAD, Fontainebleau, France

Thorsten Beck European University Institute, Fiesole, Italy

Lee Benham Ronald Coase Institute and Washington University, St. Louis, MO, USA

Andrew Bennett Georgetown University, Washington, DC, USA

Alexander Bernier University of Toronto, Toronto, ON, Canada

Jean Beuve Lille Economie Management, University of Lille, Lille, France

Samuel Bowles Santa Fe Institute and CORE Econ, Santa Fe, NM, USA

John M. Carey Department of Government, Dartmouth College, Hanover, NH, USA

Jung-Kyoo Choi Kyungpook National University, Daegu, South Korea

Ronald H. Coase University of Chicago Law School, Chicago, IL, USA

Gary W. Cox Department of Political Science, Stanford University, Stanford, CA, USA

Desiree Desierto Department of Economics, George Mason University, Fairfax, VA, USA

Georgy Egorov Kellogg School of Management, Northwestern University, Evanston, IL, USA

Stanley L. Engerman Department of Economics, University of Rochester, Rochester, NY, USA

Faith Fatchen University of Chicago, Chicago, IL, USA

Hubert Gatignon INSEAD, Fontainebleau, France

Ricard Gil IESE Business School, Barcelona, Spain

Peter Grajzl Department of Economics, The Williams School of Commerce, Economics, and Politics, Washington and Lee University, Lexington, VA, USA
CESifo, Munich, Germany

Avner Greif Stanford University, Stanford, CA, USA

Gillian K. Hadfield University of Toronto, Toronto, ON, Canada

Paul L. Joskow Department of Economics, Massachusetts Institute of Technology, Cambridge, MA, USA

Philip Keefer Inter-American Development Bank, Washington, DC, USA

Peter G. Klein Department of Entrepreneurship and Corporate Innovation, Baylor University, Waco, TX, USA

Mark Koyama Department of Economics, CEPR, and Mercatus Center, George Mason University, Fairfax, VA, USA

Ross Levine Stanford University, Stanford, CA, USA

Gary D. Libecap National Bureau of Economic Research, University of California, Santa Barbara, CA, USA

John List University of Chicago, Chicago, IL, USA

Dean Lueck Department of Economics, Indiana University, Bloomington, IN, USA

Rocco Macchiavello London School of Economics and Political Science, London, UK

Robbert Maseland Radboud University, Nijmegen, Netherlands

João Fernando Rossi Mazzoni Department of Entrepreneurship and Corporate Innovation, Baylor University, Waco, TX, USA

Claude Ménard Centre d'Economie de la Sorbonne, Pantheon-Sorbonne University, Paris, France

Juan Pablo Micozzi Instituto Tecnológico Autónomo de México (ITAM), Mexico City, Mexico

Joel Mokyr Departments of Economics and History, Northwestern University, Evanston, IL, USA

Berglas School of Economics, Tel Aviv University, Tel Aviv, Israel

Bernardo Mueller University of Brasilia, Brasilia, Brazil

Peter Murrell Department of Economics, University of Maryland, College Park, MD, USA

Roger B. Myerson University of Chicago, Chicago, IL, USA

Douglass C. North Washington University in St. Louis, St. Louis, MO, USA

Elinor Ostrom Indiana University, Bloomington, IN, USA

Martin Peitz Department of Economics, University of Mannheim and MaCCI, Mannheim, Germany

Jean-Philippe Platteau Department of Economics and Centre for Research in the Economics of Development (CRED), University of Namur, Namur, Belgium

Mark J. Roe School of Law, Harvard University, Cambridge, MA, USA

Carol M. Rose Yale Law School, New Haven, CT, USA

University of Arizona College of Law, Tucson, AZ, USA

Jared Rubin Chapman University, Orange, CA, USA

Stéphane Saussier Sorbonne Business School, Institut d'administration des entreprises, University Paris I Panthéon-Sorbonne, Paris, France

Mary M. Shirley The Ronald Coase Institute, Philomont, VA, USA

Alberto Simpser ITAM, Mexico City, Mexico

Kenneth L. Sokoloff University of California, Los Angeles, CA, USA

Konstantin Sonin Harris School of Public Policy, University of Chicago, Chicago, IL, USA

Daniel Treisman Department of Political Science, University of California, Los Angeles, CA, USA

Barry R. Weingast Stanford University, Stanford, CA, USA

Oliver E. Williamson University of California, Berkeley, CA, USA

Lixin Colin Xu Cheung Kong Graduate School of Business, Beijing, China

Giorgio Zanarone Hautes Etudes Commerciales, University of Lausanne, Lausanne, Switzerland

Introduction: Handbook of New Institutional Economics

1

Claude Ménard and Mary M. Shirley 

Contents

1.1	What Is New Institutional Economics?	2
1.2	Why a Second Edition of the Handbook on New Institutional Economics?	3
1.3	The Domain of NIE	4
1.4	Political Institutions	6
1.5	Legal Institutions of a Market Economy	8
1.6	Organizations and Governance	10
1.7	Contracts	13
1.8	Institutional Change	15
1.9	Culture and Informal Institutions	16
1.10	Methodologies	18
	References	20

Abstract

New institutional economics (NIE) studies the rules and norms that humans devise to reduce uncertainty and control their environment, both formal institutions, such as written contracts and laws, and informal institutions, such as societal norms. NIE is an exceptionally dynamic field and, as this introductory chapter attests, this second edition of the Handbook covers a wealth of new research and novel methodologies. It includes a new section on culture and informal institutions and new chapters on local politics, authoritarian institutions, platforms, relational contracts, and more. Many chapters from the first edition incorporate new discoveries in this rapidly changing field. This edition, like the first, covers both branches of NIE. As North defines it, “Institutions are the rule of

C. Ménard (✉)

Centre d'Economie de la Sorbonne, Pantheon-Sorbonne University, Paris, France
e-mail: claudemenard@univ-paris1.fr

M. M. Shirley

The Ronald Coase Institute, Philomont, VA, USA
e-mail: mshirley@coase.org

© The Author(s) 2025

C. Ménard, M. M. Shirley (eds.), *Handbook of New Institutional Economics*,
https://doi.org/10.1007/978-3-031-50810-3_1

the game in a society,” while organizations are the players, a distinction which delineates the two branches. The “Northeastern” branch focuses on the fundamental ground rules of the institutional environment and the “Williamsonian” branch on organizations and their governing institutions.

Keywords

New institutional economics · Organizations · Transaction costs · Property rights · Contracts · Governance

1.1 What Is New Institutional Economics?

New institutional economics (NIE) studies the rules and norms that humans devise to reduce uncertainty and control their environment. These include formal institutions, such as written agreements that govern contractual relations and corporate governance, constitutions, laws and regulations that govern politics, government, finance, and society more broadly, as well as informal institutions, such as unwritten codes of conduct, norms of behavior and conventions. As North defines it, “Institutions are the rule of the game in a society,” which define the “framework within which human interaction takes place” (1990, pp. 3–4). If institutions are the rules of the game, organizations are the players. Organizations are “specific groups of individuals pursuing a mix of common and individual goals” (North et al., 2009, p. 15). The distinction between institutions and organizations “delineates the two main branches of NIE: one focuses on the fundamental ground rules of the institutional environment . . . the ‘Northeastern branch’ . . . and the other on organizations and their governing institutions, or the ‘Williamsonian’ branch” (Ménard & Shirley, 2022, pp. 2–3). This Handbook covers both branches: after four classic chapters on the domain of NIE in Sections I, II and III deal with the “macro” or Northeastern side of NIE, and Sections IV and V deal with the “micro” institutions common to the Williamsonian branch. The remaining sections cover transversal topics.

NIE is defined not only by its content, but by its assumptions. New institutional economics abandons the standard neoclassical assumptions that individuals have perfect information and unbounded rationality and that transactions are costless. NIE assumes instead that individuals have incomplete information and limited mental capacity and because of this they face uncertainty about unforeseen events and outcomes and incur transaction costs to acquire and process information. To reduce risk and transaction costs humans create institutions and develop modes of organization. For new institutionalists, the performance of a market economy depends upon the formal and informal institutions and modes of organization that facilitate private transactions and cooperative behavior. NIE focuses on how such institutions emerge, operate, and evolve, how they shape the different arrangements that support production and exchange, as well as how these arrangements act in turn to change the rules of the game.

Because NIE considers choices to be embedded in institutions, it has a much broader reach than neoclassical economics, which has been largely concerned with prices and outcomes. But unlike old institutional economics, NIE does not abandon neoclassical economic theory. While new institutionalists reject the neoclassical assumption of perfect information and instrumental rationality, they accept orthodox assumptions of scarcity and competition. Both Arrow and Williamson attributed the rising influence of NIE to its acceptance of the successful core of neoclassical economics. As Kenneth Arrow observed, unlike the older institutionalist school, New Institutional Economics does “not consist of giving new answers to the traditional questions of economics—resource allocation and the degree of utilization. Rather it consists of answering new questions, why economic institutions emerged the way they did and not otherwise . . .” (1987, p. 734).

As Arrow suggested, NIE tries to answer questions that neoclassical economics does not address, and this has given NIE a distinct identity and a strong following. North pointed out (North, 2005) that neoclassical economics was not created to explain the process of economic change, much less political or social change. Institutionalists in contrast aim to understand change by understanding human incentives and intentions and the beliefs, norms, and rules that they create in pursuit of their goals (see North, 2005).

Answering new questions requires institutionalists to devise new methodologies. Elinor Ostrom’s chapter (4) argues that unlike much of social science, institutional analysis cannot simply hold constant other institutions because “the impact on incentives and behavior of one type of rule is not independent of other rules.” There are numerous examples of these interaction effects throughout this Handbook. For instance, Section II on political institutions illustrates how electoral procedures, political party norms, and constitutional laws and structures interact with one another to shape the incentives of politicians and voters and, ultimately, to influence policy decisions and organizational choices.

NIE’s breadth and innovation have fostered a multidisciplinary approach. Institutional analysts adapt useful concepts and methodology from political science, sociology, law, anthropology, cognitive science, evolutionary biology, and any other discipline that sheds light on the rules, norms, and beliefs that govern human interactions in the process of production and exchange. A number of the authors in this Handbook are not economists, but all are social scientists who share an interest in the scientific analysis of institutions.

1.2 Why a Second Edition of the Handbook on New Institutional Economics?

New institutional economics has continued to expand and diversify rapidly since the first edition was published in 2005. Indeed, since the term new institutional economics was first coined by Oliver Williamson in 1975, its influence over scholarly research has mushroomed. This influence is not limited to economists. Increasing numbers of legal scholars, political scientists, sociologists, anthropologists,

management specialists, and others are doing research in new institutional economics. NIE is also attracting new researchers, from many different countries. The exploding diversity and widening scope of NIE since 2005 cried out for a new edition to update existing information and add new subjects.

As with the first edition, this new edition is a synthetic book that provides interested readers with an overview of recent developments and broad orientations of new institutional economics. Much institutional research is published in journals that may not be familiar to others working in the field, and new institutionalists may be unaware of discoveries from disciplines other than their own. This updated edition of the Handbook acquaints readers with the scope of NIE, the recent trends, and the progress made by scholars from other fields. Also, young researchers may want guidance about what the topic means and how it is being researched. This volume, written by some of the foremost NIE specialists, gives new researchers an introduction to the topic and a reference book for their research.¹

This new edition of the Handbook has been extensively updated and revised. Eighteen of the 41 chapters are entirely new to this edition, including four chapters in a new section on methodology, and another 14 new additions on local government, religion, culture, China, relational contracts, platforms, and more. Of the 23 chapters that appeared in the earlier edition, all but nine have been substantially updated and revised. The nine unchanged chapters are classics whose relevance continues unabated.

The book opens with four of these classic chapters by four Nobel laureates that give the reader a sense of the scope of new institutional economics and the issues fundamental to the study of economic institutions (Section I). As we mentioned, the next sections deal with the macro institutions that shape the functioning of markets, firms, and other modes of organization, specifically, political institutions (Section II) and the legal system (Section III). This is followed by sections on the micro institutions that govern firms (Section IV) and their contractual relations (Section V). Increasingly institutional economics has focused on how institutions, both macro and micro, change: how they emerge, evolve and die (Section VI) and on culture and informal institutions (Section VII). Because NIE is addressing new questions or new aspects of old questions, its future is also being shaped by its choice of methodology, the subject of Section VIII.

1.3 The Domain of NIE

Douglass North's chapter argues that one of NIE's main inputs to economics has been to remove the fiction of the frictionless market by adding institutions but that NIE has the potential to perform an equally, or more powerful, service: changing neoclassical economics from a static to a dynamic theory. To understand economic change and how to improve economic performance, it is not enough to understand

¹Useful background on NIE can be found in Furubotn & Richter, 1998.

the basic rules of the game or even customs, norms, and habits, North maintains. We must also understand what people believe and how they arrive at those beliefs—how people learn. North was a leader in the movement to study the broader institutional framework that shapes the functioning of markets, to add beliefs and norms to the study of institutions, and to incorporate aspects of cognition in order to understand institutional change, as reflected in this classic chapter.

Ronald Coase's chapter on the institutional structure of production hearkens back to his seminal work that lies at the heart of New Institutional Economics: "The Nature of the Firm" (1937) and "The Theory of Social Cost" (1960) (reprinted in Coase, 1988). Coase's chapter describes his unwillingness to treat the firm as a "black box" that takes prices at one end and produces outputs at the other. This unwillingness led him famously to ask why firms exist, why are not all transactions done through the market? He famously answered that firms exist to economize on transaction costs. We find firms when it is cheaper to organize activities under a governing hierarchy than to try to conduct them in the market place and pay the cost of using the price mechanism to coordinate. Coase's argument that the level of transaction costs depends upon the institutional setting within which economic actors operate set the stage for the NIE. Coase's emphasis on empirical analysis of real economic phenomena using practical, even if sometimes inelegant, methodologies has also been influential in the variety of themes, approaches, and disciplines that characterize the NIE.

Transaction cost economics is a direct descendent of Coase's "Big Idea," as Oliver Williamson (following Varian) terms the theory of the firm in his chapter. Transaction cost economics is well named; it is concerned with transactions, specifically: (i) the extent to which the assets involved are specific to a transaction, (ii) how disturbances or changes may affect the transaction, and (iii) how frequently the transaction will reoccur. The nature of transactions affects contracts and the way in which economic activities are allocated between firms, markets, or other modes of organization. These in turn affect whether incentives are high- or low-powered, whether administration is hands on or off; and whether dispute resolution relies on courts or private ordering. As a consequence, the relative advantages of a specific arrangement can only be assessed comparatively, taking into account the characteristics of the transactions at stake and the institutional environment within which they are conducted.

Elinor Ostrom's chapter presents a challenge to new institutionalists as daunting as North's. Ostrom calls for institutionalists in all social sciences to seek out universal components for markets and hierarchies and develop theories of human behavior in diverse settings. Ostrom draws on the foundations of many disciplines to devise a framework (IAD, Institutional Analysis and Development) that can be used in analyzing any type of institutional arrangement, which she and others have applied to a variety of different arenas. The theoretical and empirical tasks she sets are difficult and complex, but her own large body of research and much subsequent contributions show that they are feasible if social scientists are ready to rise above the specialized language, knowledge, and assumptions of their sub-disciplines.

1.4 Political Institutions

The level of transaction costs depends upon the institutional setting according to Ronald Coase, and political institutions are among the foremost factors that shape that setting. As North has pointed out, political institutions can play an important role in reducing transaction costs by improving the security of property rights and enforcement of contracts (North, 1997, p. 150). But states do not necessarily play this role; indeed, many are ineffectual at protecting rights or securing contracts and many others are themselves an important threat to the security of property rights and a prime violator of contracts. Understanding how politics affect the transactional environment, how economic and political markets interact, and how, when, and why states enforce or violate property rights and contracts are important tasks for NIE research.

As Barry Weingast makes clear in his chapter on federalism, NIE approaches these themes by dropping the traditional economic assumption of government as benevolent and the opposite assumption of government as Leviathan, focusing instead on how different institutional arrangements affect the incentives and performance of government. In particular NIE considers a fundamental dilemma: investment increases when property rights are protected, but a state strong enough to protect property rights is also strong enough to expropriate them (Weingast, 1993). One way to limit government is by putting institutional constraints on the executive, such as competitive elections and to separate state powers into branches (executive, legislative, judicial) or hierarchies (federal, provincial, local) and require them to compete or balance each other.

As the chapter on electoral institutions by Gary Cox makes clear, the variance in electoral rules is large and the effects of different designs significant. Yet electoral rules are not often analyzed as determinants of outcomes and few papers compare the effects of different electoral rules. This is a major gap. Within democracies, electoral systems differ in how they allocate the number of votes per voter, the number of seats per district, and the proportionality of votes to seats. These three factors—number of votes, district magnitude, and proportionality—affect how electoral competitors try to coordinate, persuade, and mobilize voters, and this in turn has implications for who gets elected, the types of promises politicians make to voters, and both the extent of turnout and which groups are likely to vote. Policy choices are strongly influenced by which groups of constituents a policy maker must account to during elections, and this ultimately has a profound effect on economic performance.

The choice between a presidential or parliamentary government is also highly consequential for any democracy. This seems like a stark choice between two polar options, but it is not. As John Carey and Juan Pablo Micozzi's chapter shows, there are many hybrid arrangements, especially in newer democracies. New democracies have overwhelmingly adopted systems that directly elect the chief executive, but many have also adopted mechanisms to try to ensure that the president maintains support from the legislative body, by adding an office of prime minister or requiring the president to survive confidence votes, for example. Even more striking are actions by legislatures to replace presidents. These hybrids between presidential

and parliamentary systems are attempts to preserve a separation of powers while reducing the risk that a president loses all cooperation from the legislature, which can threaten the very survival of a fragile democratic system. The stability and policies of presidential and hybrid systems are affected by electoral rules, such as whether the president and legislature are elected at the same time, and by the rules governing the power of the president to set the legislative agenda, issue decrees, veto laws, and be reelected.

The design of federal systems, as Barry Weingast shows in his chapter, is another complex source of variance. Federalism varies in the number and character of layers in the hierarchy of a country's governments, the types of power delegated to its sub-national governments, the extent and regulation of the common market shared by its different sub-national governments, and the institutions that protect the federal arrangement from encroachment. These choices have profound implications for performance since they affect whether governments serve private or public ends. For example, Weingast shows how federal revenue sharing rules can affect levels of corruption. When local governments raise funds locally, they have a stronger incentive to focus on market enhancing public goods because it increases their tax base and allows local governments to provide more of these goods. To the extent that revenues are raised nationally and distributed to the local governments according to national political criteria, this incentive is reduced and local governments will focus more on private rents. The design of the federal rules also determines whether federalism itself can survive. The center of a federal state needs sufficient power to police common pool problems among the sub-national governments, yet the more powerful the center, the more likely it is to abuse its power. In addressing this puzzle, Weingast's chapter exemplifies how an institutional focus poses different questions and seeks new answers. Traditionally economists asked what powers should be assigned to what level of government and answered variously that power should be assigned to the level with the most information, to maximize competition, to produce public goods most efficiently, etc. NIE asks instead how different federalist designs affect the incentives and objectives of government officials to further citizens' welfare and whether federalism is self-enforcing.

Roger Myerson's chapter continues the discussion of local government, describing how local government has roots in communities formed during the earliest evolution of human societies when communal endeavors required cooperation and coordination, including agreements about rights to crops and land with the advent of agriculture. The formation of wider states promising to protect more extensive territories against raiding by outsiders led to the rise of "mandarins," agents who expected to be rewarded for serving the state above any local connections. At the same time, Myerson argues, the stability and prosperity of the wider state hinged on its ability to win the trust of the local elites or "gentry." Local and national leadership can be aligned in two fundamentally different ways: by top-down influence of national leaders in local politics or by bottom-up influence of local leaders in national politics. Top-down influence is dominant in successful autocratic states. In successful democracies, national leaders need bottom-up approval from local groups throughout the nation, and local leaders who perform well can become

competitive candidates for national leadership. Myerson examines the different ways autocracies and modern democracies have established connections between local and national politics and explores the policy implications for international support for democratic development.

The next two chapters deal with non-democracies. Autocratic regimes were once neglected by NIE but are now the subject of a large and growing literature as clarified in the chapter by Georgy Egorov and Konstantin Sonin. Egorov and Sonin explore some of the reasons why autocracy emerges and spreads and why it seems to be resurgent today. In their view the political institutions of authoritarianism are designed to ensure the elite's lasting hold of power. They might be growth-enhancing at certain stages of a country's development and become a major obstacle to economic progress in other stages, but their primary role is to keep the elite unaccountable. Egorov and Sonin also analyze some of the tools dictators employ to maintain power, such as institutions of disenfranchisement and prevention of collective action with a primary role played by information control, and the challenges of building elite coalitions and balancing power.

The transition from democracy to autocracy and vice versa is another concern for institutionalists, made even more salient with recent apparent trends toward dictatorship. Daniel Treisman's chapter examines the variety of explanations for political regime change, some emphasizing macro processes of economics and institutional change and others focusing on micro games of political actors. He finds limitation in both theories and explores new research suggesting ways in which they might be usefully combined. For example, micro interventions might trigger change while its direction is set by structural factors. The chapter also reviews arguments about international influences and how they may combine with domestic processes, as well as recent work disaggregating democracy into constituent elements and examining the historical processes by which these emerge. A final section addresses literature on democratic breakdowns and "backsliding" to authoritarian governance, where Treisman once again finds a combination of macro and micro explanations more convincing.

1.5 Legal Institutions of a Market Economy

An important regularity in NIE, which goes back to Coase (1960), is the critical role played by rule of law in the development and health of a market economy. Several of the chapters in this section focus on how legal institutions support market economies by enforcing contractual obligations and protecting private property from state predation.

Contractual agreements can be enforced in many ways, as the chapter by Gillian Hadfield and Alexander Bernier describes, but many forms of enforcement are not credible to all parties or have high costs. When enforcement is not credible or too costly, many otherwise lucrative transactions will not occur and economic performance will suffer. Even though very few disputes are actually pursued in court, an effective and efficient legal system alters people's incentives to behave

opportunistically, improves the efficacy of other forms of contract enforcement, and increases the number of profitable transactions. Hadfield and Bernier show that the term “legal system” covers a complex, interwoven structure of laws, doctrines, norms, organizations, professions, and individuals facing major incentive and coordination problems, nicely illustrating Ostrom’s argument that interactions between institutions cannot be safely ignored. In identifying clearly the different components of a legal system, Hadfield and Bernier also explore the interdependencies between domestic and transnational systems suggesting that in a globalized world, well-functioning domestic legal systems may no longer be sufficient to assure low cost access to contract law.

From the outset, an important focus of the NIE research agenda was analysis of the issues surrounding the delineation, allocation, and implementation of property rights, a fundamental side of new institutional economics’ golden triangle and the subject of the next three chapters.

In characterizing property as a “social institution,” Carol Rose’s chapter on property law confronts head on the interaction between formal property law and the pervasive informal norms that compel people—sometimes—to respect other’s claims. Rose argues that property law goes beyond informal property rules because it aims to make rights universal and transactions transparent, reducing the scope for dispute and the costs of transfer. Rose also finds that property law has created property in realms unknown to informal property, such as intellectual property rights and tradeable environmental rights. She also suggests that under some circumstances both property law and informal property can contribute to democratic recognition of rights more broadly.

Lee Alston and Bernardo Mueller examine the complexity of the state’s role in upholding private property rights. Early theories argued that scarcity in resources would make secure property more valuable and create demand for the state to protect property rights. But many states do not supply secure property rights nor do they change property rights when changes in scarcity value demand it. There are political and economic transaction costs associated with the state establishing or changing property rights that are very similar to those described in other chapters of this Handbook for other actions. Central to their discussion is the influence of the institutional context and the power dynamics among stakeholders, which critically shape the provision, enforcement, and transformation of property rights. Through a discussion of four different approaches from the New Institutional Economics literature, they explore the political economy of the emergence, enforcement, and evolution of property rights and their role in economic development. Alston and Mueller’s chapter shows how NIE illuminates the causes and consequences of insecure and inefficient property rights.

Gary Libecap’s chapter examines circumstances where private property rights are less or even in-feasible—open access, common pool resources. When there are few constraints on access and use of a valuable resource, agents have an incentive to rush to capture immediate rents and to over-exploit. The resource cannot be distributed to more productive agents because there are no property rights underlying exchange to do so. Informal common-property solutions such as those proposed by Ostrom are

feasible where resource values are low, group size is small, and users are homogeneous in cost and objectives because transaction costs are minimal. Where these conditions are not met, state intervention can formally recognize and enforce group restrictions; recognize private property rights and exchange; or impose regulatory restrictions on access, inputs, and/or output. Understanding why and how governments intervene and assessing the economic, welfare, and resource outcomes requires attention to bargaining among the affected parties, the collective action costs involved, and the roles and incentives of politicians and agency officials, as Libecap illustrates in three case studies.

The chapter by Thorsten Beck and Ross Levine reminds us of another central role of legal institutions: supporting financial development. The authors review the burgeoning literature on the role legal institutions play in shaping financial development. They explore the considerable research that establishes that where legal systems enforce private property rights, support private contractual arrangements, and protect investors' legal rights, savers are more willing to finance firms, and financial markets more efficiently allocate capital. They also explore the literature on how Europe's different legal traditions were spread internationally through conquest, colonization, and imitation and help explain cross-economy differences in investor protection, the contracting environment, and financial development.

Finally, a hallmark of NIE is its concern with how and why *ex post* behavior differs from *ex ante* rules, assumptions, or agreements. Lee Benham's chapter illustrates how regulation can have consequences not anticipated in standard neo-classical models. Regulation can lead to such licit responses as substituting unregulated goods for regulated ones, or barter for regulated money exchange, or altering the organization of the market in response to regulations that raise or lower transaction costs. Regulation also stimulates a number of illicit responses, such as extralegal activity (the informal or underground economy) or corruption. Benham shows how regulation is path dependent and its long-run consequences depend heavily on the context and time period, leading to outcomes in which the effects on allocation are only a small part of regulation's total impact. His conclusions substantiate a central theme of NIE: the effects of institutions need to be assessed comparatively.

1.6 Organizations and Governance

Ronald Coase's query as to why some transactions are done in markets and others in firms initiated a rich stream of research. An important achievement of NIE in the years since Coase asked that question is to show how complex both question and answer really are. Following Coase's contribution and its transformation into a full research agenda by Williamson (see ► [Chap. 3](#)), a huge number of publications focused on what became known as the "make-or-buy" paradigm: why is it that a firm decides to make a good or service internally rather than supplying it through markets? This section comes back to this issue and to the many empirical contributions that lead to an extension and deepening of the initial model.

NIE has proved exceptionally powerful in exploring the “make-or-buy” question and particularly in explaining vertical integration. Paul Joskow’s now classic chapter contrasts traditional “neoclassical” theories that explained vertical integration by different forms of market imperfections with the transaction cost perspective that emphasizes the conditions under which integration prevails when it allows overcoming contractual incompleteness and securing relationship-specific investment. Although the chapter focuses on vertical integration Joskow makes it clear upfront that markets and hierarchies are only polar solutions among a wide variety of solutions to the organization of transactions. Comparing transaction cost theory to explanations provided by industrial organization as well as by the so-called property rights theory, which both tend to consider switching among alternative organizational solutions as costless, he points out that one clear advantage of transaction cost explanations is that it has produced testable hypotheses and spawned a wealth of empirical studies based on the fundamental assumption that there are benefits and costs associated to all solutions. Joskow then discusses methodological steps in the empirical implementation of this research program and some of the main challenges it faces, particularly when it comes to measurement.

Since Joskow’s chapter from 2005 was published, significant progress has been made in the TCE approach to organizational issues, including the “make-or-buy” trade-off, as substantiated by Peter Klein and João Fernando Rossi Mazzoni. Coming back to the problem of firm boundaries and the empirical findings inspired by the Coasian approach, these authors proceed to a detailed discussion of methodological issues already raised by Joskow, with an entire section on “strategies for empirical research.” Beyond TCE, which remains central for the analysis of make-or-buy decisions, Klein and Mazzoni point out complementary lessons that can be drawn from the resource-based view and its emphasis on internal resources and capabilities and from the “new” property rights approach that pointed out the problem raised by ownership arrangements in the decision whether to invest. Their review of recent empirical contributions also shows the progress made on the analysis of formal contracts as tools to coordinate intra-firm as well as inter-firm transactions and the increased attention paid to the role of informal agreements. They also take note of the new challenges posed by technological transformation with respect to firm boundaries and regulatory activities, as illustrated by the case of platforms, an issue which Martin Peitz explores in depth in another chapter. They conclude by noting that our understanding of the interaction between the determinants of organizational choices and the institutional environment remains very limited, opening room for future research.

In his chapter, Claude Ménard pushes further the analysis of a long-neglected category of organizations, the “non-standard” arrangements that Williamson identified as “hybrids.” Coming back to the fundamental contribution of NIE and its core concepts, Ménard suggests that the heuristic model initially proposed by Williamson (1985) can be usefully complemented by considerations of the allocation of rights. Integrating this dimension more explicitly improves our understanding of the tradeoffs among the variety of transactional solutions. New contributions over recent decades have enriched our understanding of these tradeoffs and also improved

identification of the complex motivations that push partners to go hybrid. In exploring the conditions under which hybrids facing unexpected challenges manage to adapt and remain sustainable in the long run, the literature has also paid increasing attention to the ex-post modalities of governance. Nevertheless, more research is needed on how contracts have been complemented by specific administrative regimes in which managerial provisions and practices play a key coordinating role. Additional analysis of the functions played by managers in organizational arrangements in which parties share some assets and related rights while competing on others also throws light on the conditions under which hybrids might prevail over integration as well as market solutions in efficiently organizing transactions.

As suggested by the previous chapters, many questions in the study of transactions and their governance remain unanswered. One of them has to do with the internal organization of firms and its impact on transaction costs, particularly in large business firms characterized by a dispersed ownership that makes their modalities of coordination the potential source of administrative costs that could transform them into mammoths condemned to disappear. In his contribution, Mark Roe examines this issue as two different aspects of property rights: the vertical problem of governance between shareholders and managers and the horizontal problem of governance between a controlling shareholder and distant shareholders. In other words, he focuses on agency problems between the shareholders, the board, and the managers. Roe reviews extensively the diversified institutional modalities, from information gathering and sharing to incentive compensation and corporate laws, that have been adopted to circumvent the risks of opportunistic behavior. His contribution shows how the existing literature often neglects issues surrounding the institutional embeddedness of these rules and mechanisms, as well as the way corporate governance is affected by political institutions or the more basic institutions of property and contracts.

The last two chapters of this section are about the different but very important topic of innovation, which often combines technological and organizational dimensions. Erin Anderson and Hubert Gatignon turn to the dynamic problem of how firms create new markets for both existing and new products. They consider markets as institutional arrangements resulting from interactions among firms and between firms and potential customers. Market creation thus results from both the internal challenge of governing to encourage innovation and the external challenge of acquiring innovations through acquisition or appropriation. New markets can be developed through in-house marketing efforts or by franchising or other forms of partnership, all requiring safeguards against the risk of opportunism. Uncertainty in the firm's environment as well as internally is a key issue here. This analysis substantiates the NIE perspective that markets should not be treated as black boxes: they can be analyzed as diverse outcomes of complex adjustments and innovations over time. Anderson and Gatignon show how NIE is the right analytical tool to understand the comparative business logic of these governance choices and the effects of path dependence and bounded rationality.

The final chapter of this section is about the substantial transformation of the organization of transactions through platforms and its consequence on regulation.

Martin Peitz delivers a complete and subtle analysis of the different dimensions of a phenomenon that has dramatically changed and continues to impact the way users have access to products and services and how they interact, leading to different network effects. Platforms create digital ecosystems that raise issues regarding price and non-price strategies as well as how to manage interactions on platforms. Platforms can benefit from horizontal as well as vertical scope economies, which lead to the adoption of very different strategies toward competitors as well as users. The diversity of practices, for example, platform exclusivity or bundling and tying choices, can lead to anti-competitive effects that demand new approaches from public regulation that often require going beyond the national boundaries of traditional regulatory institutions.

1.7 Contracts

Contracts have had a very special status in NIE ever since the pioneering article by Williamson (1971) on “The Vertical Integration of Production.” The previous sections discuss contracts as part of what Davis and North called the “institutional environment” or as important issues for organizations. This section focuses on recent developments in the theory of contracts, paying particular attention to the agreements known as “relational contracts,” a notion that has considerably evolved since its introduction to economics in the mid-1970s.

Rocco Macchiavello’s chapter first reminds us of the widespread role of relational contracts as well as their ubiquitous and messy characteristics since they are based on the trust required to sustain commitments among parties motivated by the prospect of future gains. Macchiavello then discusses recent research that improves our understanding of how relational contracts work in practice. His chapter includes (1) a section on the different methodologies for the empirical study of relational contracts, including game theory for framing models of “sustainable” relations and surveys for identifying and measuring the relational dimension, and (2) a review of some key empirical studies. Macchiavello shows that most of these studies have focused on the benefits that relational contracts bring to parties involved. However relational contracts may also favor “bad” transactions, for example, facilitating collusion, a behavior notoriously difficult to identify. In an innovative section, Macchiavello also discusses the potentially “ugly” face of relational contracts with their potential role at the core of undesirable transactions such as the “protection” secured by criminal organizations through the development of networks and a very specific organizational culture. The diversity of empirical studies reported in this chapter confirms the pervasive role of relational contracts in economic life.

In the same vein, Ricard Gil and Giorgio Zanarone provide a simple and clear exposition of the “canonical” model of relational contracting developed by Baker-Gibbons-Murphy and Levin, emphasizing the central role of delegation of authority in these contracts. This is also their point of entry to the analysis of the interactions between relational contracts and formal ones. One important lesson from the model and its illustration through several empirical studies is that parties may engage in

contracts even if they are imperfectly enforceable by courts, thanks to formal contracts that provide complementary support to the relational agreement. The authors also push the analysis further in showing that relational contracts may require specific organizational designs in order to successfully delegate authority. The complementarity between relational and formal contracts also generates specific challenges. Among them is the issue of lack of clarity, with ambiguities and language imperfections making both types of contracts incomplete and susceptible to misinterpretations, potentially leading to conflicts and even breaks in the relationship. Several cases illustrate the importance of managerial practices and of employees and/or external parties' commitment to build trust and reduce temptations to renege. The authors conclude with a call for more empirical studies on these issues.

While the previous contributions focused almost entirely on the role of contracts within firm or between firms, Jean Beuve and Stephane Saussier take a different stance: they look at how contracts in public procurement differ from private ones. A main difference comes from the interaction between the public authorities initiating and implementing such contracts and the role of political opponents and interest groups that interfere in all steps of the contractual process. What makes things complex is the diversity of motivation of these parties, some motivated by the specific interest of their constituencies, others by the need for transparency to limit risks of corruption, etc. Third-party opportunism might also interfere in the contractual process, for example, political opponents in the context of an electoral competition might systematically criticize a contract to win votes. An important consequence might be excessive rigidity of public contracts, which raises costs when it reduces capacity to adapt to unexpected events. These constraints also challenge the legitimacy of renegotiations, which could allow adaptation but also exposes the agreement to opportunistic behavior either by public authorities (as with corruption) or by opponents in the context of political contests. The result is a tradeoff between rules and discretion, with the benefits of greater discretion depending in part on the quality of public institutions and institutional actors. Here again, more empirical studies are needed.

Douglass Allen and Dean Lueck's chapter analyzes the strategic role played by contracts in agriculture. This issue has become particularly challenging in the context of ill-defined property rights in so many countries in which agriculture plays a central role and the transformations related to climate change. Focusing on farming contracts in North America, the authors show the importance of contracts at all biological stages of production. Beside the special attention paid by economists to crop share and cash rent contracts, Allen and Lueck also discuss the role of contracts in other steps of the production chain, for example, contracts limited by seasonal constraints. Contrary to rival explanations, transaction cost analyses do not treat contracting parties as principal and agent but instead examine the incentives of both parties to maximize wealth in an uncertain environment. This approach also helps us to better understanding the potential role of contracts in soil conservation, a topic that has recently attracted much attention; and it leads to considerations about how neglecting the interdependence between conservation and contract form has led to misleading policy recommendations. They conclude with the strong statement that

the transaction cost approach that has inspired so many empirical studies explains the basic facts regarding agricultural contracts.

1.8 Institutional Change

The contributions in this section confront the challenge that Doug North posed at the outset: move economics from a static to a dynamic theory by explaining how institutions change. Jean Phillipe Platteau's chapter examines the evolving ways in which institutional economists have grappled with this challenge since the end of the Second World War. His chapter explores how over the last 70 years NIE has shifted from crude evolutionism and induced institutional innovations, through equilibrium-of-the-game and political economy approaches, to the most recent, so-called gradual change approach, resting on rather sophisticated evolutionary and co-evolutionary models. He concludes that all existing approaches have their merits and domains of application and can be seen as essentially complementary and an eclectic attitude toward them is fully warranted.

Like Platteau, Mary Shirley's chapter on institutions and development agrees that NIE has made progress in understanding underdevelopment and development but still has a long way to go. This question remains pressing. Only 14% of the global population lives in high-income countries with strong institutions, and despite drastic declines in global poverty, that number has barely budged. Shirley's chapter summarizes how NIE addresses four questions: which institutions must function effectively for countries to converge with high-income countries? What are the obstacles to developing such strong institutions? How can countries improve their institutions? And can outsiders, including foreign aid agencies, help them develop? While research has made much progress in answering these questions, persistent puzzles remain. Thus, while there has been much progress in understanding the importance of and obstacles to achieving institutions that enforce contracts and protect investors and entrepreneurs against expropriation by the state or elites, considerably more research is needed to understand how countries achieve such institutions in the modern era.

Stanley Engerman and Kenneth Sokoloff concur that institutions are critical to any explanation of economic development but also suggest that no one institution is indispensable for growth. They emphasize how very different institutional structures have often been found to be reasonable substitutes for each other, both in dissimilar as well as similar contexts. Furthermore, they cite evidence that there are systematic patterns to the ways institutions evolve that undercuts the idea that exogenous change in institutions is what powers growth. They argue that institutions are profoundly influenced by the political and economic environment and that if any aspect of institutions is crucial for growth, it is that institutions change over time as circumstances change. Engerman and Sokoloff suggest that perhaps what matters for growth is not any particular institutional design but how well institutions are adapted to their specific settings and how flexible they are in adapting to changing circumstances.

Desiree Desierto and Mark Koyama survey NIE's various approaches to institutional change. They find the literature generally follows two perspectives. One suggests that institutions perform useful functions and change when circumstances make these functions obsolete except when transaction costs render this change too costly. The other suggests that both cooperation and conflict drive institutional change and that institutional change need not be efficient. Desierto and Koyama propose a framework for simultaneously depicting institutional change driven by both cooperation and conflict. They also show that understanding institutional changes requires detailed historical and institutional knowledge, which they exemplify in their chapter's analysis of important case studies, such as serfdom, feudalism, and the Magna Carta. Finally, they examine the role of individuals in driving institutional change.

In the last chapter of this section, Avner Greif examines the factors determining the dynamics of markets and market-supporting institutions. His analysis describes the key role played by two sets of institutions. First are "contract-enforcement institutions," the complex set of institutions required for securing exchanges. "Contract-enforcement institutions" can be organic, private-order institutions that arise spontaneously from the pursuit of individual interests or designed private- or public-order institutions that are intentionally created to secure contracts. Second are "coercion-constraining institutions," rules that constrain those with coercive power from abusing the property rights of others. Without coercion-constraining institutions, economic actors will be unwilling to bring their goods to market for fear that the rulers or other powerful actors will expropriate them. Greif describes how markets and political institutions co-evolve through the dynamic interaction of these two sets of institutions. Referring to historical examples, such as the contrast in the organic institutions implemented by the Maghribi traders and the designed institutions used by the Genoese, he shows the different forms that contract enforcement and coercion-constraining institutions can take and how they explain the dynamics of markets and political institutions. This problem of controlling coercive power is an issue central to several other contributions as well.

1.9 Culture and Informal Institutions

NIE in its infancy was focused mainly on what North termed "formal" institutions, written constitutions, laws, and rules enforced by third parties such as courts. Data on formal institutions was more available and although institutionalists always recognized the critical role of unwritten, "informal" institutions such as norms and conventions, they struggled to analyze them rigorously. Circumstances have changed thanks to new datasets and new methodologies (see the next section), and we have added a section (VII) to this edition of the Handbook to reflect the richness of this evolving scholarship on culture and informal institutions.

In their chapter, Samuel Bowles and Jung-Kyoo Choi argue that the earliest institutions—cooperative and egalitarian practices and other enduring social structures—have governed social interactions since the appearance of biologically

modern humans hundreds of thousands of years ago. Further, the subsequent emergence of private property as a fundamental institution predates states by many millennia. In the absence of states, institutions such as sharing the necessities of life beyond the family and respecting the property rights of others resembled linguistic conventions, where conformity with the rules was a mutual best response conditional on most others conforming. The authors also examine evidence that what the first states accomplished was not to put an end to Hobbes' chaotic state of nature and wars but instead to have expanded the scope of coordinated social interactions and to have inaugurated unprecedented and enduring levels of inequality of wealth.

Joel Mokyr's chapter tackles head on the distinction between institutions and culture by asking: is there a connection between the Industrial Revolution and the eighteenth century Enlightenment? The Enlightenment was a very broad cultural movement with many moving parts, but Mokyr focuses on the Industrial Enlightenment, a movement that directly advocated institutions that supported and promoted the accumulation and dissemination of technological progress and economic growth. These institutions included those that directly helped incentivize the accumulation of useful knowledge such as patronage and reputation effects, as well as patents, prizes, and other rewards to successful inventors. The Industrial Enlightenment also supported dissemination of knowledge through scholarly societies, publications, and various informal networks. Mokyr concludes that both this culture and the institutions it supported created a fertile ground for the rapid economic growth ignited by the Industrial Revolution.

Institutional economists have long been preoccupied by China, which early on was a subject of great interest to Ronald Coase. One reason for this fascination: China presents an institutional puzzle. How have private entrepreneurs driven its extraordinary growth despite the apparent weakness of the very institutions that NIE portrays as fundamental to economic dynamism, such as rule of law and secure private property rights? Part of the answer is China's robust informal institutions. Mary Shirley and Lixin Colin Xu examine the formal and informal institutions that explain both the rapid growth and current stagnation of China's private sector. They survey the deep historical roots of China's institutions and their evolution since the death of Mao in 1976. They then analyze the specific formal and informal institutions enforcing contracts between firms and providing (weaker) protections against expropriation by the state and elites, including such informal institutions as, for example, kinship ties, business networks, and links with local government officials, combined with improvements in some formal institutions such as property law and the judicial system. The chapter concludes with a consideration of the adverse effects of recent political changes on China's private sector and the future trajectory of China's growth.

Jared Rubin maintains that while it is undoubtedly true that religion has been among the most important social forces historically shaping individual behavior, it has only been in the last decade or two that practitioners of New Institutional Economics have taken religion seriously. Rubin's chapter surveys the growing NIE literature examining the role that religion and religious institutions have played in historical and contemporary economic development. With its focus on religious

institutions, Rubin's chapter covers issues at the intersection of religion and finance, law, political economy, and human capital. For example, Rubin explores how the roles of different religions in conferring legitimacy on the state have had large political economy effects. Thus, after the rise of Protestantism reduced the role of the church in providing legitimacy for the monarchy, rulers of England turned to economic elites and Parliament for legitimacy and ended up enacting policies more favorable to commerce. Given the centrality of institutions in the role of religion on outcomes, Rubin argues that scholars of the New Institutional Economics tradition should be front and center in major advances in our understanding.

Philip Keefer's chapter examines a paradox at the heart of the New Institutional Economics: institutions resolve problems of mistrust, but mistrust undermines the efficacy of institutions. Mistrust and opportunistic behavior drive the principal-agent and collective action dilemmas that are central to the New Institutional Economics but also interfere with the institutional and organizational reforms that can resolve these dilemmas. Keefer argues that understanding the extrinsic and intrinsic drivers of untrustworthy, opportunistic behavior are key to addressing the paradox. The NIE agenda focuses most on reforms that reduce extrinsic incentives (reputational, financial, judicial, or social) to be untrustworthy and act opportunistically. Yet Keefer maintains that intrinsically motivated trustworthy behavior is also key to solving the two dilemmas and that institutions that support social networks and reduce social distance will strengthen this intrinsic motivation.

In line with the earlier chapters finding increased attention to informal institutions, Alberto Simpser's chapter argues that informal institutions are increasingly viewed as an important part of the study of corruption. Simpser defines corruption as the abuse of entrusted power for private gain. As he discusses, corruption has long been understood to respond to costs and benefits shaped by formal, legal institutions. Yet Simpser suggest that increasingly informal norms are also receiving attention as drivers of corrupt behavior and as potential policy targets to combat corruption. Simpser surveys existing scholarship including empirical evidence about how social norms affect behavior, how they interact with the law, and how they might be influenced through policy interventions to combat corruption. For example, he cites evidence from randomized experiments showing the importance of social norms in corrupt or dishonest behavior. While recognizing that there is still much debate about how corruption can be curbed and a history of many unsuccessful interventions, Simpser's chapter concludes on a more optimistic note, highlighting promising ideas, suggestive evidence, and avenues for future research and policy development.

1.10 Methodologies

The last section of the Handbook deals with methodologies for analyzing institutions, considering both new techniques and established methodologies. Institutional economics presents methodological challenges for researchers because, as we described earlier, it rejects the simplifying assumptions of neoclassical economics,

such as perfect information, zero transaction costs, complete rationality, an objective and efficient legal system, a powerful state aiming to maximize human welfare, a black box firm seamlessly transforming inputs into outputs, etc. NIE has greatly enriched our view of the economy, while making it much more complicated to study. Adding to the complexity, NIE has focused on “big” questions of economic development and organization. “Thus, even as neoclassical models have been excluding questions and assuming away frictions and institutions to meet the requirements of more and more sophisticated mathematical tools. . . NIE scholars have prioritized content” (Ménard & Shirley, 2022, p. 15). The chapters in this section assess how NIE has worked to find new sources of data and adapt analytical tools to allow causal identification without ignoring institutions’ multifaceted and multidimensional characteristics.

Machine learning and other advanced computing techniques open the way for the process and analysis of a wealth of new data especially well-suited for institutional inquiries. In their chapter, Peter Grajzl and Peter Murrell show how machine learning can extend the scope of historical institutional and cultural analysis. After providing an overview of some of literature using this method, they use their own work on pre-nineteenth-century English caselaw and print culture to illustrate the possibilities and challenges of machine learning as a tool for systematic quantitative inquiry into the origins, change, and impact of institutions and culture. They show the power of this tool for distilling core facts and generating datasets amenable to conventional econometric analysis. Machine learning allowed them to explore the deep institutional roots of specific legal and cultural ideas, analyze the coevolution of ideas within caselaw and culture, examine the impact of caselaw on economic development both before and during the Industrial Revolution, and discern critical junctures in England’s legal and cultural development. Their chapter provides clear and detailed background for scholars interested in employing machine learning and illuminates the types of lessons that can be learned from the application of this method in new institutional economics. It also suggests a pathway to finding a practical, implementable set of methods among a proliferation of new techniques.

Despite a poor reputation among some economists, case studies have always been an important tool for institutional economists. Andrew Bennett’s chapter points out a number of advantages of case studies for studying economic institutions, including their strengths in analyzing path dependency, equifinality, high order interactions, anomalies, substantively important cases, complex concepts that are hard to measure, and causal mechanisms and processes. Bennett’s chapter outlines two new case study methods: typological theorizing and formal Bayesian process tracing, which he usefully illustrates using a study of state building and international finance in the global south in the 1900s. By illustrating how to use typological theorizing to analyze this case, Bennett provides detailed guidance to scholars interested in applying this technique. He concludes that wider and more rigorous use of case study methods can contribute greatly to new institutional economics and the field of economics more generally.

NIE has long been concerned with the factors that determine the nature and evolution of institutions. Robert Maseland’s chapter examines this diverse literature

from a methodological perspective. Maseland first categorizes the main institutional determinants in this literature as follows: (1) natural environment, (2) historical events, and (3) culture. He then assesses how each of the three determinants is theorized to impact contemporary outcomes and discusses the dominant methodological approaches used to identify those impacts. Maseland argues that the literature's emphasis on causal inference has led to a relative neglect of questions about the validity and interpretation of the data, as well as questions about the evolutionary mechanisms linking past determinants to present outcomes. He suggests that a turn to historical narratives may remedy some of these concerns but argues that a fuller appreciation of the disciplinary contexts from which the literature borrows (history and ethnography in particular) is imperative if the literature is to translate its intriguing findings into fuller insights on institutional development.

The final chapter by Omar Al-Ubaydli, Faith Fatchen, and John List explores whether field experiments would be a useful addition to the NIE toolkit of methodologies. After brief background on the basics of experiments, the authors describe some fundamental barriers to using field experiments to understand the impact of institutions on economic growth. Not only are such microeconomic experiments infeasible, they raise serious ethical questions. Yet while field experiments may be limited in answering questions in institutional economics with macroeconomic outcomes, the authors see great potential in employing field experiments to answer micro founded questions, such as issues surrounding property rights, state capacity, political accountability, legal systems, and so on. The authors survey some significant scholarly contributions using field experiments to study institutions and propose some future methods for using field experiments within institutional economics, emphasizing the utility of mixed methods combining field experiments with quantitative and qualitative analysis; interviews, surveys, observational data, and quasi-experiments.

We concluded the introduction to the first edition of this Handbook by challenging its readers to “fill in the blanks,” to find answers to the many unanswered questions that leaders in the field considered crucial. This second edition explores the wealth of new research that has met that challenge as well as methodologies providing better tools for exploring answers to these and new questions. As is true of any scientific endeavor, with expanded knowledge and fresh insights comes new puzzles and challenges. Fittingly, these chapters describe extraordinary progress coupled with an ever more ambitious research agenda.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Arrow, K. J. (1987). Reflections on the essays. In G. Feiwel (Ed.), *Arrow and the foundations of the theory of economic policy* (pp. 727–734). New York University Press.
- Coase, R. H. (1960). The problem of social cost. *The Journal of Law and Economics*, 3, 1–44.

- Coase, R. H. (1988). *The firm, the market and the law*. The University of Chicago Press.
- Furubotn, E. G., & Richter, R. (1998). *Institutions and economic theory: The contribution of the new institutional economics*. The University of Michigan Press.
- Ménard, C., & Shirley, M. M. (2022). *Advanced introduction to new institutional economics*. Edward Elgar.
- North, D. C. (1990). *Institutions, institutional change, and economic performance*. Cambridge University Press.
- North, D. C. (1997). Transaction costs through time. In C. Ménard (Ed.), *Transaction cost economics: Recent developments*. Edward Elgar.
- North, D. C. (2005). *Understanding the process of economic change*. Princeton University Press.
- North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.
- Weingast, B. R. (1993). Constitutions as governance structures: The political foundations of secure markets. *Journal of Institutional and Theoretical Economics (JITE)*, 149(1), 286–311.
- Williamson, O. E. (1971). The vertical integration of production: Market failure considerations. *American Economic Review*, 61, 112–123.
- Williamson, O. E. (1985). *The Economic Institutions of Capitalism*. The Free Press

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part I

The Domain of New Institutional Economics

Institutions and the Performance of Economies Over Time

2

Douglass C. North

Contents

2.1	Introduction	26
2.2	Interactions in a World of Uncertainty	26
2.2.1	Let Me Expand on These Propositions	27
2.3	Process of Economic Change	29
2.4	Implications and Lacunae	33
2.5	Conclusion	34
	References	34

Abstract

In this chapter, I intend to provide an approach to the study of the process of economic change. The discipline of economics is made up of a static body of theory that explores the efficiency of resource allocation at an instant of time and under the restrictive assumptions of frictionless markets. But the first constraint of static analysis severely hinders our ability to analyze and improve the performance of economies in a world of continuous change. There is still much that we do not understand about the process of economic change, but this chapter provides an analytical framework that does, I believe, highlight the problems that must be confronted in order to understand and improve economic performance. I first describe the intentional nature of human interaction in a world of pervasive uncertainty (Sect. 2) before going on to describe the process of economic change (Sect. 3). I conclude with drawing some implications from this approach to the process of change which highlights the lacunae in our understanding of this process (Sect. 4).

Douglass North sadly passed away before this second edition of the *Handbook of New Institutional Economics* was created. His chapter from the first edition remains a classic and emblematic of his invaluable contributions to the field.

D. C. North
Washington University in St. Louis, St. Louis, MO, USA

Keywords

Institutions · Institutional change · Economic change · Incentive structure · Path dependency · Belief systems · Norms

2.1 Introduction

The discipline of economics is made up of a static body of theory that explores the efficiency of resource allocation at an instant of time and under the restrictive assumptions of frictionless markets. Recent research has explored the nature of the frictions by incorporating institutions, transaction costs, and political economy into economic analysis, thereby providing the theory with a bridge to the real world of real economies. But the first constraint of static analysis severely hinders our ability to analyze and improve the performance of economies in a world of continuous change. And, in fact, the employment of static theory as a source of policy recommendation in a setting of dynamic change is a prescription for the policies producing unanticipated and undesirable results. In this chapter, I intend to provide an approach to the study of the process of economic change. There is still much that we do not understand about the process but this chapter provides an analytical framework that does, I believe, highlight the problems that must be confronted in order to understand and improve economic performance. I first describe the intentional nature of human interaction in a world of pervasive uncertainty (Sect. 2.2) before going on to describe the process of economic change (Sect. 2.3). I conclude with drawing some implications from this approach to the process of change which highlights the lacunae in our understanding of this process (Sect. 2.4).

2.2 Interactions in a World of Uncertainty

In contrast to standard theory that draws its inspiration from physics, modeling the process of change must derive its inspiration from evolutionary biology but in contrast to Darwinian theory in which the selection mechanisms are not informed by beliefs about the eventual consequences, human evolution is guided by the perceptions of the players in which choices—decisions—are made in the light of these perceptions with the intent of producing outcomes downstream that will reduce the uncertainty of the organizations—political, economic, and social—in pursuit of their goals.¹ Institutional change, therefore, is a deliberate process shaped by the perceptions of the actors about the consequences of their actions. The immediate vehicle by which the actors attempt to shape their environment is by altering the

¹This section is drawn from my essay “Five Propositions about Institutional Change,” in Knight and Sened (1995).

institutional framework in order to improve their (and their organizations') competitive position. Let me state five propositions that describe this process:

1. The continuous interaction between institutions and organizations in the economic setting of scarcity and hence competition is the key to institutional change.
2. Competition forces organizations continually to invest in new skills and knowledge to survive. The kind of skills and knowledge individuals and their organizations acquire will shape evolving perceptions about opportunities and hence choices that will incrementally alter institutions.
3. The institutional framework provides the incentive structure that dictates the kinds of skills and knowledge perceived to have the maximum payoff.
4. Perceptions are derived from the mental constructs of the players.
5. The economies of scope, complementarities, and network externalities of an institutional matrix make institutional change overwhelmingly incremental and path dependent.

2.2.1 Let Me Expand on These Propositions

1. Institutions are the rules of the game—both formal rules, informal norms, and their enforcement characteristics. Together they define the way the game is played. Organizations are the players. They are made up of groups of individuals held together by some common objectives. Economic organizations are firms, trade unions, cooperatives, etc.; political organizations are political parties, legislatures, and regulatory bodies; educational organizations are universities, schools, and vocational training centers. The immediate objective of organizations may be profit maximizing (for firms) or improving reelection prospects (for political parties); but the ultimate objective is survival because all organizations live in a world of scarcity and hence competition.
2. New or altered opportunities may be perceived to be a result of exogenous changes in the external environment which alter relative prices to organizations or a consequence of endogenous competition among the organizations of the polity and the economy. In either case the ubiquity of competition in the overall economic setting of scarcity induces entrepreneurs and the members of their organizations to invest in skills and knowledge. Whether through learning by doing on the job or the acquisition of formal knowledge, improving the efficiency of the organization relative to that of rivals is the key to survival.

While idle curiosity surely is an innate source of acquiring knowledge among human beings, the rate of accumulating knowledge is clearly tied to the payoffs. Secure monopolies, be they organizations in the polity or in the economy, simply do not have to improve to survive. But firms, political parties, or even institutions of higher learning faced with rival organizations must strive to improve their efficiency. When competition is muted (for whatever reasons), organizations will have less incentive to invest in new knowledge and in consequence will not induce rapid institutional change. Stable institutional structures will be the result.

Vigorous organizational competition will accelerate the process of institutional change.

3. There is no implication in proposition 2 of evolutionary progress or economic growth—only of change. The institutional matrix defines the opportunity set, be it one that makes income redistribution the highest payoff in an economy or one that provides the highest payoffs to productive activity. While every economy provides a mixed set of incentives for both types of activity, the relative weights (as between redistributive and productive incentives) are crucial factors in the performance of economies. The organizations that come into existence will reflect the payoff structure. More than that, the direction of their investment in skills and knowledge will equally reflect the underlying incentive structure. If the highest rate of return in an economy comes from piracy, we can expect that the organizations will invest in skills and knowledge that will make them better pirates. Similarly if there are high returns to productive activities, we will expect organizations to devote resources to investing in skill and knowledge that will increase productivity (the new growth economics literature can become relevant at this point).

The immediate investment of economic organizations in vocational and on-the-job training obviously will depend on the perceived benefits; but an even more fundamental influence on the future of the economy is the extent to which societies will invest in formal education, schooling, the dissemination of knowledge, and both applied and pure research which will mirror the perceptions of the entrepreneurs of political and economic organizations.

4. The key to the choices that individuals make is their perceptions about the payoffs, which are a function of the way the mind interprets the information it receives. The mental constructs individuals form to explain and interpret the world around them are partly a result of the genetic evolution of the mind, partly of their cultural heritage, partly a result of the local everyday problems they confront and must solve, and partly a result of non-local learning. The mix among these sources in interpreting one's environment obviously varies as between, for example, a Papuan tribesman on the one hand and an economist in the United States on the other (although there is no implication that the latter's perceptions are independent of his or her cultural heritage).

The implication of the foregoing paragraph is that individuals from different backgrounds will interpret the same evidence differently; they may, in consequence, make different choices. If the information feedback of the consequences of choices were complete, then individuals with the same utility function would gradually correct their perceptions and over time converge to a common equilibrium; but as Frank Hahn has succinctly put it, "There is a continuum of theories that agents can hold and act upon without ever encountering events which lead them to change their theories" (Hahn, 1987, p. 324). The result is that multiple equilibria are possible due to different choices by agents with identical tastes.

5. The viability, profitability, and indeed survival of the organizations of a society typically depend on the existing institutional matrix. That institutional structure has brought them into existence; and their complex web of inter-dependent contracts and other relationships has been constructed on it. Two implications follow. Institutional change is typically incremental and is path dependent.

This institutional change is occurring in a world of pervasive uncertainty or ambiguity which by definition is one in which one cannot derive a probability distribution of possible outcomes—such as is possible with decision-making in the face of risk (in the Knightian definitions). This uncertainty persists because the “human landscape” in which humans are interacting is continually undergoing change—change induced in part by non-human action (e.g., changes in climate, natural disasters) but primarily by the human actors themselves.

Humans attempt to reduce that uncertainty (or convert it into risk) by learning. The cumulative learning of a society embodied in language, beliefs, myths, and ways of doing things—in short the culture of a society—not only determines societal performance at a moment of time but through the way in which it constrains the choices of the players contributes to the nature of the process through time. Humans scaffold both the mental models they possess—belief systems—and the external environment, institutions. The focus of our attention, therefore, must be on human learning, on what is learned and how it is shared among the members of a society, on the incremental process by which the beliefs and preferences change through time, and on the way in which they shape the performance of economies through time.

We can describe that performance by innumerable statistics on its demographic, economic, technological, and institutional features; but what we really need to know is what is the interplay between all these features that makes it work. The foundations of the interplay are three: the demography, which describes the quantity and quality of human beings; the stock of knowledge that the society possesses, which determines the human command over nature; and the institutional framework that determines the rules of the game. The demographic characteristics include not only the fertility, mortality, and migration characteristics and the labor force composition but also the stock of human capital (derived from the stock of knowledge). The stock of knowledge includes not only the scientific knowledge that a society possesses, its distribution in the society, and its application to solving problems of scarcity, but also the beliefs that the society holds that influence the choices made. That stock of knowledge determines the potential upper bound of the well-being of the society. The institutional framework determines the incentive structure of the society. It is the interplay between these three that shapes the features of the economy. We know very little about this interaction, although we do have some limited hypotheses about parts of the interaction. Self-conscious modeling of this interaction at a moment of time, much less over time, has not been part of the agenda of economists, development economists, or economic historians. But with this background we are now ready to explore the process of economic change.

2.3 Process of Economic Change

A bare-bones description of that process is straightforward. The “reality” of a political-economic system is never known to anyone, but humans do construct elaborate beliefs about the nature of that “reality”—beliefs that are both a positive

model of the way the system works and a normative model of how it should work. The belief system may be broadly held within the society, reflecting a consensus of beliefs; or widely disparate beliefs may be held, reflecting fundamental divisions in perception about the society. The dominant beliefs—those of political and economic entrepreneurs in a position to make policies—produce over time an elaborate structure of institutions, both formal rules and informal norms, that determines economic/political performance. The resultant institutional matrix imposes severe constraints on the choice set of entrepreneurs when they seek to introduce new or modified institutions in order to improve their economic or political positions. The resultant path dependence typically makes change incremental. But change is continually occurring (although the rate will depend on the degree of competition among organizations and their entrepreneurs) as entrepreneurs enact policies to improve their competitive position—policies that result in alterations of the institutional matrix described in the previous section. The result is revised perceptions of reality and in new efforts by entrepreneurs to improve their position. The process of change is never ending. Change can also come from non-human-induced changes in the human landscape, such as natural disasters; but overwhelmingly it is humans themselves who are incrementally altering the human landscape, as even the most cursory overview of human history will attest.

It is one thing to be able to provide a summary description of the process of economic change; it is something else to provide sufficient content to this description to give us an understanding of this process. What do we mean by reality? How do beliefs get formed? How do they change? What is the relationship between beliefs and institutions?

I have nothing to add to the age-old question of philosophers—what is reality? But I do have a direct pragmatic interest in just what it is that we are trying to model in our theories, beliefs, and ideologies. The pragmatic concern is with the degree to which our beliefs coincide with “reality.” To the extent that they do, then there is some prospect that the policies we enact will produce the intended result. The model is always a very imperfect reflection of how the economy really works. In some cases the defects are fatal, as in the case of the communist economies that disintegrated in 1989.

Beliefs and the way they evolve are at the heart of understanding the process of change. For the most part, economists, with a few important exceptions like Hayek, have ignored the role of ideas in making choices. While the rationality assumption has served economists well for a limited range of issues in micro-theory, it has devastating limitations in dealing with the process of economic change. The way we perceive the world and construct our explanations about that world requires that we delve into how the mind and brain work—the subject matter of cognitive science. We are some distance from a theory of learning that would account for how the mind works but we can at least outline the nature of the process.

The first level of learning entails developing a structure by which to make sense of the varied signals received by the senses. The initial architecture of the structure is genetic, but its subsequent development is a result of the experiences of the individual. This architecture can be thought of as generating an event space which gets us to interpret the data provided by the world. The experiences can be classified into two

kinds—those from the physical environment and those from the socio-cultural linguistic environment (Hutchins & Hazlehurst, 1992). The event space structure consists of categories—classifications that gradually evolve from earliest childhood on in order to organize our perceptions and keep track of our memory of analytic results and experiences. Building on these categories, we form mental models to explain and interpret the environment, typically in ways relevant to some goals (Holland et al., 1986, p. 22). Both the categories and the mental models will evolve to reflect the feedback derived from new experiences—feedback that may strengthen and confirm our initial categories and models or that may lead to modifications. Thus, the event space may be continually redefined with experience, including contact with others' ideas.

Learning which preserves the categories and concepts intact but provides changed ideas about details and the applicability of the existing knowledge is the second level of learning. Together, learning within a given set of concepts and learning which changes the structure of concepts and mental models suggest an approach to the dynamics of learning.

The belief systems that evolve from learning induce political and economic entrepreneurs in a position to make choices that shape micro- and macroeconomic performance to erect an elaborate structure of rules, norms, conventions, and beliefs embodied in constitutions, property rights, and informal constraints; these in turn shape economic performance. This “scaffolding” not only constrains the choice set at a moment of time but is the source of path dependence. Thus when political or economic entrepreneurs seek to alter some aspect of economic performance, they make choices that are constrained not only by the standard constraints of technology and income but also by this scaffolding. The process of institutional change described above is intended to alter performance in a particular direction. The aggregate of such institutional changes is continually altering the way the economy works. In turn that leads to gradual alterations of the models we devise in a never-ending process of economic change.

Throughout history, humans have typically gotten it (at least partly) wrong in (1) their understanding of the way the economy works, (2) the synthetic frameworks they construct, or (3) the policies they enact (at best blunt instruments to serve their purposes) which produce unanticipated consequences. We may write economic history as a great success story of the enormous increase in material well-being which has reflected the secular growth in the stock of knowledge. But it is also a vast panorama of decisions that have produced death, famine, starvation, defeat in warfare, economic decline and stagnation, and indeed the total disappearance of civilizations. And even the decisions made in the success stories have typically been an admixture of luck intermingled with shrewd judgments and unanticipated outcomes. Take American economic history. From the earliest attempts at settlement, through the colonial era, to the perceptions leading to the revolutionary war, the colonists had it, at best, half right. The Constitution, surely a classic of shrewd judgment, was aided by chance (the events of the 1780s), luck (the boycott of the Convention by the anti-federalists), and unanticipated decisions (the development of the independent judiciary and the Marshall court).

I wish to emphasize the limits to our understanding because there is a certain amount of hubris evident in the annual surveys by the World Bank and in the writing of orthodox economists who think that now we have it right. But it is important that we understand that even if we did have it right for one economy it would not necessarily be right for another economy and even if we have it right today it would not necessarily be right tomorrow. I am not suggesting that we haven't learned a good deal about determinants of economic performance. We have; but the implications of my brief survey of the sequence of steps from our understanding of an economy, to the scaffold we erect, to the policies we then enact to alter economic performance are that there are innumerable junctures where we can and do get it wrong. Crucial junctures, critical to the issues of improving the performance of economies, have resulted from the way scaffolds have evolved and policies were formed as well as the way time has affected the formation of beliefs.

Scaffolds include the political structure that specifies how we develop and aggregate political choices, the property rights structure that defines the formal incentives in the economy, and the informal constraints of norms, conventions, and internally held beliefs. They have evolved over many generations, reflecting, as Hayek has reminded us (1960), the trial and error process which has sorted out those behavioral patterns that have worked from those that have failed. Because the experience of every society has been unique, they will differ for each economy. They constrain the choice set not only because the organizations of that economy have been built on the foundations of that institutional structure and therefore their survival depends on its continuance but also, and perhaps more fundamentally, because the belief system that is a complementary part of that scaffolding tends to change very slowly. This scaffolding is what makes path dependence so important. When the scaffolding crumbles, as it did in Eastern Europe in 1989, the problems of constructing a new framework have exposed our limited understanding of the process of change.

Equally crucial are the policies that we enact to alter the performance of an economy. Even when we have a "correct" understanding of the economy and the (more or less) "correct" theory about its operation, the policies at our disposal are very blunt instruments. They consist of alterations in the formal rules only, when in fact the performance of an economy is an admixture of the formal rules, the informal norms, and their enforcement characteristics. Changing only the formal rules will produce the desired results only when the informal norms that are complementary to that rule change and enforcement is either perfect or at least consistent with the expectations of those altering the rules.

Finally, time is important because it is the dimension in which human learning occurs and there is no implication in the foregoing brief description of the process of learning that suggests that we get it right. Indeed throughout history, we have gotten it wrong far more often than we have gotten it right. The rise and fall of communism in the twentieth century is only a recent illustration. It is probably correct that if "reality" stayed constant the feedback from the policies we enacted would gradually lead us to get it right, but change and therefore persistent uncertainty is our lot which guarantees that we will continue to get it wrong at least part of the time.

2.4 Implications and Lacunae

The implications of the foregoing brief outline of the process of economic change are straightforward. If our objective is to improve the long run performance of economies, we are in possession of the essential characteristics of successful economies. The best single predictor of the growth of an economy remains its investment rate. The new growth economics literature highlights some of the specific features of successful development. What is glaringly missing from this literature is the incentive structure to realize these objectives. But we do know a good deal about the institutional foundations of successful development. A number of recent empirical studies have made clear the importance of the institutional matrix (see Knack & Keefer, 1995 for a good summary). That matrix broadly comprises the incentive structure which will determine the quantity and quality of investment. What is still missing is how to get there. The key is the way path dependence will constrain the process of institutional and economic change.

The implication of the foregoing analysis is that path dependence can and will produce a wide variety of patterns of development depending on the cultural heritage and specific historical experience of the economy. Indeed the success of TVEs (township and village enterprises—a form of organization that is neither a firm nor a cooperative) in China does not fit our preconceptions about successful institutional/organizational structures and has been a sobering reminder of how much we still have to learn about the process. A description of that process in China from the enactment of the household responsibility system traces a unique path which has produced (so far) rapid economic growth (although even that success is tempered by growing problems of TVEs). I would hope that this paper puts to rest for good any simplistic general nostras such as “big bang” or “shock therapy” theories to magically overcome lack of development.

If path dependence can help us to understand the variety of development patterns, it also speaks forcefully to the constraints that the scaffolds erected in an economy impose on institutional change. The historically derived constraints are supported not only by the existing organizations that oppose change but also by the belief system that has evolved to produce those constraints. The rate and direction of change will be determined by the “strength” of the existing organizations and belief system.

The demise of communism in Eastern Europe in 1989 reflected a collapse of the existing belief system and consequent weakening of the supporting organizations. Policy makers were confronted not only by the problems of restructuring an entire society but also by the blunt instrument that is inherent in policy changes that can only alter the formal rules but cannot alter the accompanying norms and even have had only limited success in inducing enforcement of policies. The relative success of policy measures—such as the auctioning of state assets and the reestablishment of a legal system—in the Czech Republic compared to Russia resulted from the heritage of informal norms that made for the relatively harmonious establishment of the new rules (although even here the system for shifting assets from public to private hands in the Czech Republic produced some adverse and unanticipated downstream consequences).

One of the shortcomings of research is the lack of attention paid to the polity and the problem of aggregating choices through the political system. We simply have no good models of polities in third-world, transition, or other economies. The interface between economics and politics is still in a primitive state in our theories but its development is essential if we are to implement policies consistent with intentions.

2.5 Conclusion

Let me conclude by talking again about time. If you accept the crude schematic outline of the process of change I laid out in Sect. 2.3 above, it is clear that change is an ongoing continuous affair and that typically our institutional prescriptions reflect the learning from past experience. But there is no guarantee that the past experiences are going to equip us to solve new problems. Indeed an historic dilemma of fundamental importance has been the difficulties of economies shifting from a political economy based on personal exchange to one based on impersonal exchange. An equally wrenching change can be the movement from a “command” economy to a market economy. In both cases the necessity to restructure institutions—both economic and political—has been a major obstacle to development; it still is the major obstacle for third-world and transition economies. The belief system that has evolved as a result of the cumulative past experiences of a society has not equipped the members to confront and solve the new problems.

We are just beginning systematically to explore the process of economic change. Our laboratory is not only our history but, particularly, what we are learning in the ongoing efforts to improve the performance of third-world and transition economies. We have made some progress but we still have a long way to go.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute

References

- Hahn, F. H. (1987). Information, dynamics, and equilibrium. *Scottish Journal of Political Economy*, 34(4), 321–334.
- Hayek, F. A. (1960). *The constitution of liberty*. The University of Chicago Press.
- Holland, J. H., Idalyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. MIT Press.
- Hutchins, E., & Hazlehurst, B. (1992). Learning in the cultural process. In C. G. Langston, C. Taylor, I. Farmer, & S. Rasmussen (Eds.), *Artificial life II* (pp. 682–706). Addison-Wesley.
- Knack, S., & Keefer, P. (1995). Institutional and economic performance: Cross country tests using alternative institutional measures. *Economics and Politics*, 7(3), 207–227.
- Knight, J., & Sened, I. (1995). *Exploring social institutions*. The University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Institutional Structure of Production

3

Ronald H. Coase

Abstract

My contribution to economics has been to urge the inclusion in our analysis of features of the economic system so obvious that they have tended to be overlooked. Nonetheless, once included in the analysis, they will, as I believe, bring about a complete change in the structure of economic theory, at least in what is called price theory or microeconomics. What I have done is to show the importance for the working of the economic system of what may be termed the institutional structure of production. I shall explain why, in my view, these features of the economic system were ignored and why their recognition will lead to a change in the way we analyze the working of the economic system and in the way we think about economic policy, changes which are already beginning to occur. I will also address the empirical work that needs to be done if this transformation in our approach is to increase our understanding.

Keywords

Institutions · Institutional structure of production · Transaction costs · Economic systems · Economic policy · Price theory

In my long life I have known some great economists, but I have never counted myself among their number nor walked in their company. I have made no innovations in high

This paper is published with minor changes made by Ronald Coase, with the kind permission of the Nobel Foundation, 1991.

Ronald Coase sadly passed away before this second edition of the *Handbook of New Institutional Economics* was created. His Nobel speech, which constitutes this chapter from the first edition, remains a classic and emblematic of his invaluable contributions to the field.

R. H. Coase
University of Chicago Law School, Chicago, IL, USA

theory. My contribution to economics has been to urge the inclusion in our analysis of features of the economic system so obvious that, like the postman in G. K. Chesterton's Father Brown tale, "The Invisible Man," they have tended to be overlooked. Nonetheless, once included in the analysis, they will, as I believe, bring about a complete change in the structure of economic theory, at least in what is called price theory or microeconomics. What I have done is to show the importance for the working of the economic system of what may be termed the institutional structure of production. In this lecture I shall explain why, in my view, these features of the economic system were ignored and why their recognition will lead to a change in the way we analyze the working of the economic system and in the way we think about economic policy, changes which are already beginning to occur. I will also speak about the empirical work that needs to be done if this transformation in our approach is to increase our understanding. In speaking about this transformation, I do not wish to suggest that it is the result of my work alone. Oliver Williamson, Harold Demsetz, and Steven Cheung, among others, have made outstanding contributions to the subject, and without their work and that of many others, I doubt whether the significance of my writings would have been recognized. While it has been a great advantage of the creation of the Prize in Economic Sciences in Memory of Alfred Nobel that, by drawing attention to the significance of particular fields of economics, it encourages further research in them, the highlighting of the work of a few scholars, or, in my case, one scholar, tends to obscure the importance of the contributions of other able scholars whose researches have been crucial to the development of the field.

I will be speaking of that part of economics which has come to be called industrial organization, but to understand its present state, it is necessary to say something about the development of economics in general. During the two centuries since the publication of *The Wealth of Nations*, the main activity of economists, it seems to me, has been to fill the gaps in Adam Smith's system, to correct his errors, and to make his analysis vastly more exact. A principal theme of *The Wealth of Nations* was that government regulation or centralized planning were not necessary to make an economic system function in an orderly way. The economy could be coordinated by a system of prices (the "invisible hand") and, furthermore, with beneficial results. A major task of economists since the publication of *The Wealth of Nations*, as Harold Demsetz (1988, volume I, p. 145) has explained, has been to formalize this proposition of Adam Smith. The given factors are technology and the tastes of consumers, and individuals, who follow their own interest, are governed in their choices by a system of prices. Economists have uncovered the conditions necessary if Adam Smith's results are to be achieved and where, in the real world, such conditions do not appear to be found, they have proposed changes which are designed to bring them about. It is what one finds in the textbooks. Harold Demsetz has said rightly that what this theory analyzes is a system of extreme decentralization. It has been a great intellectual achievement, and it throws light on many aspects of the economic system. But it has not been by any means all gain. The concentration on the determination of prices has led to a narrowing of focus which has had as a result the neglect of other aspects of the economic system. Sometimes, indeed, it seems as though economists conceive of their subject as being concerned only with the pricing system and anything outside this is considered as no part of their business. Thus, my

old chief and wonderful human being, Lionel Robbins, wrote, in *The Nature and Significance of Economic Science*, about the “glaring deficiencies” of the old treatment of the theory of production with its discussion of peasant proprietorships and industrial forms: “It suggests that from the point of view of the economist ‘organisation’ is a matter of internal industrial (or agricultural) arrangement—if not internal to the firm, at any rate internal to ‘the’ industry. At the same time it tends to leave out completely the governing factor of all productive organisation—the relationship of prices and cost. . .” (Robbins, 1932, p. 70). What this comes down to is that, in Robbins’s view, an economist does not interest himself in the internal arrangements within organizations but only in what happens on the market, the purchase of factors of production, and the sale of the goods that these factors produce. What happens in between the purchase of the factors of production and the sale of the goods that are produced by these factors is largely ignored. I do not know how far economists today share Robbins’s attitude but it is undeniable that microeconomics is largely a study of the determination of prices and output; indeed, this part of economics is often called price theory.

This neglect of other aspects of the system has been made easier by another feature of modern economic theory—the growing abstraction of the analysis, which does not seem to call for a detailed knowledge of the actual economic system or, at any rate, has managed to proceed without it. Bengt Holmstrom and Jean Tirole (1989) writing on “The Theory of the Firm” in the recently published *Handbook of Industrial Organization* conclude at the end of their article of 63 pages that “the evidence/theory ratio. . . is currently very low in this field” (p. 126). Sam Peltzman (1991) has written a scathing review of the *Handbook* in which he points out how much of the discussion in it is theory without any empirical basis. What is studied is a system which lives in the minds of economists but not on earth. I have called the result “blackboard economics.” The firm and the market appear by name but they lack any substance. The firm in mainstream economic theory has often been described as a “black box.” And so it is. This is very extraordinary given that most resources in a modern economic system are employed within firms, with how these resources are used dependent on administrative decisions and not directly on the operation of a market. Consequently, the efficiency of the economic system depends to a very considerable extent on how these organizations conduct their affairs, particularly, of course, the modern corporation. Even more surprising, given their interest in the pricing system, is the neglect of the market or more specifically the institutional arrangements which govern the process of exchange. As these institutional arrangements determine to a large extent what is produced, what we have is a very incomplete theory. All this is beginning to change, and in this process I am glad to have played my part. The value of including such institutional factors in the corpus of mainstream economics is made clear by recent events in Eastern Europe. These ex-communist countries are advised to move to a market economy, and their leaders wish to do so, but without the appropriate institutions no market economy of any significance is possible. If we knew more about our own economy, we would be in a better position to advise them.

What I endeavored to do in the two articles cited by the Royal Swedish Academy of Sciences was to attempt to fill these gaps or more exactly to indicate the direction

in which we should move if they are ultimately to be filled. Let me start with "The Nature of the Firm." I went as a student to the London School of Economics in 1929 to study for a Bachelor of Commerce degree, specializing in the Industry group, supposedly designed for people who wished to become works managers, a choice of occupation for which I was singularly ill-suited. However, in 1931 I had a great stroke of luck. Arnold Plant was appointed Professor of Commerce in 1930. He was a wonderful teacher. I began to attend his seminar in 1931, some 5 months before I took the final examinations. It was a revelation. He quoted Sir Arthur Salter: "The normal economic system works itself." And he explained how a competitive economic system coordinated by prices would lead to the production of goods and services which consumers valued most highly. Before being exposed to Plant's teaching, my notions on how the economy worked were extremely woolly. After Plant's seminar, I had a coherent view of the economic system. He introduced me to Adam Smith's "invisible hand." As I had taken the first year of university work while still at high school, I managed to complete the requirements for a degree in 2 years. However, University regulations required 3 years of residence before a degree could be granted. I had therefore a year to spare. I then had another stroke of luck. I was awarded a Cassel traveling scholarship by the University of London. I decided to spend the year in the United States, this being treated as a year's residence at the London School of Economics, the regulations being somewhat loosely interpreted.

I decided to study vertical and lateral integration of industry in the United States. Plant had described in his lectures the different ways in which various industries were organized, but we seemed to lack any theory which would explain these differences. I set out to find it. There was also another puzzle which, in my mind, needed to be solved and which seemed to be related to my main project. The view of the pricing system as a coordinating mechanism was clearly right, but there were aspects of the argument which troubled me. Plant was opposed to all schemes, then very fashionable during the Great Depression, for the coordination of industrial production by some form of planning. Competition, according to Plant, acting through a system of prices, would do all the coordination necessary. And yet we had a factor of production, management, whose function was to coordinate. Why was it needed if the pricing system provided all the coordination necessary? The same problem presented itself to me at that time in another guise. The Russian Revolution had taken place only 14 years earlier. We knew then very little about how planning would actually be carried out in a communist system. Lenin had said that the economic system in Russia would be run as one big factory. However, many economists in the West maintained that this was an impossibility. And yet there were factories in the West, and some of them were extremely large. How did one reconcile the views expressed by economists on the role of the pricing system and the impossibility of successful central economic planning with the existence of management and of these apparently planned societies, firms, operating within our own economy?¹

¹A fuller account of these events will be found in Oliver E. Williamson and Sidney G. Winter (editors), 1991, pp. 34–47.

I found the answer by the summer of 1932. It was to realize that there were costs of using the pricing mechanism. What the prices are have to be discovered. There are negotiations to be undertaken, contracts have to be drawn up, inspections have to be made, arrangements have to be made to settle disputes, and so on. These costs have come to be known as transaction costs. Their existence implies that methods of coordination alternative to the market, which are themselves costly and in various ways imperfect, may nonetheless be preferable to relying on the pricing mechanism, the only method of coordination normally analyzed by economists. It was the avoidance of the costs of carrying out transactions through the market that could explain the existence of the firm in which the allocation of factors came about as a result of administrative decisions (and I thought it did). In my 1937 article I argued that in a competitive system there would be an optimum of planning since a firm, that little planned society, could only continue to exist if it performed its coordination function at a lower cost than would be incurred if it were achieved by means of market transactions and also at a lower cost than this same function could be performed by another firm. To have an efficient economic system, it is necessary not only to have markets but also areas of planning within organizations of the appropriate size. What this mix should be we find as a result of competition. This is what I said in my article of 1937. However, as we know from a letter I wrote in 1932 which has been preserved, all the essentials of this argument had been presented in a lecture I gave in Dundee at the beginning of October 1932 (see Williamson & Winter, 1991, pp. 34–5). I was then 21 years of age, and the sun never ceased to shine. One could never have imagined that these ideas would become some 60 years later a major justification for the award of a Nobel Prize. And it is a strange experience to be praised in my 80s for work I did in my 20s.

There is no doubt that the recognition by economists of the importance of the role of the firm in the functioning of the economy will prompt them to investigate its activities more closely. The work of Oliver Williamson and others has led to a greater understanding of the factors which govern what a firm does and how it does it². But it would be wrong to think that the most important consequence for economics of the publication of “The Nature of the Firm” has been to direct attention to the importance of the firm in our modern economy, a result which, in my view, would have come about in any case. What I think will be considered in the future to have been the important contribution of this article is the explicit introduction of transaction costs into economic analysis. I argued in “The Nature of the Firm” that the existence of transaction costs leads to the emergence of the firm. But the effects are pervasive in the economy. Businessmen in deciding on their ways of doing business and on what to produce have to take into account transaction costs. If the costs of making an exchange are greater than the gains which that exchange would bring, that exchange would not take place and the greater production that would flow from specialization would not be realized. In this way transaction costs affect not only contractual

² Amendment made by the author, R.H.C., June 2003.

arrangements, but also what goods and services are produced. Not to include transaction costs in the theory leaves many aspects of the working of the economic system unexplained, including the emergence of the firm, but much else besides. In fact, a large part of what we think of as economic activity is designed to accomplish what high transaction costs would otherwise prevent or to reduce transaction costs so that individuals can freely negotiate and we can take advantage of that diffused knowledge of which Hayek has told us.

I know of only one part of economics in which transaction costs have been used to explain a major feature of the economic system and that relates to the evolution and use of money. Adam Smith pointed out the hindrances to commerce that would arise in an economic system in which there was a division of labor but in which all exchange had to take the form of barter. No one would be able to buy anything unless he possessed something that the producer wanted. This difficulty, he explained, could be overcome by the use of money. A person wishing to buy something in a barter system has to find someone who has this product for sale but who also wants some of the goods possessed by the potential buyer. Similarly, a person wishing to sell something has to find someone who both wants what he has to offer and also possesses something that the potential seller wants. Exchange in a barter system requires what W. S. Jevons called “this double coincidence.” Clearly the search for partners in exchange with suitable qualifications is likely to be very costly and will prevent many potentially beneficial exchanges from taking place. The benefit brought about by the use of money consists of a reduction in transaction costs. The use of money also reduces transaction costs by facilitating the drawing up of contracts as well as by reducing the quantity of goods that need to be held for purposes of exchange. However, the nature of the benefits secured by the use of money seems to have faded into the background so far as economists are concerned, and it does not seem to have been noticed that there are other features of the economic system which exist because of the need to mitigate transaction costs.

I now turn to that other article cited by the Swedish Academy, “The Problem of Social Cost,” published some 30 years ago. I will not say much here about its influence on legal scholarship, which has been immense, but will mainly consider its influence on economics, which has not been immense, although I believe that in time it will be. It is my view that the approach used in that article will ultimately transform the structure of microeconomics—and I will explain why. I should add that in writing this article I had no such general aim in mind. I thought that I was exposing the weaknesses of Pigou’s analysis of the divergence between private and social products, an analysis generally accepted by economists, and that was all. It was only later and in part as a result of conversations with Steven Cheung in the 1960s that I came to see the general significance for economic theory of what I had written in that article and also to see more clearly what questions needed to be further investigated.

Pigou’s conclusion and that of most economists using standard economic theory was (and perhaps still is) that some kind of government action (usually the imposition of taxes) was required to restrain those whose actions had harmful effects on others (often termed negative externalities). What I showed in that article, as

I thought, was that in a regime of zero transaction costs, an assumption of standard economic theory, negotiations between the parties would lead to those arrangements being made which would maximize wealth and this irrespective of the initial assignment of rights. This is the infamous Coase theorem, named and formulated by George Stigler, although it is based on work of mine. Stigler argues that the Coase theorem follows from the standard assumptions of economic theory. Its logic cannot be questioned, only its domain (Stigler, 1989, pp. 631–3). I do not disagree with Stigler. However, I tend to regard the Coase theorem as a stepping stone on the way to an analysis of an economy with positive transaction costs. The significance to me of the Coase theorem is that it undermines the Pigovian system. Since standard economic theory assumes transaction costs to be zero, the Coase theorem demonstrates that the Pigovian solutions are unnecessary in these circumstances. Of course, it does not imply, when transaction costs are positive, that government actions (such as government operation, regulation, or taxation, including subsidies) could not produce a better result than relying on negotiations between individuals in the market. Whether this would be so could be discovered not by studying imaginary governments but what real governments actually do. My conclusion: let us study the world of positive transaction costs.

If we move from a regime of zero transaction costs to one of positive transaction costs, what becomes immediately clear is the crucial importance of the legal system in this new world. I explained in “The Problem of Social Cost” that what are traded on the market are not, as is often supposed by economists, physical entities, but the rights to perform certain actions, and the rights which individuals possess are established by the legal system. While we can imagine in the hypothetical world of zero transaction costs that the parties to an exchange would negotiate to change any provision of the law which prevents them from taking whatever steps are required to increase the value of production, in the real world of positive transaction costs, such a procedure would be extremely costly and would make unprofitable, even where it was allowed, a great deal of such contracting around the law. Because of this, the rights which individuals possess, with their duties and privileges, will be, to a large extent, what the law determines. As a result, the legal system will have a profound effect on the working of the economic system and may in certain respects be said to control it. It is obviously desirable that these rights should be assigned to those who can use them most productively and with incentives that lead them to do so and that, to discover (and maintain) such a distribution of rights, the costs of their transference should be low, through clarity in the law and by making the legal requirements for such transfers less onerous. Since this can come about only if there is an appropriate system of property rights (and they are enforced), it is easy to understand why so many academic lawyers (at least in the United States) have found so attractive the task of uncovering the character of such a property-rights system and why the subject of “law and economics” has flourished in American law schools. Indeed, work is going forward at such a pace that I do not consider it overoptimistic to believe that the main outlines of the subject will be drawn within 5 or 10 years.

Until quite recently, most economists seem to have been unaware of this relationship between the economic and legal systems except in the most general way. Stock and produce exchanges are often used by economists as examples of perfect or near-perfect competition. But these exchanges regulate in great detail the activities of traders (and this quite apart from any public regulation there may be). What can be traded, when it can be traded, the terms of settlement, and so on are all laid down by the authorities of the exchange. There is, in effect, a private law. Without such rules and regulations, the speedy conclusion of trades would not be possible. Of course, when trading takes place outside exchanges (and this is almost all trading) and where the dealers are scattered in space and have very divergent interests, as in retailing and wholesaling, such a private law would be difficult to establish, and their activities will be regulated by the laws of the State. It makes little sense for economists to discuss the process of exchange without specifying the institutional setting within which the trading takes place, since this affects the incentives to produce and the costs of transacting. I think this is now beginning to be recognized and has been made crystal-clear by what is going on in Eastern Europe today. The time has surely gone in which economists could analyze in great detail two individuals exchanging nuts for berries on the edge of the forest and then feel that their analysis of the process of exchange was complete, illuminating though this analysis may be in certain respects. The process of contracting needs to be studied in a real-world setting. We would then learn of the problems that are encountered and of how they are overcome, and we would certainly become aware of the richness of the institutional alternatives between which we have to choose.

Oliver Williamson has ascribed the non-use or limited use of my thesis in "The Nature of the Firm" to the fact that it has not been made "operational," by which he means that the concept of transaction costs has not been incorporated into a general theory. I think this is correct. There have been two reasons for this. First, incorporating transaction costs into standard economic theory, which has been based on the assumption that they are zero, would be very difficult, and economists who, like most scientists, as Thomas Kuhn has told us, are extremely conservative in their methods, have not been inclined to attempt it. Second, Williamson has also pointed out that although I was correct in making the choice between organization within the firm or through the market the centerpiece of my analysis, I did not indicate what the factors were that determined the outcome of this choice and thus made it difficult for others to build on what is often described as a "fundamental insight." This also is true. But the interrelationships which govern the mix of market and hierarchy, to use Williamson's terms, are extremely complex, and in our present state of ignorance it will not be easy to discover what these factors are. What we need is more empirical work. In a paper written for a conference of the National Bureau of Economic Research, I explained why I thought this was so. This is what I said: "An inspired theoretician might do as well without such empirical work, but my own feeling is that the inspiration is most likely to come through the stimulus provided by the patterns, puzzles, and anomalies revealed by the systematic gathering of data, particularly when the prime need is to break our existing habits of thought" (Coase, 1988, p. 71). This statement was made in 1970. I still think that in essentials

it is true today. Although much interesting and important research was done in the 1970s and 1980s and we certainly know much more than we did in 1970, there is little doubt that a great deal more empirical work is needed. However, I have come to the conclusion that the main obstacle faced by researchers in industrial organization is the lack of available data on contracts and the activities of firms. I have therefore decided to do something about it.

Believing that there is a great deal of data on contracts and the activities of firms in the United States available in government departments and agencies in Washington, DC, and that this information is largely unknown to economists, I organized a conference at the University of Chicago Law School in the summer of 1990 at which government officials presented papers in which they described what data was available and how to get access to it and also reported on some of the research being carried out within their departments. The audience consisted of academic economists. It was, as a colleague remarked, a case of supply meeting demand. The proceedings of this conference will be published in a special issue of the *Journal of Law and Economics*³. This availability of data and the encouragement given to all researchers working on the institutional structure of production by the award to me of the Nobel Prize should result in a reduction in that elegant but sterile theorizing so commonly found in the economics literature on industrial organization and should lead to studies which increase our understanding of how the real economic system works.

My remarks have sometimes been interpreted as implying that I am hostile to the mathematization of economic theory. This is untrue. Indeed, once we begin to uncover the real factors affecting the performance of the economic system, the complicated interrelations between them will clearly necessitate a mathematical treatment, as in the natural sciences, and economists like myself, who write in prose, will take their bow. May this period soon come.

I am very much aware that many economists whom I respect and admire will not agree with the opinions I have expressed, and some may even be offended by them. But a scholar must be content with the knowledge that what is false in what he says will soon be exposed and, as for what is true, he can count on ultimately seeing it accepted, if only he lives long enough.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Coase, R. H. (1988). *The firm, the market, and the law*. University of Chicago Press.
- Demsetz, H. (1988). *Ownership, control, and the firm* (Vol. I). Blackwell.
- Holmstrom, B., & Jean, T. (1989). The theory of the firm. In R. Schmalensee & R. D. Willig (Eds.), *Handbook of industrial organization* (pp. 61–128). North-Holland.

³ Amendment made by the author, R.H.C., June 2003.

- Peltzman, S. (1991). The handbook of industrial organization: A review article. *Journal of Political Economy*, 99, 201–217.
- Robbins, L. (1932). *The nature and significance of economic science*. Macmillan.
- Stigler, G. J. (1989). Two notes on the Coase theorem. *Yale Law Journal*, 99, 631–633.
- Williamson, O. E., & Winter, S. G. (Eds.). (1991). *The nature of the firm, origins, evolution, and development*. Oxford University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Transaction Cost Economics

4

Oliver E. Williamson

Contents

4.1	Introduction	48
4.2	The Lenses of Choice and Contract	49
4.2.1	Big Ideas	49
4.2.2	The Sciences of Choice and Contract	49
4.3	The Paradigm Problem: Vertical Integration	52
4.4	The Rudiments	53
4.4.1	Human Actors	53
4.4.2	Unit of Analysis	54
4.4.3	Main Purpose	54
4.4.4	Governance Structures	55
4.4.5	Predictions and Empirical Testing	58
4.5	Variations on a Theme	59
4.5.1	The Simple Contractual Schema	59
4.5.2	Applications	61
4.6	Public Policy	64
4.6.1	Antitrust	64
4.6.2	Regulation/Deregulation	65
4.6.3	Other Public Policy	65
4.7	Conclusions	66
	References	68

Oliver Williamson sadly passed away before this second edition of the *Handbook of New Institutional Economics* was created. His chapter from the first edition remains a classic and emblematic of his invaluable contributions to the field.

O. E. Williamson
University of California, Berkeley, CA, USA

Abstract

This chapter analyzes how transaction cost economics provides an understanding of complex economic organization by selectively joining law, economics, and organization theory. As against neoclassical economics, which is predominantly concerned with price and output, relies extensively on marginal analysis, and describes the firm as a production function (which is a technological construction), transaction cost economics (TCE) is concerned with the allocation of economic activity across alternative modes of organization (markets, firms, bureaus, etc.), employs discrete structural analysis, and describes the firm as a governance structure (which is an organizational construction). I begin by contrasting the lens of contract (out of which TCE works) with the lens of choice (orthodoxy). Vertical integration, which is the paradigm problem for TCE, is then examined. The operationalization of TCE is discussed in Sect. 4.4. Variations on a theme are sketched in Sect. 4.5. Public policy is discussed in Sect. 4.6. Concluding remarks follow.

Keywords

Transaction cost · Organizational economics · Contracting · Vertical integration · Governance

4.1 Introduction

Transaction cost economics is an effort to better understand complex economic organization by selectively joining law, economics, and organization theory. As against neoclassical economics, which is predominantly concerned with price and output, relies extensively on marginal analysis and describes the firm as a production function (which is a technological construction), transaction cost economics (TCE) is concerned with the allocation of economic activity across alternative modes of organization (markets, firms, bureaus, etc.), employs discrete structural analysis, and describes the firm as a governance structure (which is an organizational construction). Real differences notwithstanding, orthodoxy and TCE, are in many ways complements—one being more well-suited to aggregation in the context of simple market exchange and the other being more well-suited to the microanalytics of complex contracting and nonmarket organization.

I begin by contrasting the lens of contract (out of which TCE works) with the lens of choice (orthodoxy). Vertical integration, which is the paradigm problem for TCE, is then examined. The operationalization of TCE is discussed in Sect. 4.4. Variations on a theme are sketched in Sect. 4.5. Public policy is discussed in Sect. 4.6. Concluding remarks follow.

4.2 The Lenses of Choice and Contract¹

4.2.1 Big Ideas

Hal Varian has recently distinguished between important ideas and Big Ideas and describes Ronald Coase's classic paper, "The Nature of the Firm" (1937) as a Big Idea (2002, p. C2). Although there is widespread agreement on this, the nature of the big idea took a long time to register. Thus as of 1972, 35 years after the publication of "The Nature of the Firm," Coase described his 1937 article as "much cited and little used" (1972, p. 63). It was much cited because it was onto something important, perhaps even big. But it was little used because the big idea was only dimly perceived and/or lacked operationalization (Coase, 1992, pp. 716–718).

The essence of the Coasian contribution has been variously described (Williamson, 1994, p. 202; North, 2000, p. 37; Werin, 2000, p. 45). For the purposes of TCE, I contend that the overarching big idea was to move from choice to contract: bring the lens of contract systematically to bear on economic phenomena of all kinds. For many transactions, of which the make-or-buy decision is one (Coase, 1937), the contractual structure is easily recognized. Other transactions, such as the externality problem (Coase, 1960), needed to be reformulated to bring out their latent contractual features. The object, in these and other cases described herein, is to uncover previously neglected but, often, consequential features, the discovery of which often leads to different and, sometimes, deeper understandings than the orthodox lens of choice affords. If, as James Buchanan declares, "mutuality of advantage from voluntary exchange is . . . the most fundamental of all understandings in economics" (2001, p. 29), then at least some of us should be thinking of economics as the "science of exchanges" (Buchanan, 2001, p. 28).²

4.2.2 The Sciences of Choice and Contract

Economics throughout the twentieth century has been developed predominantly as a science of choice. As Lionel Robbins famously put it in his book, *The Nature and Significance of Economic Science*, "Economics is the science which studies human behavior as a relationship between ends and scarce means which have alternative uses" (1932, p. 16). Choice has been developed in two parallel constructions: the theory of consumer behavior, in which consumers maximize utility, and the theory of

¹This Subsection Is Based on Williamson, 2002b

²Students of the history of thought will remind us that cattallactics—meaning "the science of exchanges"—has much earlier origins. Indeed, a book by E. B. de Condillac on this subject was published in 1776, which is when *The Wealth of Nations* first appeared (see Murray Rothband (1987, pp. 377–378) for an historical sketch). Recurrent interest in the science of contract notwithstanding, it has operated in the shadows of the science of choice. Why the disparity? Here as elsewhere, good ideas need to be operationalized. Contractual analysis has gotten under way in a sustained way only during the past 40 years.

the firm as a production function, in which firms maximize profit. Economists who work out of such setups emphasize how quantities are influenced by changes in relative prices and available resources, a project which became the “dominant paradigm” for economics throughout the twentieth century (Reder, 1999, p. 48).

But the science of choice is not the only lens for studying complex economic phenomena, nor is it always the most instructive lens. The other main but less fully developed approach is the science of contract. Indeed, Buchanan (1975, p. 225) avers that economics as a discipline went “wrong” in its preoccupation with the science of choice and the optimization apparatus associated therewith. What was needed was the parallel development of a science of contract. Awaiting this, some phenomena would go unnoticed, others would be poorly understood, and public policy error would result.

As perceived by Buchanan, the principal needs for a science of contract were to the field of public finance and took the form of public ordering: “Politics is a structure of complex exchange among individuals, a structure within which persons seek to secure collectively their own privately defined objectives that cannot be efficiently secured through *simple market exchanges*” (1987, p. 296; emphasis added). Thinking contractually in the public ordering domain leads into a focus on the rules of the game. Issues of a constitutional economics kind are posed (Buchanan & Tullock, 1962; Brennan & Buchanan, 1985).

Whatever the rules of the game, the lens of contract is also usefully brought to bear on the play of the game. This latter is what I refer to as *private ordering*, which entails self-help efforts by the immediate parties to a transaction to align incentives and craft governance structures that are better attuned to their exchange needs. John R. Commons’ prescient statement on the nature of the economic problem provides the unifying theme: “the ultimate unit of activity. . . must contain in itself the three principles of conflict, mutuality, and order. This unit is the transaction” (1932, p. 4).³ Not only does transaction cost economics take the transaction to be the basic unit of analysis, but governance is the means by which to infuse *order*, thereby to mitigate *conflict* and realize *mutual gain*.

Although market competition serves these governance purposes in the context of the “simple market exchanges” to which Buchanan made reference (which is wholly in the spirit of orthodox price theory), transaction cost economics is predominantly concerned with complex market exchange where there are small numbers of parties on each side of the transaction. Rather than examine such issues with the price-theoretic apparatus of oligopoly or oligopsony, transaction cost economics focuses instead on uncovering and explicating the *strategic hazards* that are posed by small numbers exchange in the context of incomplete contracting and the cost-effective deployment of governance to mitigate these hazards. Strategic issues that had been

³Not everyone associated with the lens of contract would agree. Coase, for example, contends that “American institutionalism,” of which Commons was a prominent part, “is a dreary subject. All it had was a stance of hostility to the standard economic theory. It certainly led to nothing” (1984, pp. 229–230). My view is that Commons was ahead of his time. He had a lens of contract conception of economics as early as the 1920s.

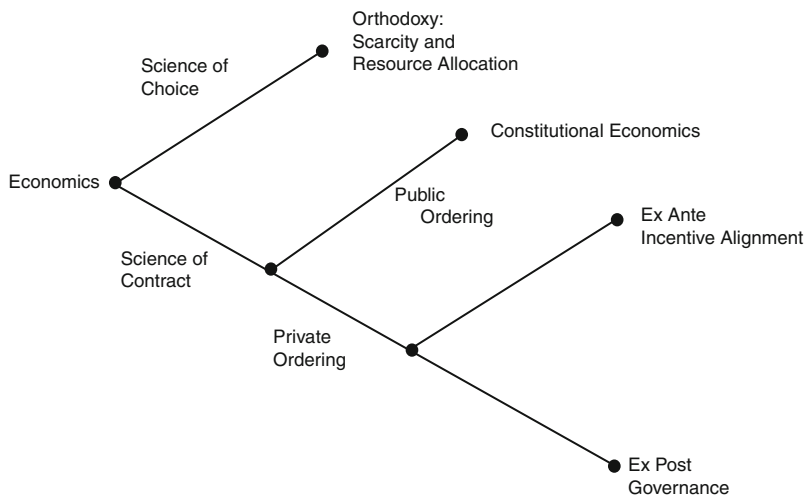


Fig. 4.1 The sciences of choice and contract

ignored by neoclassical economists from 1870 to 1970 now make their appearance (Makowski & Ostroy, 2001, pp. 482–483, 490–491).

Figure 4.1 sets out the main distinctions. The initial divide is between the science of choice (orthodoxy) and the science of contract. The latter then divides into public (constitutional economics) and private ordering parts, where the second is split into two related branches. One branch deals with ex ante incentive alignment (mechanism design, agency theory, the formal property rights literature) while the second features the ex post governance of contractual relations (contract implementation). Albeit related, these two are in tension. Thus whereas transaction cost economics locates the main analytical action in the ex post stage of contract (where maladaptation problems appear), the formal incentive alignment literature annihilates ex post governance.⁴ One device for accomplishing this is to assume common knowledge of payoffs and costless bargaining.

The use of strong assumptions (of which common knowledge of payoffs and costless bargaining are two) to strip away inessentials and get at the essence is, to be

⁴Contract theorists who concentrate the analytical action on the ex ante incentive alignment stage of contracting might complain that TCE makes strong assumptions also, the effect of which is to annihilate the ex ante incentive alignment stage. For example, TCE assumes that contracting parties in intermediate product markets are risk neutral, whence efficient risk bearing plays no role in incentive alignment. Contract theorists who rely on risk aversion for their main results might protest against risk neutrality. Be that as it may, TCE also assumes that contracting parties look ahead, recognize consequential contractual hazards that arise during contract implementation, and factor these into the ex ante contractual design—by pricing out unrelieved contractual hazards and by introducing credible commitments (in cost effective degree)—so ex ante and ex post stages of contract are definitely joined. What TCE disallows are assumptions which vaporize maladaptation and strategizing during contract implementation—of which common knowledge of payoffs and costless bargaining are two.

sure, vital to the scientific enterprise. Simplifications, however, that lose contact with core issues are deeply problematic: “A model can be right in...[a]mechanical sense...[yet be] unenlightening because...[it is]imperfectly suited to the subject matter. It can obscure the key interactions, instead of spotlighting them” (Solow, 2001, p. 112). In the degree to which the core issues that are posed by contractual incompleteness are those of maladaptation, formalizations that preserve rather than annihilate *ex post* governance are needed (Bajari & Tadelis, 2001).

4.3 The Paradigm Problem: Vertical Integration

Contract is an encompassing concept. Rather than treat the issue in its full generality, it will be instructive to begin with a specific puzzle of economic organization, ideally one for which other contractual issues will turn out to be variations on a theme.

The obvious transaction with which to begin is that of vertical integration or, in more mundane terms, the make-or-buy decision. Not only is the make-or-buy decision the transaction on which Coase focused in 1937, but it has a prior and continuing history of importance within economics. Examining the intermediate product market transactions (within and between firms) also has an advantage in that it relieves many of the asymmetry conditions—of information, budget, legal talent, risk aversion, and the like—that complicate the analysis of transactions in final product markets.⁵

Coase’s classic article opens with a basic puzzle: Why does a firm emerge at all in a specialized exchange economy? If the answer resides in entrepreneurship, why is coordination “the work of the price mechanism in one case and the entrepreneur in the other” (Coase, 1937, p. 389)? Coase appealed to transaction cost economizing as the hitherto missing factor for explaining why markets were used in some cases and hierarchy in other cases and averred that “The main reason why it is profitable to establish a firm would seem to be that there is a cost of using the price mechanism, the most obvious...[being] that of discovering what the relevant prices are” (1937, p. 391). That sounds plausible but is it truly comparative? How is it that internal procurement by the firm avoids the cost of price discovery?

The “obvious” answer is that sole-source internal supply avoids the need to consult the market about prices because internal accounting prices of a formulaic kind (say, of a cost-plus kind) can be used to transfer a good or service from one internal stage to another. If, however, that is the source of the advantage of internal organization over market procurement, the obvious lesson is to apply this same practice to outside procurement. The firm simply advises its purchasing office to turn a blind eye to the

⁵Final product market transactions and transactions between suppliers and distributors need to be distinguished. The former refer to transactions between firms (as suppliers) and final consumers (buyers). Serious asymmetry conditions for which consumer protections are sometimes warranted arise for final product market transactions. By contrast, the transactions between the manufacturer and the distributor are between firms. Thus, the manufacturer can sell outright to distributors (which is a market transaction). Or the manufacturer can integrate forward into distribution (so both stages are under unified ownership/hierarchy). Or the manufacturer can create franchisees (which are a hybrid mode of contracting).

market by placing orders, period by period, with a qualified sole-source external supplier who agrees to sell on cost-plus terms. In that event, firm and market are put on a parity in price discovery respects—which is to say that the price discovery burden that Coase ascribes to the market does not survive comparative institutional scrutiny.

Even, however, if price discovery did survive comparative institutional scrutiny, that seems to be a thin basis upon which to rest the case for using firms rather than markets if, as I contend, firm and market differ in kind rather than in degree. What economic purposes are served by the discrete structural changes that distinguish market and hierarchy? Does the move from choice (where prices are focal) to contract implicate other, possibly more basic, considerations? What rudiments inform the logic of contract and comparative economic organization?

4.4 The Rudiments

This last invites the student of economic organization to step back and address contract “on its own terms.” What are the attributes of human actors that bear on the efficacy of contract? What unit of analysis should be employed? Of the many purposes of contract, which are salient? How are alternative modes of governance described? What refutable implications accrue upon reformulating the problem of economic organization in comparative contractual terms? Are the data corroborative? What are the public policy ramifications?

4.4.1 Human Actors

If “Nothing is more fundamental in setting our research agenda and informing our research methods than our view of the nature of the human beings whose behavior we are studying” (Simon, 1985, p. 303), then economists and other social scientists are well-advised to describe the key cognitive, self-interest, and other attributes of human actors on which their analyses rest. Simon’s view of cognition is that the usual hyper-rationality assumption be supplanted by bounded rationality—behavior that is “intendedly rational but only limitedly so” (1957, p. xxiv). He further recommends that self-interest be described as “frailty of motive” (1985, p. 305). TCE concurs that bounded rationality is the appropriate cognitive assumption and takes the chief lesson of bounded rationality for the study of contract to be that *all complex contracts are unavoidably incomplete*. But TCE also takes a further step, which takes exception with the common view that bounded rationality implies that human actors are myopic. As against myopia, human actors are assumed to have the capacity to look ahead, uncover possible contractual hazards, and work out the contractual ramifications (Shultz, 1995).

TCE also pushes beyond frailty of motive to make provision for opportunism. This latter does not deny that most people will do what they say and some will do more most of the time. Opportunism, however, has reference to exceptions—outliers where there is a lot at stake and parties are often observed to defect from the spirit of cooperation to insist on the letter of the incomplete contract. Strategic considerations are introduced upon making provision for opportunism.

4.4.2 Unit of Analysis

The natural unit of analysis for lens of contract purposes is the transaction. Naming a unit of analysis is always much easier, however, than identifying the critical dimensions for describing the unit of analysis—as witness the fact that the key attributes for so many would-be units of analysis are never identified.⁶ Awaiting dimensionalization, transaction cost economics remained a largely tautological construction.

To be sure, transactions can be variously described—depending on the purpose. Transaction cost economics holds that three dimensions that have pervasive ramifications for governance are asset specificity (which takes a variety of forms—physical, human, site, dedicated, brand name—and is a measure of non-redeployability), the disturbances to which transactions are subject (and to which potential maladaptations accrue), and the frequency with which transactions recur (which bears both on the efficacy of reputation effects in the market and the incentive to incur the cost of specialized internal governance). The absence of asset specificity describes the ideal transaction in law and economics for which competition works well: “sharp in by clear agreement; sharp out by clear performance” (Macneil, 1974, p. 734). As asset specificity builds up, however, bilateral dependency develops and, in combination with uncertainty (which pushes incomplete contracts out of alignment), the aforementioned contractual complications set in.

4.4.3 Main Purpose

Interestingly, both the economist Friedrich Hayek (1945) and the organization theorist Chester Barnard (1938) were in agreement that adaptation is the central problem of economic organization. Hayek (pp. 526–527) focused on the adaptations of economic actors who adjust spontaneously to changes in the market, mainly as signaled by changes in relative prices: upon looking “at the price system as . . . a mechanism for communicating information,” the marvel of the market resides in “how little the individual participants need to know to be able to take the right action.” By contrast, Barnard featured coordinated adaptation among economic actors working through administration (hierarchy). The latter is accomplished not spontaneously but in a “conscious, deliberate, purposeful” way (p. 9) and comes into play when the simple market exchanges on which Hayek focused break down.

In effect, the adaptations to which Hayek refers are *autonomous* adaptations in which individual parties respond to market opportunities as signaled by changes in relative prices whereas the adaptations of concern to Barnard are *cooperative* adaptations accomplished through administration within the firm. Because a high performance economic system will display adaptive capacities of both kinds, an

⁶Examples of would-be units of analysis for which operational content is missing include the role (see Simon’s critique (1957, p. xxx), the decision-premise (which is Simon’s candidate, but which has found little application outside of cognitive psychology (Newell & Simon, 1972)), and the routine (Nelson & Winter, 1982). The next two paragraphs and Sect. 4.5 are based on Williamson (2002a).

understanding and appreciation for both markets *and* hierarchies (rather than the comparative economic systems dichotomy between markets *or* hierarchies) is needed. The firm for these purposes is described not as a production function (which is a technological construction) but as a governance structure (which is an organizational construction). And the market is described similarly. The lens of contract, as against the lens of choice, is made the cutting edge.

One of the advantages of focusing on *adaptation as the main case* is that it brings added meaning to the idea of mutual gain. It is elementary that gains from trade will always be realized by moving onto the contract curve. But how is this to be accomplished in a world where complex contracts are incomplete and are implemented over time in the face of disturbances for which contingent provisions either have not been made or, if made, are often in error? Crafting governance structures that *are attuned to the hazards and help the parties to restore efficiency* (return to the shifting contract curve) where otherwise a costly impasse would develop is needed in these circumstances. More attention to designing processes that have good adaptive properties (and less to concentrating all of the action in the ex ante incentive alignment stage) is thus one of the central lessons of TCE.

4.4.4 Governance Structures

Examining economic organization through the lens of contract both places the spotlight on ex post adaptation and, in the process, gives prominence to the role of governance. Specifically, TCE holds that each generic mode of governance is defined by a syndrome of internally consistent attributes to which different adaptive strengths and weaknesses accrue.

The three attributes of principal importance for describing governance structures are (1) incentive intensity, (2) administrative controls, and (3) contract law regime. Spot markets and hierarchy differ with respect to these attributes as follows: the high-powered incentives of markets are supplanted by low-powered incentives when transactions are organized within firms; market exchange is a hands-off control mechanism whereas hierarchy involves considerable hands on administrative involvement; and whereas disputes in markets are treated in a legalistic way and rely on court ordering, courts refuse to hear (most) internal disputes, whereupon the firm becomes its own court of ultimate appeal. Firms have access to fiat that markets do not because of these dispute resolution differences.⁷

Governance, moreover, is an encompassing concept. Going beyond polar forms, all modes of organization within which (or with the support of which) transactions are managed come under scrutiny. Hybrid modes of contracting to which credible

⁷One of the reasons why markets lack fiat is that attempts to award decision rights to autonomous agents by contract (e.g., A will decide disputed matters of type X; B will decide disputed matters of type Y; etc.) are commonly unenforceable. That is because one of the parties to a market contract can invoke (invent) a "technicality," the effect of which is to bring the dispute before the courts.

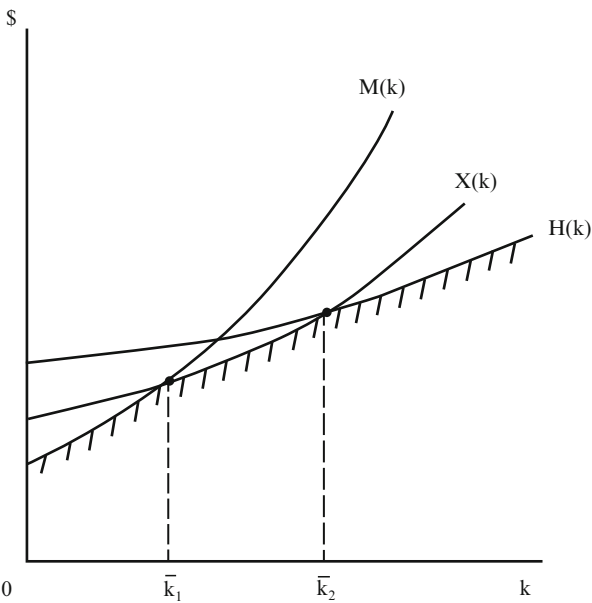
commitment supports have been crafted (penalties against premature termination, specialized dispute settlement mechanisms and the like) are especially important. Table 4.1 summarizes the key attributes of (spot) market, hybrids, and hierarchies. As developed in the text accompanying Figs. 4.2 and 4.3, the clusters of attributes that define these three alternative modes of governance, give rise to differential transaction costs among modes, conditional on the attributes of the transactions to be organized.

An unremarked governance complication also needs to be introduced. This is that organization, like the law, has a life of its own (Selznick, 1950, p. 10). Issues of internal organization that have been featured by organization theorists (Barnard, 1938; Simon, 1947; March & Simon, 1958) and by sociologists (Michels, 1962; Merton, 1936; Selznick, 1949; Scott, 1992) thus arise. Note in this connection that while it is relatively easy to show that internal organization is subject to incentive limits and bureaucratic distortions, it is much more difficult to show that

Table 4.1 Attributes of leading generic modes of governance

<i>Governance modes</i>			
Governance attributes	Market	Hybrid	Hierarchy
Incentives	High-powered	Less high-powered	Low-powered
Administrative support by bureaucracy	Nil	Some	Much
Contract law regime	Legalistic	Contract as framework	Firm as own court of ultimate appeal (fiat)

Fig. 4.2 Transaction costs and asset specificity



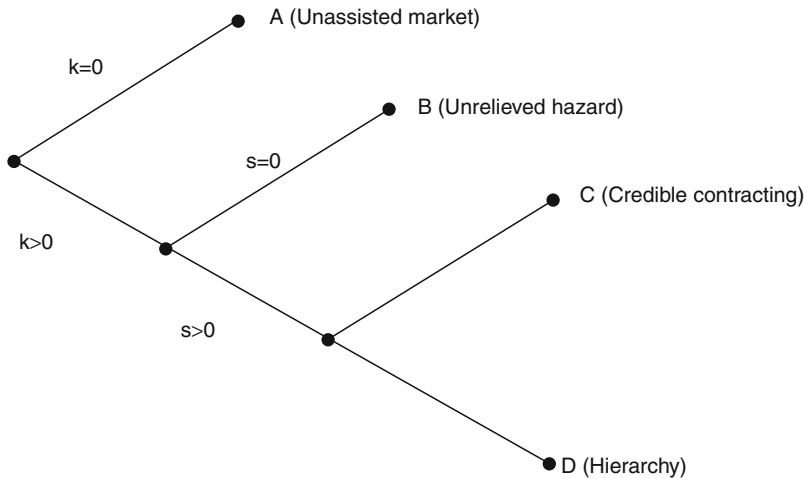


Fig. 4.3 Simple contracting schema

comparative cost consequences accrue upon taking a transaction out of the market and organizing it internally (Williamson, 1985, Chap. 6). That is because the goods and services traded in a market are produced within firms. The question of make-or-buy is thus whether the costs (including bureaucratic costs) are greater in two autonomous firms than in one combined entity.

TCE uncovers and explicates the incentive and bureaucratic cost consequences that attend the move from market to hierarchy by postulating two processes: replication and selective intervention. Were it that the firm could replicate the market in all circumstances where market procurement works well and intervene selectively (if expected net gains can be projected) where markets break down, then the firm could never do worse than the market (through replication) and would sometimes do better (through selective intervention). As it turns out, it is impossible to realize this ambition. Incentives are unavoidably compromised and added bureaucratic costs are unavoidably incurred upon taking a transaction out of the market and organizing it internally.⁸ The upshot is that the move from market to hierarchy is attended by tradeoffs. Discriminating alignment is thus needed.

⁸The issues here are rather involved. The interested reader is referred to Williamson (1985, Chap. 6) for a discussion. Briefly the argument is that (1) replication is essential if parity between firm and market is to be assumed, (2) such replication implies that incentives are unchanged in each of the separable stages upon taking a transaction out of the market and organizing it internally, yet (3) unchanged incentive intensity cannot be accomplished if the acquiring stage exercises control over the accounting system (to include transfer prices, overhead rates, depreciation, and the like) and cannot credibly commit to behaving non-strategically, to include intervening always but only for good cause (selective intervention). But there is more. Not only is incentive intensity unavoidably weakened when transactions move from market to hierarchy, but cooperative adaptation across successive stages is promoted by intentionally weakening incentive intensity within the firm.

4.4.5 Predictions and Empirical Testing

The main engine from which the predictions of TCE are derived is that of discriminating alignment, according to which transactions, which differ in their attributes, are aligned with governance structures, which differ in their cost and (adaptive) competence, so as to affect a transaction cost economizing result. The upshot is that there is a place for each generic mode of governance, yet each should be kept in its place. The TCE answer to the Coasian puzzle of which transactions go where and why resides precisely in discriminating alignment, the efficacy of which relies in part on weak-form natural selection (Simon, 1983, p. 69) to penalize errors of inefficient alignment.

It will be convenient here to focus on three modes: spot markets (M), hybrid modes of contracting (X) into which credible commitments have been introduced, and hierarchies (H). The basic argument is that (1) markets are well-suited to making autonomous adaptations, firms enjoy the advantage for comparative adaptation purposes, and hybrids are located in between, (2) the needs for adaptation vary among transactions (especially with reference to asset specificity), and (3) bureaucratic cost burdens increase as transactions move from market, to hybrid, to hierarchy.

In a heuristic way, the transaction cost consequences of organizing transactions in markets (M) and hierarchies (H) as a function of asset specificity (k) are shown in Fig. 4.2. As shown, the bureaucratic burdens of hierarchy place it at an initial disadvantage ($k = 0$), but the cost differences between $M(k)$ and $H(k)$ narrow as asset specificity builds up and eventually reverse as the need for cooperative adaptation becomes especially great ($k \gg 0$). As indicated, moreover, the hybrid mode of organization $X(k)$ is viewed as a market-preserving credible contracting mode that possesses adaptive attributes located between classical markets and hierarchies. Incentive intensity and administrative control thus take on intermediate values and Karl Llewellyn's (1931) concept of contract as framework applies. As shown in Fig. 4.2, $M(0) < X(0) < H(0)$ (by reason of bureaucratic cost differences) while $M' > X' > H'$ (which reflects the differential ability of these three modes to implement coordinated adaptation, the needs for which increase as asset specificity builds up).⁹ The least cost mode of governance is thus the market for $k < \bar{k}_1$, the hybrid for $\bar{k}_1 < k < \bar{k}_2$, and hierarchy for $k > \bar{k}_2$.

Whereas many theories of vertical integration do not invite empirical testing, the transaction cost theory of vertical integration invites and has been the subject of considerable empirical analysis. Both the theory of the firm (Holmstrom & Tirole, 1989, p. 126) and the field of industrial organization (Peltzman, 1991) have been criticized for lack of empirical testing; yet empirical research in transaction cost economics has grown exponentially during the past 20 years. (For surveys, see Shelanski and Klein (1995), Lyons (1996), Crocker and Masten (1996), Rindfleisch

⁹ M' , X' , and H' refer to the marginal costs of market, hybrid, and hierarchical governance with respect to changes in asset specificity (k).

and Heide (1997), Masten and Saussier (2000) and Boerner and Macher (2002).) Added to this are numerous applications to public policy, especially antitrust and regulation, but also to economics more generally (Dixit, 1996) and to the contiguous social sciences (especially political science). The upshot is that the theory of the firm as governance structure has become a “much used” construction.¹⁰

4.5 Variations on a Theme

Vertical integration turns out to be a paradigm. Thus, although many of the empirical tests and public policy applications have reference to the make-or-buy decision and vertical market restrictions, this same conceptual framework has application to contracting more generally. Specifically, the contractual relation between the firm and its “stakeholders”—customers, suppliers, and workers along with financial investors—turn out to be variations on the theme set out in the simple contractual schema.

4.5.1 The Simple Contractual Schema

Assume that a firm can make or buy a component and assume further that the component can be supplied by either a general purpose technology or a special purpose technology. Again, let k be a measure of asset specificity. The transactions in Fig. 4.3 that use the general purpose technology are ones for which $k = 0$. In this case, no specific assets are involved and the parties are essentially faceless. Those transactions that use the special purpose technology are ones for which $k > 0$. As earlier discussed, bilaterally dependent parties have incentives to promote continuity and safeguard their specific investments. Let s denote the magnitude of any such safeguards, which include penalties, information disclosure and verification

¹⁰Reprints of leading articles on vertical integration (by Benjamin Klein on Fisher Body-GM (1988), Paul Joskow on long-term coal contracting (1988), Kirk Monteverde and David Teece on automobile integration (1982), Scott Masten on aerospace production (1984), Erin Anderson and David Schmittlein on sales force organization (1984), George John and Barton Weitz on forward integration into distribution (1988), and Scott Masten, James Meehan, and Edward Snyder on the cost of organization (1991)) can be found in Williamson and Masten, Vol. II (1995). This same volume includes empirical contracting articles (by Thomas Palay on rail freight contracting (1984), Victor Goldberg and John Erickson on long term petroleum coke contracts (1987), Paul Joskow on contract duration (1987), Harold Mulherin on natural gas contracting (1986), Scott Masten and Keith Crocker on take-or-pay provisions (1985), Keith Leffler and Randal Rucker on timber (1991), and Keith Crocker and Scott Masten on the long-term contracting process (1991). Empirical studies of regulation and positive political economy include Oliver Williamson (1976), Victor Goldberg (1976), George Priest (1993), Brian Levy and Pablo Spiller (1994), Barry Weingast and William Marshall (1988), and Rafael Gely and Pablo Spiller (1990). Antitrust applications include Oliver Williamson (1979), Roy Kenney and Benjamin Klein (1983), and Scott Masten and Edward Snyder (1993).

procedures, specialized dispute resolution (such as arbitration), and, in the limit, integration of the two stages under unified ownership. An $s = 0$ condition is one for which no safeguards are provided; a decision to provide safeguards is reflected by an $s > 0$ result.

Node A in Fig. 4.3 corresponds to the ideal transaction in law and economics: there being an absence of dependency, governance is accomplished through competitive market prices and, in the event of disputes, by court awarded damages. Node B poses unrelieved contractual hazards, in that specialized investments are exposed ($k > 0$) for which no safeguards ($s = 0$) have been provided. Such hazards will be recognized by farsighted players, who will price out the implied risks of contractual breakdown.

Added contractual supports ($s > 0$) are provided at nodes C and D. At node C, these contractual supports take the form of interfirm contractual safeguards. Should, however, costly contractual breakdowns continue in the face of best bilateral efforts to craft safeguards at node C, the transaction may be taken out of the market and organized under unified ownership (vertical integration) instead. Because added bureaucratic costs accrue upon taking a transaction out of the market and organizing it internally, internal organization is usefully thought of as the organization form of last resort: try markets, try hybrids, and have recourse to the firm only when all else fails. Node D, the unified firm, thus comes in only as higher degrees of asset specificity and added uncertainty pose greater needs for cooperative adaptation.

Note that the price that a supplier will bid to supply under node C conditions will be less than the price that will be bid at node B. That is because the added security features serve to reduce the risk at node C, as compared with node B, so the contractual hazard premium will be reduced. One implication is that suppliers do not need to petition buyers to provide safeguards. Because buyers will receive product on better terms (lower price) when added security is provided, buyers have the incentive to offer cost-effective credible commitments. Also note that whereas such commitments are sometimes thought of as a user-friendly way to contract, the analytical action resides in the hard-headed use of credibility to support those transactions where asset specificity and contractual hazards are at issue. Such supports are without purpose for transactions where generic technologies are employed.

The foregoing schema can be applied to virtually all transactions for which the firm is in a position to own as well as to contract with an adjacent stage—backward into raw materials, laterally into components, forward into distribution. But for some activities, ownership is either impossible or very rare.¹¹ For example, firms cannot own their workers or their final customers (although worker cooperatives and consumer cooperatives can be thought of in ownership terms). Also, firms rarely own their suppliers of finance. Node D drops out of the schema in cases where ownership is

¹¹ Closely complementary technologies are commonly relegated to the “core technology” (Thompson, 1967, pp. 19–23) and are effectively exempt from comparative institutional analysis, it being “obvious” that these are done within the firm.

either prohibited by law or is otherwise rare. I begin with forward integration into distribution, after which relationships with other stakeholders of the firm, including labor, finance, and public utility regulation are successively considered.¹²

4.5.2 Applications

4.5.2.1 Forward Integration into Distribution

I will set aside the case where mass marketers integrate backward into manufacturing and focus on forward integration into distribution by manufacturers of products or owners of brands. Specifically, consider the contractual relation between a manufacturer and large numbers of wholesalers and, especially, of retailers for the good or service in question.

Many such transactions are of a generic kind. Although branded goods and services are more specific, some require only shelf space, since advertising, promotion, and any warranties are done by the manufacturer. Since the obvious way to trade with intermediaries for such transactions is through the market, in a node A fashion, what is to be inferred when such transactions are made subject to vertical market restrictions—such as customer and territorial restrictions, service restrictions, tied sales, and the like?

Price discrimination, to which allocative efficiency benefits were often ascribed, was the usual price theoretic (science of choice) explanation for such restrictions. Such efficiency claims, however, are problematic once the transaction costs of discovering customer valuations and deterring arbitrage are taken into account (Williamson, 1975, pp. 11–13). Not only are the benefits problematic, but price discrimination is a needlessly narrow interpretation.

Viewed through the lens of contract, vertical market restrictions often have the purpose and effect of infusing order into a transaction where the interests of the system and the interests of the parts are otherwise in conflict. For example, the Schwinn Bicycle Company imposed non-resale restrictions upon franchisees. The concern was that the integrity of the brand, which was a system asset, would be compromised by franchisees who perceived local opportunities to realize individual gain by selling to discounters, who would then sell a “bike in a box,” without service or support (Williamson, 1985, pp. 183–189). More generally, the argument is this: in circumstances where market power is small, where simple market exchange (at node A) would compromise the integrity of differentiated products, and where

¹²Natural monopoly and government bureaus can be interpreted in terms of the schema in Fig. 4.3 as follows: First, given natural monopoly, the three “evils” to which Milton Friedman (1962, p. 128) referred—unregulated monopoly, regulation, and nationalization—correspond, roughly, to nodes B, C, and D, respectively. Also, for a good or service for which the government is the buyer, nodes A, B, and C are all market nodes (spot market, unrelieved hazard, and long-term (often cost-plus) contracting, respectively), while node D is the government bureau deciding to do the task itself. The government bureau has especially low-powered incentives, is highly bureaucratized (by design), and has its own dispute settlement machinery (Williamson, 1999).

forward integration into distribution (at node D) would be especially costly, the use of vertical market restrictions to effect credible commitments (at node C) has much to recommend it.

4.5.2.2 Relationship with Labor

Because the firm is unable to own its labor, node D is irrelevant and the comparison comes down to nodes A, B, and C. Node A corresponds to the case where labor is easily redeployed to other uses or users without loss of productive value ($k = 0$). Thus, although such labor may be highly skilled (as with many professionals), the lack of firm-specificity means that, transition costs aside, neither worker nor firm has an interest in crafting penalties for unwanted quits/terminations or otherwise creating costly internal labor markets (ports of entry; promotion ladders), costly information disclosure and verification procedures, and costly firm-specific dispute settlement machinery. The mutual benefits simply do not warrant the costs.

Conditions change when $k > 0$, since workers who acquire firm-specific skills will lose value if prematurely terminated (and firms will incur added training costs if such employees quit). Here, as elsewhere, unrelieved hazards (as at node B) invite governance responses to which mutual gains accrue. Because continuity has value to both firm and worker, governance features that deter termination (severance pay) and quits (nonvested benefits) and which address and settle disputes in an orderly way (grievance systems) to which the parties ascribe confidence have a lot to recommend them. These can, but need not, take the form of “unions.” Whatever the name, the object is to craft a collective organizational structure (at node C) in which the parties have mutual confidence and that enhances efficiency (Baron & Kreps, 1999, pp. 130–138; Williamson, 1975, pp. 27–80, 1985, pp. 250–262).¹³

4.5.2.3 Relationship with Sources of Finance

Viewed through the lens of contract, the board of directors is interpreted as a security feature that arises in support of the contract for equity finance (Williamson, 1988). More generally, debt and equity are not merely alternative modes of finance, which is the law and economics construction (Easterbrook & Fischel, 1986; Posner, 1986) but are also alternative modes of governance.

Suppose that a firm is seeking cost-effective finance for the following series of projects: general-purpose, mobile equipment; a general-purpose office building located in a population center; a general-purpose plant located in a manufacturing center; distribution facilities located somewhat more remotely; special-purpose equipment; market and product development expenses; and the like. Suppose further that debt is a governance structure that works almost entirely out of a set of rules:

¹³ The emphasis on collective organization as a governance response is to be distinguished from the earlier work of Gary Becker, where human asset specificity is responsible for upward sloping age-earnings profiles (1962). Becker's treatment is more in the science of choice tradition whereas mine views asset specificity through the lens of contract. These two are not mutually exclusive. They do, however, invite different public policy interpretations and redirect the empirical research agenda.

(1) stipulated interest payments will be made at regular intervals; (2) the business will continuously meet certain liquidity tests; (3) principal will be repaid at the loan-expiration date; and (4) in the event of default, the debtholders will exercise preemptive claims against the assets in question. In short, debt is unforgiving if things go poorly.

Such rules-based governance is well-suited to investments of a generic kind ($k = 0$), since the lender can redeploy these to alternative uses and users with little loss of productive value. Debt thus corresponds to market governance at node A. But what about investment projects of more specific (less redeployable) kinds?

Because the value of holding a preemptive claim declines as the degree of asset specificity deepens, rule-based finance of the above-described kind will be made on more adverse terms. In effect, using debt to finance such projects would locate the parties at node B, where a hazard premium must be charged. The firm in these circumstances has two choices: sacrifice some of the specialized investment features in favor of greater redeployability (move back to node A) or embed the specialized investment in a governance structure to which better terms of finance will be ascribed. What would the latter entail?

Suppose that a financial instrument called equity is invented and assume that equity has the following governance properties: (1) it bears a residual-claimant status to the firm in both earnings and asset-liquidation respects; (2) it contracts for the duration of the life of the firm; and (3) a board of directors is created and awarded to equity that (a) is elected by the pro-rata votes of those who hold tradeable shares, (b) has the power to replace the management, (c) decides on management compensation, (d) has access to internal performance measures on a timely basis, (e) can authorize audits in depth for special follow-up purposes, (f) is apprised of important investment and operating proposals before they are implemented, and (g) in other respects bears a decision-review and monitoring relation to the firm's management (Fama & Jensen, 1983).¹⁴ So construed, the board of directors is awarded to the holders of equity so as to reduce the cost of capital by providing safeguards for projects that have limited redeployability (by moving shareholders from node B to node C).

4.5.2.4 Regulation and Natural Monopoly

The market-oriented approach to natural monopoly is to auction off the franchise to the highest bidder (Demsetz, 1968; Posner, 1972). But while this is an imaginative proposal, it is not an all-purpose construction. Viewed through the lens of contract, whether this works well or poorly depends on the nature of the transaction and the particulars of governance. The action, once again, resides in the details (Williamson, 1976)—although others counsel that to “expound the details of particular regulations and proposals. . . would serve only to obscure the basic issues” (Posner, 1972, p. 98).

¹⁴ It will not go unnoticed that this is a rather normative way to describe the board of directors. In practice, many boards are rubber stamps to the management (in exchange for handsome fees). Takeover by tender offer is important for precisely because it is a means by which to replace a protective/complacent/compliant board.

Going beyond the initial bidding competition (“competition for the market”), the governance approach insists upon examining the contract implementation stage. Transactions to which the Fundamental Transformation applies—namely, those requiring significant investments in specific assets and that are subject to considerable market and technological uncertainty—are ones for which the efficacy of simple franchise bidding is problematic. If what had been a large numbers franchise bidding competition at the outset becomes, in effect, a small number supply relation during contract implementation and at the contract renewal interval, then the purported efficiency of franchise bidding is problematic.

This is not to say that franchise bidding never works. Neither is it to suggest that decisions to regulate ought not to be revisited—as witness the successful deregulation of trucking (which never should have been regulated to begin with) and more recent efforts to deregulate “network industries” (Peltzman & Whinston, 2000). I would nevertheless urge that examining deregulation through the lens of contracting is instructive for both—as it is for assessing efforts to deregulate electricity in California, where too much deference was given to the (assumed) efficacy of smoothly functioning markets and insufficient attention to potential investment and contractual hazards and appropriate governance responses thereto. As Joskow (2000, p. 51) observes: “Many policy makers and fellow travelers have been surprised by how difficult it has been to create wholesale electricity markets Had policy makers viewed the restructuring challenge using a TCE framework, these potential problems are more likely to have been identified and mechanisms adopted *ex ante* to fix them.”

Here as elsewhere, the lesson is to think contractually: look ahead, recognize potential hazards, and fold these back into the design calculus. Paraphrasing Robert Michels (1962, p. 370) on oligarchy, nothing but a serene and frank examination of the contractual hazards of deregulation will enable us to mitigate these hazards.

4.6 Public Policy

The initial public policy applications of TCE were in the field of industrial organization, especially to antitrust and regulation. These were followed by public policy applications to labor, health, agriculture, public finance, economic development, and reform, and the list goes on. If, indeed, any problem that arises as or can be reformulated as a contracting problem can be examined to advantage through the lens of contracting, then the list of applications is unending.

4.6.1 Antitrust

A long-standing puzzle for antitrust was what to make of vertical integration. If the “natural” way to procure a good or service was in the market, why take a transaction out of the market and organize it internally? For that matter, what explained efforts by firms to go beyond simple market exchange and impose customer, territorial, and other vertical market restrictions on distributors? Issues of both kinds arose during

the year that I spent as Special Economic Assistant to the head of the Antitrust Division of the US Department of Justice (1966–1967). Stringent vertical merger guidelines were issued by the Department in 1968. And the Justice Department mistakenly ascribed anticompetitive purposes to prohibitions against franchisee resale of bicycles in arguing the Schwinn case in 1967.¹⁵ More generally, the prevailing view on vertical market restrictions was that these were to be interpreted “not hospitably in the common law tradition, but inhospitably in the tradition of antitrust.”¹⁶ If the natural boundary of the firm was defined by technology (the firm being viewed as a production function), then what useful purpose was served by interfering with nature?

All well and good in the context of simple market exchange. It overreaches, however, to prohibit node C governance by forcing all market transactions for which $k > 0$ onto node B because the “market is a marvel.” Organization in all of its forms is a marvel—once we understand which transactions go where and why. Monopoly purpose enters into the calculus only when the requisite preconditions for monopoly are satisfied, which is a special case.

4.6.2 Regulation/Deregulation

As discussed above, the obvious regulatory problem to which to bring TCE to bear was the imaginative proposal to use franchise bidding as the solution to natural monopoly. In a TCE world where all feasible forms of organization are flawed, it is not surprising that franchise bidding can be expected to work well in some circumstances but not in all.

Other applications of comparative contractual reasoning to regulation (broadly construed) include the much condemned sugar program (Williamson, 1996, pp. 197–210) and consumer health and safety and labor health and safety (Williamson & Bercovitz, 1996, pp. 343–347) issues. More generally, the lens of contract can be applied to the full range of regulatory issues (at local, state, and federal levels) and such uses have been growing. As matters stand presently, however, TCE is still underused in relation to its potential.

4.6.3 Other Public Policy

Avinash Dixit opens his monograph on *The Making of Economic Policy: A Transaction-Cost Politics Perspective* (1996) with a contrast with old-style public finance, where the government was described as an omniscient, omnipotent, and benevolent dictator (1996, p. 8), with the lens of contract approach, according to

¹⁵For a discussion see Williamson (1985, pp. 183–189).

¹⁶The quotation is attributed to Turner by Stanley Robinson, 1968, N.Y. State Bar Association, Antitrust Symposium, p. 29.

which all feasible forms of organization are flawed. This latter is an immediate ramification of describing human actors as boundedly rational, which disallows omniscience, and given to subgoal pursuit (opportunism), which disallows benevolence. Upon recognizing implementation obstacles, moreover, omnipotence also drops out.

Two crucial TCE moves endorsed by Dixit are the view of the firm as governance structure and to insist upon remediableness. He observes with reference to the first that the standard normative approach to policy analysis views the process as taking place within (1996, p. 9):

... a social welfare maximizing black box, exactly as the neoclassical theory of production and supply viewed the firm as a profit-maximizing black box. While some useful insights follow from this, it leaves some very important gaps in our understanding and gives us some very misleading ideas about the possibilities of beneficial policy intervention. Economists studying business and industrial organization have long recognized the inadequacy of the neoclassical view of the firm and have developed richer paradigms and models based on the concepts of various kinds of transaction costs. Policy analysis also stands to benefit from such an approach, opening the black box and examining the actual workings of the mechanism inside. That is the starting point, and a recurrent theme, of this monograph.

The remediableness criterion for evaluating public policy proposals is to be contrasted with that of the Pareto criterion. Whereas the latter typically scants issues of feasibility and implementation, the remediableness criterion makes express provision to both. Thus the remediableness criterion holds that an extant practice or mode of organization for which no *feasible* superior mode can be described and *implemented* with expected net gains is *presumed* to be efficient. Reference to feasibility disallows hypothetical ideals (costlessness in any of its forms, including costless bargaining, is thus disallowed). Reference to implementation entails looking ahead to uncover obstacles to implementation, after which the mechanisms are worked out. And presumptions of efficiency are rebuttable—possibly by showing that the initial conditions (often of a political kind) are not acceptable (Williamson, 1996, pp. 208–212).

Whereas lens of choice reasoning holds that a simple display of deadweight losses is dispositive of inefficiency, the lens of contract (remediableness) holds otherwise. Now the analyst is pushed to establish that the proposed reform is feasible (recall that hypothetical ideals are disallowed) and further to demonstrate that “legitimate” resistance can be overcome in a cost-effective way. Those are not impossible obstacles, but they are very demanding. Among other things, ready recourse to costless compensation of losers (of a Hicks-Kaldor kind) is disallowed.

4.7 Conclusions

Robert Solow’s prescription for doing good economics is set out in three injunctions: keep it simple; get it right; make it plausible (2001, p. 111). Keeping it simple entails stripping away the inessentials and going for the main case (the jugular). Getting it right “includes translating economic concepts into accurate mathematics (or diagrams, or words) and making sure that further logical operations are correctly

performed and verified” (Solow, 2001, p. 112). Making it plausible entails describing human actors in (reasonably) veridical ways and maintaining meaningful contact with the phenomena of interest (contractual or otherwise).

To this, moreover, I would add a fourth injunction: derive refutable implications to which the relevant (often microanalytic) data are brought to bear. Nicholas Georgescu-Roegen has a felicitous way of putting it: “The purpose of science in general is not prediction, but knowledge for its own sake,” yet prediction is “the touchstone of scientific knowledge” (1971, p. 37).

Why the fourth injunction? This is necessitated by the need to choose among alternative theories that purport to deal with the same phenomenon—say vertical integration—and (more or less) satisfy the first three injunctions. Thus, assume that all of the models are tractable, that the logic of each hangs together, and that agreement cannot be reached as to what constitutes veridicality and meaningful contact with the phenomena. Does each candidate theory then have equal claims for our attention? Or should we be more demanding? This is where refutable implications and empirical testing come in: ask each would-be theory to stand up and be counted.

Why more economists are not insistent upon deriving refutable implications and submitting these to empirical tests is a puzzle. One possibility is that the world of theory is set apart and has a life of its own. A second possibility is that some economists do not agree that refutable implications and testing are important. Another is that some theories are truly fanciful and their protagonists would be discomfited by disclosure. A fourth is that the refutable implications of favored theories are contradicted by the data. And perhaps there are still other reasons. Be that as it may, a multiplicity of theories, some of which are vacuous and others of which are fanciful, is an embarrassment to the pragmatically oriented members of the tribe. Among this subset, insistence upon the fourth injunction—derive refutable implications and submit these to the data— is growing.

TCE responds to the injunction of keeping it simple by taking economizing on transaction costs to be the main case. The logic of economic organization is that of discriminating alignment (the implementation of which requires that the key attributes of transactions and governance structures be named and the logic of efficient alignment be worked out). The main response to the plausibility injunction is to describe human actors in more veridical terms—in cognitive, self-interestedness, and feasible foresight respects. And TCE is insistent upon deriving refutable implications to which the data are thereafter brought to bear.

As described elsewhere, TCE has progressed from informal into pre-formal, semi-formal, and fully formal stages. As matters stand presently, however, some efforts at fully formal modeling lose contact with key issues.¹⁷ Specifically, if adaptation (of autonomous and cooperative kinds) is truly the central problem of

¹⁷I have reference to the Property Rights Theory of firm and market organization (Hart, 1995), which Bernard Salani  mistakenly describes as formalizing “the intuitions of transaction cost economics, as created by Coase and Williamson” (1997, p. 176). For a discussion, see Williamson (2002a).

economic organization, then to annihilate ex post maladaptation (by making implausible assumptions of common knowledge of payoffs and costless bargaining), thereby to focus entirely on ex ante incentive alignment, is deeply problematic. Recent formal models have nevertheless begun to restore attention to ex post governance (Bajari & Tadelis, 2001).

However, such fully formal modeling shapes up, there is broad agreement that work of a transaction cost economics kind has helped to transform our understanding of complex contracting and economic organization (in both theoretical and public policy respects) and that applications outside of economics are growing and will continue.

Acknowledgments This is not the first time that I have attempted to describe the transaction cost economics project. For earlier treatments, see Williamson (1989, 1998). My thanks to Claude Ménard and the referees for constructive suggestions and to the Rockefeller Study and Conference Center at Bellagio for tranquility.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Anderson, E., & Schmittlein, D. C. (1984). Integration of the sales force: An empirical examination. *RAND Journal of Economics*, 15(3), 385–395.
- Bajari, P., & Tadelis, S. (2001). Incentives versus transaction costs: A theory of procurement contracts. *RAND Journal of Economics*, 32, 387–407.
- Barnard, C. I. (1938). *The functions of the executive*. Harvard University Press.
- Baron, J. N., & Kreps, D. M. (1999). *Strategic human resources: Frameworks for general managers*. Wiley.
- Becker, G. (1962). Investment in Human Capital: Effects on earnings. *Journal of Political Economy*, 70, 9–49.
- Boerner, C., & Macher, J. (2002). Transaction cost economics: A review and assessment of the empirical literature. Unpublished manuscript.
- Brennan, G., & Buchanan, J. (1985). *The reason of rules*. Cambridge University Press.
- Buchanan, J. M. (1975). A contractarian paradigm for applying economic theory. *American Economic Review*, 65, 225–230.
- Buchanan, J. M. (1987). The constitution of economic policy. *American Economic Review*, 77, 243–250.
- Buchanan, J. M. (2001). Game theory, mathematics, and economics. *Journal of Economic Methodology*, 8, 27–32.
- Buchanan, J. M., & Tullock, G. (1962). *The calculus of consent: Logical foundations of constitutional democracy*. University of Michigan Press.
- Coase, R. H. (1937). The nature of the firm. *Economica*, 4, 386–405.
- Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 3, 1–44.
- Coase, R. H. (1972). Industrial organization: A proposal for research. In V. R. Fuchs (Ed.), *Policy issues and research opportunities in industrial organization* (pp. 59–73). National Bureau of Economic Research.
- Coase, R. H. (1984). The new institutional economics. *Journal of Institutional and Theoretical Economics*, 140, 229–231.

- Coase, R. H. (1992). The institutional structure of production. *American Economic Review*, 82, 713–719.
- Commons, J. R. (1932). The problem of correlating law, economics and ethics. *Wisconsin Law Review*, 8, 3–26.
- Crocker, K. J., & Masten, S. E. (1996). Regulation and administered contracts revisited: Lessons from transaction-cost economics for public utility regulation. *Journal of Regulatory Economics*, 8, 5–39.
- Crocker, K. J., & Masten, S. E. (1991). Pretia ex Machina? Prices and process in long-term contracts. *Journal of Law and Economics*, XXXIV(1), 69–99.
- Demsetz, H. (1968). Why regulate utilities? *Journal of Law and Economics*, 11, 55–66.
- Dixit, A. K. (1996). *The making of economic policy: A transaction-cost politics perspective*. MIT Press.
- Easterbrook, F., & Fischel, D. (1986). Close corporations and agency costs. *Stanford Law Review*, 38, 271–301.
- Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. *Journal of Law and Economics*, 26, 301–326.
- Friedman, M. (1962). *Capitalism and freedom*. University of Chicago Press.
- Gely, R., & Spiller, P. T. (1990). A rational choice theory of supreme court statutory decisions with applications to the *State Farm* and *Grove City* cases. *Journal of Law, Economics, and Organization*, 6(2), 263–300.
- Georgescu-Roegen, N. (1971). *The entropy law and economic process*. Harvard University.
- Goldberg, V. P. (1976). Regulation and administered contracts. *Bell Journal of Economics*, 7(2), 426–448.
- Goldberg, V. P., & Erickson, J. R. (1987). Quantity and price adjustment in long-term contracts: A case study of petroleum coke. *Journal of Law and Economics*, XXX(2), 369–398.
- Hart, O. (1995). *Firms, contracts and financial structure*. Oxford University Press.
- Hayek, F. (1945). The use of knowledge in society. *American Economic Review*, 35, 519–530.
- Holmstrom, B., & Tirole, J. (1989). The theory of the firm. In R. Schmalensee & R. Willig (Eds.), *Handbook of industrial organization*. Oxford University Press.
- John, G., & Weitz, B. A. (1988). Forward integration into distribution: An empirical test of transaction cost analysis. *Journal of Law, Economics, and Organization*, 4(2), 337–355.
- Joskow, P. L. (1987). Contract duration and relationship-specific investments: Empirical evidence from coal markets. *American Economic Review*, 77(1), 168–185.
- Joskow, P. L. (1988). Asset specificity and the structure of vertical relationships. *Journal of Law, Economics and Organization*, 4(1), 95–117.
- Joskow, P. L. (2000). Transaction cost economics and competition policy. Unpublished manuscript.
- Kenney, R. W., & Klein, B. (1983). The economics of block booking. *Journal of Law and Economics*, XXVI(3), 497–540.
- Klein, B. (1988). Vertical integration as organizational ownership: The fisher body-general motors relationship revisited. *Journal of Law, Economics, and Organization*, 4(1), 199–213.
- Leffler, K. B., & Rucker, R. R. (1991). Transactions costs and the efficient organization of production: A study of timber-harvesting contracts. *Journal of Political Economy*, 99(5), 1060–1087.
- Levy, B., & Spiller, P. T. (1994). The institutional foundations of regulatory commitment: A comparative analysis of telecommunications regulation. *Journal of Law Economics, and Organization*, 10(2), 201–246.
- Llewellyn, K. N. (1931). What price contract? An essay in perspective. *Yale Law Journal*, 40 (May), 704–751.
- Lyons, B. R. (1996). Empirical relevance of efficient contract theory: Inter-firm contracts. *Oxford Review of Economic Policy*, 12(4), 27–52.
- Macneil, I. R. (1974). The many future of contracts. *Southern California Law Review*, 72, 854–906.
- Makowski, L., & Ostroy, J. (2001). Perfect competition and the creativity of the market. *Journal of Economic Literature*, 32, 479–535.
- March, J. G., & Simon, H. A. (1958). *Organizations*. New York: Wiley.

- Masten, S. E. (1984). The Organization of Production: Evidence from the aerospace industry. *Journal of Law and Economics*, XXVII(2), 403–417.
- Masten, S. E., & Crocker, K. J. (1985). Efficient adaptation in long-term contracts: Take-or-pay provisions for natural gas. *American Economic Review*, 75(5), 1083–1093.
- Masten, S. E., & Saussier, S. (2000). Econometrics of contracts: An assessment of developments in the empirical literature on contracting. *Revue D'Economie Industrielle*, 92, 215–236.
- Masten, S. E., & Snyder, E. A. (1993). United States versus united shoe machinery corporation: On the merits. *Journal of Law and Economics*, XXXVI(1 (Part 1)), 33–70.
- Masten, S. E., Meehan, J. W., Jr., & Snyder, E. A. (1991). The costs of organization. *Journal of Law, Economics, and Organization*, 7Z(1), 1–25.
- Merton, R. (1936). The unanticipated consequences of purposive social action. *American Sociological Review*, 1, 894–904.
- Michels, R. (1962). *Political parties*. Glencoe (Il.): Free Press.
- Monteverde, K., & Teece, D. J. (1982). Supplier switching costs and vertical integration in the automobile industry. *Bell Journal of Economics*, 13(1), 206–213.
- Mulherin, J. H. (1986). Complexity in long-term contracts: An analysis of natural gas contractual provisions. *Journal of Law, Economics, and Organization*, 2(1), 105–117.
- Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change*. Belknap Press.
- Newell, A., & Simon, H. (1972). *Human problem solving*. Prentice-Hall.
- North, D. (2000). A revolution in economics. In C. Ménard (Ed.), *Institutions, contracts, and organizations* (pp. 37–42). Edward Elgar.
- Palay, T. M. (1984). Comparative institutional economics: The governance of rail freight contracting. *Journal of Legal Studies*, XIII(2), 265–287.
- Peltzman, S. (1991). The handbook of industrial organization: A review article. *Journal of Political Economy*, 99, 201–217.
- Peltzman, S., & Whinston, M. (2000). *Deregulation of network industries*. Brookings Institution Press.
- Posner, R. A. (1972). The appropriate scope of regulation in the cable television industry. *Bell Journal of Economics*, 3, 98–129.
- Posner, R. A. (1986). *Economic analysis of law* (3rd ed.). Little Brown.
- Priest, G. L. (1993). The origins of utility regulation and the ‘theories of regulation’ debate. *Journal of Law and Economics*, XXXVI(1 (Part 2)), 289–323.
- Reder, M. W. (1999). *Economics: The culture of a controversial science*. University of Chicago Press.
- Rindfleisch, A., & Heide, J. (1997). Transaction cost analysis: Past, present and future applications. *Journal of Marketing*, 61, 30–54.
- Robbins, L. (1932). *An essay on the nature and significance of economic science*. New York University Press.
- Rothband, M. (1987). Catallactics. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *The new Palgrave: A dictionary of economics* (Vol. 1, pp. 377–378). The Stockton Press.
- Salanié, B. (1997). *The economics of contracts: A primer*. The MIT Press.
- Scott, R. W. (1992). *Organizations*. Prentice-Hall.
- Selznick, P. (1949). *TVA and the grass roots*. University of California Press.
- Selznick, P. (1950). The iron law of bureaucracy. *Modern Review*, 3, 157–165.
- Shelanski, H. A., & Klein, P. G. (1995). Empirical research in transaction cost economics: A review and assessment. *Journal of Law, Economics, and Organization*, 11, 335–361.
- Shultz, G. (1995). Economics in action. *American Economic Review*, 85, 1–8.
- Simon, H. A. (1947). *Administrative behavior*. The Free Press.
- Simon, H. A. (1957). *Models of man: Social and rational; mathematical essays on rational human behavior in a social setting*. Wiley.
- Simon, H. A. (1983). *Reason in human affairs*. Stanford University Press.
- Simon, H. (1985). Human nature in politics: The dialogue of psychology with political science. *American Political Science Review*, 79, 293–304.
- Solow, R. (2001). A native informant speaks. *Journal of Economic Methodology*, 8, 111–112.

- Thompson, J. S. (1967). *Organizations in action: Social science bases of administrative theory*. McGraw-Hill.
- Varian, H. (2002). If there was a new economy, why wasn't there a new economics?. *New York Times*, 17 January, C2.
- Weingast, B., & Marshall, W. J. (1988). The industrial organization of congress: Or, why legislatures, like firms, are not organized as markets. *Journal of Political Economy*, 96(1), 132–163.
- Werin, L. (2000). Ronald Coase and the new microeconomics. In C. Ménard (Ed.), *Institutions, contracts, and organizations* (pp. 42–47). Edward Elgar.
- Williamson, O. E. (1975). *Markets and hierarchies: Analysis and antitrust implications*. The Free Press.
- Williamson, O. E. (1976). Franchise bidding for natural monopolies – In general and with respect to CATV. *Bell Journal of Economics*, 7(1), 73–104.
- Williamson, O. E. (1979). Assessing vertical market restrictions: Antitrust ramifications of the transaction cost approach. *University of Pennsylvania Law Review*, 127, 953–993.
- Williamson, O. E. (1985). *The economic institutions of capitalism*. The Free Press.
- Williamson, O. E. (1988). Corporate finance and corporate governance. *Journal of Finance*, 43, 567–591.
- Williamson, O. E. (1989). Transaction cost economics. In R. Schmalensee & R. Willig (Eds.), *Handbook of industrial organization* (Vol. 1, pp. 136–184). North Holland.
- Williamson, O. E. (1994). Evaluating Coase. *Journal of Economic Perspectives*, 8, 201–204.
- Williamson, O. E., & Masten, S. E. (1995). *Transaction cost economics*. Aldershot (UK): Edward Elgar Pub.
- Williamson, O. E. (1996). *The mechanisms of governance*. Oxford University Press.
- Williamson, O. E. (1998). Transaction cost economics: How it works; where it is headed. *De Economist*, 146, 23–58.
- Williamson, O. E. (1999). Public and private bureaucracies. *Journal of Law, Economics, and Organization*, 15, 306–342.
- Williamson, O. E. (2002a). The theory of the firm as a governance structure: From choice to contract. *Journal of Economic Perspectives*, 16, 171–195.
- Williamson, O. E. (2002b). The lens of contract: Private ordering. *American Economic Review*, 92, 438–443.
- Williamson, O. E., & Bercovitz, J. (1996). The modern corporation as an efficiency instrument. In C. Kaysen (Ed.), *The American corporation today* (pp. 327–359). Oxford University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Doing Institutional Analysis Digging Deeper than Markets and Hierarchies

5

Elinor Ostrom

Contents

5.1	Introduction	74
5.2	Further Challenges Involved in Studying Institutions	77
5.2.1	Institutions Are Invisible	78
5.2.2	Multiple Definitions of Institutions	79
5.2.3	Multiple Levels of Analysis	79
5.2.4	Configural Relationships	80
5.3	Institutional Frameworks, Theories, and Models	80
5.4	The Institutional Analysis and Development (IAD) Framework	82
5.5	Explanation Viewing Action Situations as Dependent Variables	84
5.5.1	The Concept of Rules	85
5.5.2	Rule Configurations	88
5.5.3	Physical and Material Conditions	91
5.5.4	Excludability and the Free-Rider Problem	92
5.5.5	Subtractability of the Flow	93
5.5.6	Additional Attributes	94
5.5.7	Attributes of the Community	95
5.6	Linking Action Arenas	95
5.7	Conclusion	98
	References	99

Abstract

In this chapter, I present the institutional analysis and development (IAD) framework that can be used in analyzing any type of institutional arrangement. A major problem in understanding institutions relates to the complexity and diversity of contemporary life and the resulting specialization that has occurred within the

Elinor Ostrom sadly passed away before this second edition of the *Handbook of New Institutional Economics* was created. Her chapter from the first edition remains a classic and emblematic of her invaluable contributions to the field.

E. Ostrom (✉)
Indiana University, Bloomington, IN, USA

social sciences. The time is ripe for an interdisciplinary effort to examine the underlying component parts that can be used to build useful theories of human behavior in the diverse range of situations in which humans interact. This chapter asserts that there are universal components of all markets and other frequently encountered situations and regularities in human behavior that can be analyzed using the IAD framework, which was developed based on years of work with colleagues and has proven its utility in a wide diversity of empirical settings.

Keywords

Institutional analysis · Rules-in-use · Action arena · IAD framework · Free rider problem

5.1 Introduction

A major problem in understanding institutions relates to the complexity and diversity of contemporary life and the resulting specialization that has occurred within the social sciences. The central aim of the social sciences is to explain human behavior. But what kind of human behavior? Within which kinds of institutional settings?

As we go about our everyday life, we interact in a diversity of complex situations. Many of us face a morning and evening commute where we expect that others, who are traveling at great speeds, will observe the rules of the road. Our very lives depend on these expectations. Those of us who work in large organizations—universities, research centers, business firms, government offices—participate in a variety of team efforts. In order to do our own work well, we are dependent on others to do their work creatively, energetically, and predictably and vice versa.

Many of us play sports at noontime, in the early evening, and on the weekends. Here, again, we need to learn the basic rules of each of the games we play, as well as finding colleagues with whom we can repeatedly engage in this activity. During the average week, we undertake activities in various types of market settings—ranging from buying our everyday food and necessities to investing funds in various types of financial instruments. And we spend some hours each week with family and friends in a variety of activities that may involve worship, helping offspring with homework, taking care of our homes and gardens, and a long list of other activities undertaken with family and friends.

The formal study of institutions is typically divided into the study of separate kinds of situations. Students are confronted with the need to choose between the study of markets offered by departments of economics and the study of hierarchies or states offered by departments of political science. The study of communities is sometimes offered in anthropology or sociology departments but may not be offered at all in some universities. Are markets and hierarchies entirely “different” structures? Do they, in turn, share little with families, neighborhoods, and committees?

Are there no underlying universal building blocks of organized life analogous to the underlying universal building blocks of individual organisms?

In this chapter, I address the question of whether underlying components of markets and hierarchies (and, many other complex situations) constitute the elemental parts of multiple, complementary theories that explain regularities in human behavior across diverse and complex situations. In other words, I assert that there are universal components of all markets and other frequently encountered situations and provide a framework that can be used in analyzing any type of institutional arrangement.

Contemporary ways of organizing scientific knowledge have not encouraged pursuing this question. Whole disciplines have been built up around the presumption that they offer a *unique* perspective on the study of particular types of situations. Markets and hierarchies are presented in some courses as fundamentally different “pure types” of organization. Not only are these types of institutional arrangements perceived to be different, but each is presumed to require its *own* explanatory theory. Recent works in transaction cost economics do, of course, bridge these structures in the discussion of choosing firms or markets to organize activities (Williamson, 1991).

When we study or teach about these situations, behavior is presented as coming from separate worlds. Psychologists tend to study the way that single individuals acquire and process information and skills and how this affects individual choices. The theories of individual behavior developed in psychology are little known or used outside of psychology. Many economists develop their own theories of individual behavior that are not consistent with the work undertaken in contemporary psychology. Political scientists tend to study various kinds of collective-choice mechanisms (legislatures, executives, courts, and the selection of officials through voting) and are eclectic in the theories and approach they use. The discipline of law is taught in separate schools and usually made available only to those interested in becoming lawyers by profession. Social scientists rarely take any biology—and vice versa—even though recent developments in biology are highly relevant for understanding human behavior.

In each discipline, a separate set of languages is emphasized that makes diverse assumptions about the following:

- Kinds of goals that individuals seek, the images they share, and the mechanisms they use in making decisions
- Variables that are important in affecting the structure of the situations in which individuals make choices
- Capabilities of individuals and the extent of freedom that individuals have to affect the structure of these situations
- Range of outcomes that is likely to be achieved in the relevant situations

The development of separate languages is a barrier to more general explanatory frameworks and closely related theories that help analysts make cross-institutional comparisons and evaluations.

What we have learned from all of our separate disciplinary work is that there is no single cause of human behavior—even though some scholars tend to push one or another all-encompassing cause such as poverty, population, or the division of labor. Even life itself is not dependent on one cause. To live, one needs at least oxygen, water, and minimum levels of nutrition. All of these are key parts of the explanation of life, but none is a sufficient explanation of living beings. Living beings also exist at multiple levels. The human genotype is a set of instructions for building human phenotypes. Natural selection processes operating primarily at an individual level, but also in multilevel processes, have selected for some genes and selected against others. Parsing the genetic code has been one of the major breakthroughs of contemporary scientific endeavors. Our biological foundations help to explain some aspects of all of our behavior, but much of our behavior is also affected by the structure of the diverse range of situations within which we interact with others.

After decades of moving further and further apart—and engaging in sometimes futile arguments about which discipline has the best approach to understanding human behavior—a very encouraging trend over the last several decades has been the growth of interdisciplinary approaches to the study of human behavior. The number of journals with two disciplines in the title has been growing: *The Journal of Law and Economics*, *Political Sociology*, *Ecological Economics*, and many others. The Public Choice Society was started in the 1970s—originally as the “no name society” by economists, political scientists, and sociologists (see Ostrom, 1964). The International Association for the Study of Common Property was started in the mid-1980s to bring together anthropologists, economists, engineers, historians, political scientists, sociologists, and many others. The International Society for New Institutional Economics—while it has only economics in the title—includes a large number of noneconomists as key participants. This Handbook illustrates the diversity of formal disciplines that contribute to the study of institutions.

Thus, the time is ripe for an effort that attempts to draw on the foundations of many disciplines including anthropology, biology, economics, law, philosophy, political science, psychology, and sociology, to attempt to answer the core question of this chapter mentioned above. What are the underlying component parts that can be used to build useful theories of human behavior in the diverse range of situations in which humans interact? Can we identify the working parts of any kind of situation in which humans find themselves interacting on a repetitive basis? Can we use the same components to build an explanation for behavior in a commodity market as we would use to explain behavior in a university, or a religious order, or a transportation system, or an urban public economy? Can we identify the multiple levels of analysis needed to explain the regularities in human behavior that we observe? I give a positive answer to these questions based on years of work with colleagues developing the institutional analysis and development (IAD) framework.¹

¹For earlier discussions, see Kiser and Ostrom (1982), Ostrom (1986), Oakerson (1992), Ostrom et al. (1994), and Crawford and Ostrom (1995).

The publication of “The Three Worlds of Action: A Metatheoretical Synthesis of Institutional Approaches” (Kiser & Ostrom, 1982) represents the initial published attempt to describe the IAD framework intended to help integrate work undertaken by political scientists, economists, anthropologists, lawyers, sociologists, psychologists, and others interested in how institutions affect the incentives confronting individuals and their resultant behavior.² During the two decades since this publication, the framework has been developed further and applied to analyze a diversity of empirical settings.³ These include the following:

- The study of land boards in Botswana (Wynne, 1989)
- The evolution of coffee cooperatives in the Cameroon (Walker, 1998)
- The effect of rules on the outcomes of common-pool resource settings throughout the world (Ostrom, 1990, 1992; Schlager, 1990; Blomquist, 1992; Tang, 1992; Ostrom et al., 1994; Lam, 1998; National Research Council, 2002)
- The study of local public economies in urban areas (McGinnis, 1999)
- The evolution of banking reform in the United States (Polski, 2003)
- The change from group ranches to private ranches among the Maasai of Kenya (Mwangi, 2003)
- The effect of formal decentralization laws on local-level outcomes (Andersson, 2002)

Our confidence in the utility of the framework has grown steadily in light of the wide diversity of empirical settings where it has helped colleagues identify the key variables to undertake a systematic analysis of the structure of the situations that individuals faced and how rules, the nature of the events involved, and the community affected (and were affected by) these situations over time. What is certainly true is that the number of specific variables involved in each of these empirical studies is very large, and specific values of variables involved in one study (or one location in a study) differed from the specific values of variables involved in another study.

5.2 Further Challenges Involved in Studying Institutions

A set of further difficulties needs to be overcome in undertaking any form of institutional analysis. Some of these key difficulties involved in studying institutions include the following:

1. While the buildings in which organized entities are located are quite visible, institutions themselves are usually invisible.

²Elements of the framework have been used in teaching both graduate and undergraduate courses at Indiana University since the mid-1970s. Sections of this chapter are based on Ostrom (1999b).

³Ostrom, 1986, 1999b; Ostrom et al., 1994, Chap. 2; Oakerson, 1992.

2. The term “institution” is used to refer to many different types of entities including organizations as well as the rules, norms, and strategies used to structure patterns of interaction within and across organizations.
3. Given the multiple languages used across disciplines, a coherent institutional framework is needed to allow for expression and comparison of diverse theories and models of theories applied to particular puzzles and problem settings.
4. Decisions made about rules at any one level are usually made within a structure of rules existing at a different level. Thus, institutional studies need to encompass multiple levels of analysis.
5. At any one level of analysis, combinations of prescriptions, attributes of the world, and communities of individuals involved work together in a configural, rather than an additive, manner.

Let us briefly discuss these issues.

5.2.1 Institutions Are Invisible

One of the most difficult problems to overcome in the study of institutions is how to identify and measure them. Because institutions are fundamentally shared concepts, they exist in the minds of participants and sometimes are shared as implicit knowledge rather than in an explicit and written form. One of the problems facing scholars and officials is learning how to recognize the presence of institutions on the ground. The primitive physical structures that embed property-rights systems that farmers have constructed over time look flimsy to an engineer who considers only structures built out of concrete and iron to be real.⁴ These flimsy structures, however, are used by individuals to allocate resource flows to participants according to rules that have been devised in tough constitutional and collective-choice bargaining situations over time.

Unlike physical structures that are immediately visible on the horizon, rules are invisible structures that can be deeply buried under the regularities of observed behavior. We see, for example, the regularities of human behavior involved in an election campaign, an open farmers’ market, or among drivers on a major freeway. Many behavioral regularities result from the physical and technological world involved. The heavy use of television advertising in contemporary election campaigns in Western countries, for example, is due in part to the vast economies of scale involved in this mode of communicating. Some of what we see, on the other hand, results from rules. The rule that requires TV channels to give equal access to all candidates, the equal time rule, affects the amount of broadcast time assigned to all candidates.

⁴ An excellent example of what looks like “flimsy structures” that have lasted for centuries is the Banaue Rice Terraces in the Philippines (rice terraces curved from the top of mountain ranges to the valleys below and maintained for over 2000 years and still working today). UNESCO has put them on its World Heritage List. I am thankful to Ed Aralar for suggesting this example.

5.2.2 Multiple Definitions of Institutions

It is hard to make much progress in the study of institutions if scholars define the term to mean almost anything. A major confusion exists between scholars who use the term to refer to an organizational entity such as the US Congress, a business firm, a political party, a family, and scholars who use the term to refer to the accepted rules, norms, and strategies adopted by individuals operating within or across organizational settings. In this chapter, following the definition used by Douglass North (1990, this volume Chap. 2), I use the term “institution” in the latter sense to refer to the rules, norms, and strategies used by humans in repetitive situations. By rules, I mean shared prescriptions (must, must not, or may) that are mutually understood and enforced in particular situations in a predictable way by agents responsible for monitoring conduct and for imposing sanctions (see Crawford & Ostrom, 1995). By norms, I mean shared prescriptions known and accepted by most of the participants themselves involving intrinsic costs and benefits rather than material sanctions or inducements.⁵ By strategies, I mean the regularized plans that individuals make within the structure of incentives produced by rules, norms, and expectations of the likely behavior of others in a situation affected by relevant physical and material conditions.⁶

In training researchers to identify and measure institutions, we stress the concept of rules-in-use rather than focusing on rules-in-form. Rules-in-use are referred to whenever someone new (such as a new employee or a child) is being socialized into an existing rule-ordered system of behavior. They are the “do’s and don’ts” that one learns on the ground that may not exist in any written document. In some instances, they may actually be contrary to the “do’s and don’ts” that are written in formal documents. Being armed with a set of questions concerning how is X done here and why is Y not done here is a useful way of identifying rules-in-use, shared norms, and operational strategies.

5.2.3 Multiple Levels of Analysis

When individuals interact in repetitive settings, they may be in operational situations that directly affect the world, or they may be making decisions at other levels of analysis that eventually impinge on operational decision-making situations (Shepsle, 1989). Multiple sources of structure are located at diverse analytical levels as well as diverse geographic domains. Biologists took several centuries to learn how to separate the diverse kinds of relevant structures needed to analyze both communities

⁵Norms are viewed by institutional analysts drawing on sociological traditions as social facts that are not changeable by the individual (see Seo & Creed, 2002; Scott, 2001).

⁶When doing formal game-theoretical analysis, such strategies would be those identified as equilibrium strategies. Shared strategies may, however, take the form of heuristics adopted by most individuals in a society when they find themselves in particular situations.

and individual biological entities. Separating phenotypical structure from genotypical structure was part of the major neo-Darwinian breakthrough that allowed biologists to achieve real momentum and cumulation during the past century.

Besides multiple and nested action arenas at any one level of analysis, nesting of arenas also occurs across several levels of analysis. The nested structure of rules within rules, within still further rules, is a particularly difficult phenomenon to understand for those interested in institutions. Studies conducted at a macro level (see Kaminski, 1992; Ostrom, 1997; Loveman, 1993; Sawyer, 1992) tend to focus on a constitutional level of analysis (see below for further discussion of this level). Decisions made at this level affect collective-choice decisions as these impinge on the operational decisions of individuals (see Firmin-Sellers, 1996; Agrawal, 1998; Gibson, 1999). Finding ways to communicate across levels is a key challenge for all institutional theorists.

5.2.4 Configural Relationships

Successful analysis can cumulate rapidly when scholars have been able to analyze a problem by separating it into component parts that are analyzed independently and then recombining these parts in an additive fashion. Many puzzles of interest to social scientists can be torn apart and recombined. Frequently, however, the impact on incentives and behavior of one type of rule is not independent of the configuration of other rules. Thus, the impact of changing one of the current rules that is part of a state “welfare system” depends on which other rules are also in effect. Changing the minimum outside income that can be earned before losing benefits from one program, for example, cannot be analyzed independently of the effect of income on benefits derived from other programs. Similarly, analyzing the impact of changing the proportion of individuals who must agree prior to making an authoritative collective choice (e.g., 50% plus 1) depends on the quorum rule in force. If a quorum rule specifying a low proportion of members is in effect, requiring two-thirds agreement may be a less stringent decision rule than a simple majority rule combined with a quorum rule requiring a high proportion of members. *Ceteris paribus* conditions are always essential for doing any theoretical work involving institutions. In the case of institutional analysis, one needs to know the value of other variables rather than simply asserting that they are held constant. This makes institutional analysis a more difficult and complex enterprise than studies of phenomena that are strictly additive.

5.3 Institutional Frameworks, Theories, and Models

Given the need for multiple disciplines, and hence multiple disciplinary languages, and the multiple levels of analysis involved in studying configural relationships among rules, relevant aspects of the world, and cultural phenomenon, the study of institutions does depend on theoretical work undertaken at three levels of specificity that are often confused with one another. These essential foundations include

(1) frameworks, (2) theories, and (3) models. Analyses conducted at each level provide different degrees of specificity related to a particular problem.

The development and use of a general *framework* helps to identify the elements and relationships among these elements that one needs to consider for institutional analysis. Frameworks organize diagnostic and prescriptive inquiry. They provide the most general list of variables that should be used to analyze all types of institutional arrangements. Frameworks provide a metatheoretic language that is necessary to talk about theories and that can be used to compare theories. They attempt to identify the *universal* elements that any theory relevant to the same kind of phenomena would need to include. Many differences in surface reality can result from the way these variables combine with, or interact with, one another. Thus, the elements contained in a framework help the analyst generate the questions that need to be addressed when first conducting an analysis. In this chapter, I will present a framework for institutional analysis rather than specific theories.

The development and use of *theories* enable the analyst to specify which elements of the framework are particularly relevant for certain kinds of questions and to make general working assumptions about these elements. Thus, theories focus on a framework and make specific assumptions that are necessary for an analyst to diagnose a phenomenon, explain its processes, and predict outcomes. Several theories are usually compatible with any framework. Economic theory, game theory, transaction cost theory, social choice theory, covenantal theory, and theories of public goods and common-pool resources are all compatible with the IAD framework discussed in this chapter.

The development and use of *models* make precise assumptions about a limited set of parameters and variables. Logic, mathematics, game theory, experimentation and simulation, and other means are used to explore the consequences of these assumptions systematically on a limited set of outcomes. Multiple models are compatible with most theories. An effort to understand the strategic structure of the games that irrigators play in differently organized irrigation systems, for example, developed four families of models just to begin to explore the likely consequences of different institutional and physical combinations relevant to understanding how successful farmer organizations arranged for monitoring and sanctioning activities (Weissing & Ostrom, 1991). This is one set of models we have developed to analyze, in a precise manner, a subpart of the theory of common-pool resources.

For policymakers and scholars interested in issues related to how different governance systems enable individuals to solve problems democratically, the IAD framework helps to organize diagnostic, analytical, and prescriptive capabilities. It is similar in structure and intent to the “actor-centered institutionalism” framework developed by Renate Mayntz and Fritz Scharpf (1995) and applied to several national policy settings by Scharpf (1997). It also aids in the accumulation of knowledge from empirical studies and in the assessment of past efforts at reforms.

Without the capacity to undertake systematic, comparative institutional assessments, recommendations of reform may be based on naive ideas about which kinds of institutions are “good” or “bad” and not on an analysis of performance. One needs a common framework and family of theories in order to address questions of reforms

and transitions. Particular models then help the analyst to deduce specific predictions about likely outcomes of highly simplified structures. Models are useful in policy analysis when they are well-tailored to the particular problem at hand. Models are used inappropriately when applied to the study of problematic situations that do not closely fit the assumptions of the model.

5.4 The Institutional Analysis and Development (IAD) Framework

As indicated earlier, an institutional framework should identify the major types of structural variables present to some extent in *all* institutional arrangements but whose values differ from one type of institutional arrangement to another. In all the analyses presented in this chapter, one can think about a core set of building blocks—which are themselves constructed of other building blocks and affected by multiple external variables. The simplest and most aggregated way of representing all situations is shown in Fig. 5.1, where *actors* in *situations* together with exogenous factors generate *outcomes* that affect the actors in the situation and potentially others (see Ostrom et al., 1994, Chap. 2).

The outcomes that are achieved eventually feed back onto the actors and the situations and may transform both over time. When the outcomes are productive for those involved, they may increase their commitment to following the rules and norms that have evolved over time so as to continue to receive positive outcomes. When outcomes are perceived by those involved as of lower value than other outcomes that might be obtained, some actors will begin to raise questions about changing the structure of the situations through various changes in the exogenous variables or the structure of the situations themselves.

The IAD framework is a multitier conceptual map. The most basic schematic representation shown in Fig. 5.1 will be unpacked, as shown in Fig. 5.2. It can be further unpacked multiple times. The first step in using this framework to analyze a problem is to identify a conceptual unit—called an *action arena*—that can be utilized to analyze, predict, and explain behavior within institutional arrangements.⁷ Action arenas include an *action situation* and the *actors* in that situation. An action situation can be characterized using seven clusters of variables: (1) participants, (2) positions, (3) outcomes, (4) action–outcome linkages, (5) the control that participants exercise, (6) information, and (7) the costs and benefits assigned to outcomes.

An actor (an individual or a corporate actor) includes assumptions about four clusters of variables:

1. The *resources* that an actor brings to a situation
2. The *valuation* actors assign to states of the world and to actions

⁷The concept of an action arena is analogous to the concept of fields as used by Bourdieu (1977), DiMaggio and Powell (1983), and Fligstein (1998).

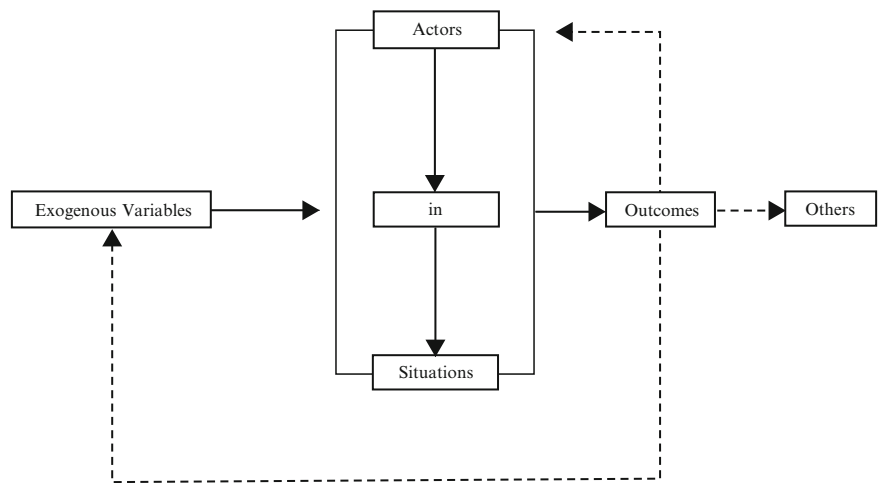


Fig. 5.1 The focal level of analysis

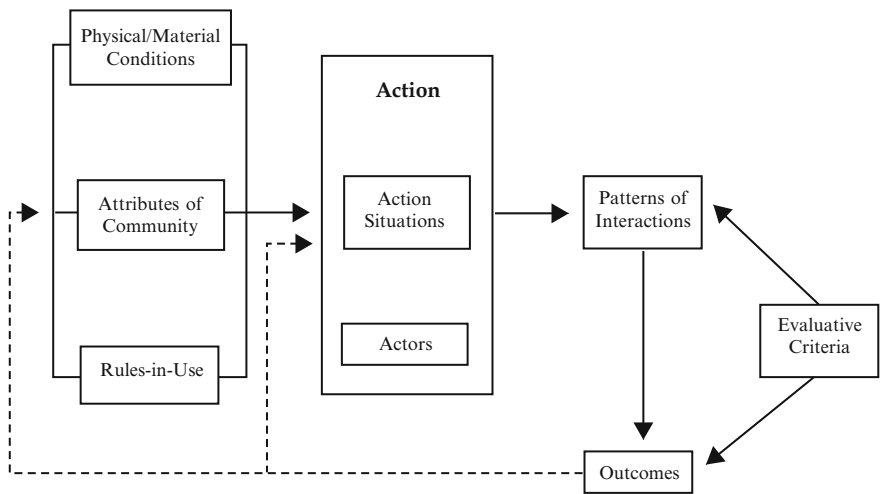


Fig. 5.2 A framework for institutional analysis. (Source: Adapted from Ostrom, Gardner, and Walker (1994, p. 37))

- 3. The way actors acquire, process, retain, and use *knowledge contingencies and information*
- 4. The processes actors use for *selection* of particular courses of action

The particular assumptions made by an analyst about each of these assumptions enable one to build a theory of self-centered rational choice, a theory of bounded rationality, or a theory of norm-driven behavior. All of these theories can be used to animate the “actor” module of the framework.

An action arena refers to the social space where individuals interact, exchange goods and services, solve problems, dominate one another, feel guilty, or fight (among the many things that individuals do in action arenas). Considerable ferment has been generated during the past several decades as a result of empirical studies—to a large degree, experimental studies—that have challenged some of the basic working assumptions used by many institutional analysts in predicting outcomes in various types of collective-action situations (Ostrom, 1998). A major proportion of theoretical work stops after predicting outcomes from a particular action arena and takes the variables specifying the situation and the motivational and cognitive structure of an actor as givens. Analysis proceeds toward the prediction of the likely behavior of individuals in such a structure.

An institutional analyst can take two additional steps after an effort is made to understand the initial structure of an action arena. One step digs deeper and inquires into the factors that affect the structure of an action arena. From this vantage point, the action arena is viewed as a set of variables dependent upon other factors. These factors affecting the structure of an action arena include three clusters of variables: (1) the rules and norms used by participants to order their relationships, (2) the attributes of states of the world that are acted upon in these arenas, and (3) the structure of the more general community within which any particular arena is placed (see Kiser & Ostrom, 1982). The second step examines how nested levels of rules affect behavior. I will now provide a brief introduction to how shared understandings of rules, states of the world, and nature of the community affect the values of the variables characterizing action arenas (see the left side of Fig. 5.2) and then turn to a discussion of the linkage of action situations.

5.5 Explanation Viewing Action Situations as Dependent Variables

Underlying the way analysts conceptualize action arenas are implicit assumptions about the *rules* individuals use to order their relationships, about attributes of *states of the world and their transformations*, and about the *nature of the community* within which the arena occurs. Some analysts are not interested in the role of these underlying variables and focus only on a particular arena whose structure is given. On the other hand, institutional analysts may be more interested in one factor affecting the structure of arenas more than they are interested in others. Sociologists tend to be more interested in how shared value systems affect the ways humans organize their relationships with one another. Environmentalists tend to focus on various ways that physical and biological systems interact and create opportunities or constraints on the situation human beings face. Political scientists tend to focus more on how specific combinations of rules affect incentives. Rules, states of the world, and the nature of the community all jointly affect the types of actions that individuals can take, the benefits and costs they attribute to these actions, and the likely outcomes achieved.

5.5.1 The Concept of Rules

The concept of rules is central to the analysis of institutions, but the term “rules” refers to many concepts with quite diverse meanings. In an important philosophical treatment of rules, Black (1962) identified four different usages of the term in everyday conversations. According to Black, the word “rule” is used to denote regulations, instructions, precepts, and principles. When used in its *regulation* sense, rules refer to something “laid down by an authority (a legislature, judge, magistrate, board of directors, university president, parent) as required of certain persons (or, alternatively, forbidden or permitted)” (ibid., p. 115). The example of a rule in the regulation sense that Black uses is: “The dealer at bridge must bid first.” When using rule in its regulation sense, one can meaningfully refer to activities such as the rule “being announced, put into effect, enforced (energetically, strictly, laxly, invariably, occasionally), disobeyed, broken, rescinded, changed, revoked, reinstated” (ibid., p. 109).

When the term “rules” is used to denote an *instruction*, it is closer in meaning to an effective strategy for how to solve a problem. An example of this usage is, “In solving quartic equations, first eliminate the cubic term” (ibid., p. 110). When speaking about a rule in this sense, one would not talk about a rule being enforced, rescinded, reinstated, or any of the other activities relevant to regulation. When rule denotes a *precept*, the term is being used as a maxim for prudential or moral behavior. An example would be: “A good rule is: to put charity ahead of justice” (ibid., p. 111). Again, one would not speak of enforcing, rescinding, or reinstating a rule in the precept sense.

The fourth sense in which the term rule is used in everyday language is to describe a law or principle. An example of this usage is: “Cyclones rotate clockwise, anticyclones anticlockwise” (ibid., p. 113). Principles or physical laws are subject to empirical test, and as such, truth values can be ascribed to them. But physical laws are not put into effect, broken, or rescinded.

Social scientists employ all four uses of the term that Black identifies. Scholars engaged in institutional analysis usually use the term to denote a regulation. Rules in the instruction sense can best be thought of as the strategies adopted by participants within ongoing situations. It is better to use the term “strategy” rather than “rule” for these plans of action. Rules in the precept sense are part of the generally accepted moral fabric of a community. I refer to these cultural prescriptions as norms. Rules in the principle sense are physical laws and are considered as part of the events in the biophysical world.

Until recently, rules have not been a central focus of most of the social sciences. Even in game theory, where “the rules of the game” seem to play an important role, there has not been much interest in examining where rules come from or how they change. Game-theoretical rules include all physical laws that constrain a situation as well as rules devised by humans to structure a situation. The rules of the game—including both physical and institutional factors—structure the game itself but have been irrelevant to many game theorists once a game can be unambiguously represented. Anatol Rapoport stated this position clearly:

Rules are important only to the extent that they allow the outcomes resulting from the choices of participants to be unambiguously specified. . . . Any other game with possibly quite different rules but leading to the same relations among the choices and the outcomes is considered equivalent to the game in question. In short, game theory is concerned with rules only to the extent that the rules help define the choice situation and the outcomes associated with the choices. Otherwise the rules of games play no part in game theory. (1966, p. 18)

Rules are shared understandings about potentially linguistic entities (Ganz, 1971; Ostrom, 1980; Commons, 1957) that refer to enforced prescriptions about what actions (or states of the world) are *required*, *prohibited*, or *permitted*. All rules are the result of implicit or explicit efforts to achieve order and predictability among humans by creating classes of persons (positions) who are then required, permitted, or forbidden to take classes of actions in relation to required, permitted, or forbidden states of the world (Crawford & Ostrom, 1995; Ostrom, 1991). Well-understood and enforced rules operate so as to *rule out* most occurrences of some actions.

Where do the rules that individuals use in action situations originate? In an open and democratic governance system, many sources of rules exist. It is not considered illegal or improper for individuals to self-organize themselves and craft their own rules if the activities they engage in are legal. In addition to the legislation and regulations of a formal central government, laws are considered passed by regional, local, and special governments. Within private firms and voluntary associations, individuals are authorized to adopt different rules for who is a member of the firm or association, how benefits are to be shared, and how decisions will be made. Each family constitutes its own rule-making body.

When individuals genuinely participate in the crafting of multiple layers of rules, some of that crafting will occur using pen and paper. Much of it, however, will occur as problem-solving individuals interact and try to figure out how to do a better job in the future than they have done in the past. Colleagues in a work team are crafting their own rules when they say to one another: “How about if you do A and I do B in the future, and before we make a decision about C again, we both discuss it and make a joint decision.” In a democratic society, problem-solving individuals do this all the time.

Thus, rules need not be written. Nor, do they need to result from formal legal procedures. Participants craft rules in order to change the structure of repetitive situations that they face so as to try to improve outcomes for themselves. Regardless of a rule’s origin, however, it is always possible to rescind it or reinstate it—or it is not a rule.

Rules-in-use are the set of rules to which participants make reference if asked to explain and justify their actions to fellow participants. They are the “do’s and don’ts” that one learns on the ground that may not exist in any written document. In some instances, they may actually be contrary to the “do’s and don’ts” that are written in formal documents. While following a rule may become a “social habit,” it is possible to make participants consciously aware of the rules they use to order their relationships. Individuals can consciously decide to adopt a different rule and change their behavior to conform to such a decision. Over time, behavior in conformance with a new rule may itself become habitual (see Shimanoff, 1980; Toulmin, 1974; Harré, 1974). The capacity of humans to use complex cognitive systems to order their own behavior at a relatively subconscious level makes it difficult for empirical researchers to ascertain what the working rules for an ongoing action arena may be.

In a system governed by a “rule of law,” the general legal framework in use will have its source in actions taken in constitutional, legislative, and administrative settings augmented by decisions taken by individuals in many different particular settings. In other words, the rules-in-form are consistent with the rules-in-use (Sproule-Jones, 1993). In a system that is not governed by a “rule of law,” there may be central laws and considerable effort made to enforce them, but individuals attempt to evade rather than obey the law. Rule-following or conforming actions are not as predictable as biological or physical behavior explained by scientific laws. All rules are formulated in human language. As such, rules share problems of lack of clarity, misunderstanding, and change that typify any language-based phenomenon (Ostrom, 1980, 1997). Words are always simpler than the phenomenon to which they refer.

The stability of rule-ordered actions is dependent upon the shared meaning assigned to words used to formulate a set of rules. If no shared meaning exists when a rule is formulated, confusion will exist about what actions are required, permitted, or forbidden. Regularities in actions cannot result if those who must repeatedly interpret the meaning of a rule within action situations arrive at multiple interpretations. Because “rules are not self-formulating, self-determining, or self-enforcing” (Ostrom, 1980, p. 342), it is human agents who formulate them, apply them in particular situations, and attempt to enforce performance consistent with them. Even if shared meaning exists at the time of the acceptance of a rule, transformations in technology, in shared norms, and in circumstances more generally change the events to which rules apply. “Applying language to changing configurations of development increases the ambiguities and threatens the shared criteria of choice with an erosion of their appropriate meaning” (ibid.).

The stability of rule-ordered relationships is also dependent upon enforcement (Gibson et al., 2005; Dietz et al., 2003). According to John R. Commons, rules “simply say what individuals must, must not, may, can, and cannot do, if the authoritative agency that decides disputes brings the collective power of the community to bear on said individuals” (1957, p. 138). Breaking rules is an option that is always available to participants in an action situation (as contrasted to players in a formal game), but associated with breaking rules is a risk of being monitored and sanctioned. If the risk is low, the predictability and stability of a situation are reduced. And instability can grow over time. If one person can cheat without fear of being caught, others can also cheat with impunity. If the risk of exposure and sanctioning is high, participants can expect that others will make choices from within the set of permitted and required actions.

The simplifying assumption is frequently made in analytical theories that individuals in an action situation will take only those actions that are lawful given the rules that apply. For many purposes, this simplifying assumption helps the analyst examine important theoretical questions not related to how well the rules are enforced. Highly complicated games, such as football, can indeed be explained with more ease because of the presence of active and aggressive on-site referees who constantly monitor the behavior of the players and assign penalties for infraction of rules. And these monitors face real incentives for monitoring consistently and for applying fair and accepted penalties. Both the fans and the managers of the

relevant sports teams may give a lot of attention to what the monitors are doing and the fairness of their judgments. In settings where a heavy investment is *not* made in monitoring the ongoing actions of participants, however, considerable difference between predicted and actual behavior can occur.

This is not to imply that the only reason individuals follow rules is because they are enforced. If individuals voluntarily participate in a situation, they must share some general sense that most of the rules governing the situation are appropriate. Otherwise, the cost of enforcement within voluntary activities becomes high enough that it is difficult, if not impossible, to maintain predictability in an ongoing voluntary activity. One can expect that it is usually difficult to maintain predictability in an ongoing activity where participants do not have the freedom to enter and leave the situation.

What rules are important for institutional analysis? A myriad of specific rules are used in structuring markets, hierarchies, committees, and other structures. Scholars have been trapped into endless cataloging of rules not related to a method of classification most useful for theoretical explanations. But classification is a necessary step in developing a science. Anyone attempting to define a useful typology of rules must be concerned that the classification is more than a method for imposing superficial order onto an extremely large set of seemingly disparate rules. The way we have tackled this problem using the IAD framework is to classify rules according to their impact on the elements of an action situation.

5.5.2 Rule Configurations

A first step toward identifying the working rules can be made, then, by overtly examining how working rules affect each of the components of an action situation listed above. A set of working rules that affect these variables should constitute the minimal but necessary set of rules needed to offer an explanation of actions and results based on the working rules used by participants to order their relationships within an action arena. Because states of the world and their transformations and the nature of a community also affect the structure of an action situation, working rules alone never provide both a necessary and sufficient explanation of the structure of an action situation and results.

Adopting this view of the task, seven groups of working rules can be said to affect the structure of any repetitive action situation, including markets, hierarchies, legislatures, common-property management systems, or competitive sports. These are *boundary rules*, *position rules*, *scope rules*, *authority rules*, *aggregation rules*, *information rules*, and *payoff rules*. These seven groups of rules directly affect the seven components of an action situation.

Boundary rules directly affect the number of *participants*, their attributes and resources, whether they can enter freely, and the conditions they face for leaving. Position rules establish *positions* in the situation. Authority rules assign sets of *actions* that participants in positions at particular nodes must, may, or may not take. Scope rules delimit the *potential outcomes* that can be affected and working

backwards, the actions linked to specific outcomes. Authority rules, combined with the scientific laws about the relevant states of the world being acted upon, determine the shape of the decision tree—the *action–outcome linkages*. Aggregation rules affect the level of *control* that a participant in a position exercises in the selection of an action at a node. Information rules affect the *knowledge-contingent information sets* of participants. Payoff rules affect the *benefits and costs* that will be assigned to particular combinations of actions and outcomes and establish the incentives and deterrents for action. The set of working rules is a *configuration* in the sense that the impact on actions chosen and outcomes of a change in one rule may depend upon the specific content of other rules-in-use.

Table 5.1 presents a simplified set of rules that could be used to constitute an open, competitive market. When a scholar makes the statement, “Let us assume an open, competitive market,” the rules constituting a market are rarely enumerated.

Table 5.1 Rules used to structure open, competitive market situations

Position rules
Positions of owner, seller, buyer, police, suspects, judge, and members of a jury are defined.
Boundary rules
Licensing requirements for individuals to become buyers and sellers are minimal.
Buyers and sellers may enter and exit the market at their own initiative.
Authority rules
Sellers are authorized to decide how many legally owned goods to offer for sale at a price.
Buyers are authorized to decide how much of a commodity to offer to buy at a price.
Police are authorized to arrest those suspected of unlawful use of goods owned by others.
Judges are authorized to determine rights and obligations of buyers and sellers in civil proceedings and of suspected thieves in criminal proceedings.
Members of juries are authorized to determine guilt or innocence of those accused of theft.
Scope rules
Actors are limited in regard to the costs they can externalize on others. (scope rules related to externalities vary substantially from market to market.)
Aggregation rules
Whenever any two actors agree to exchange goods they own, that transaction occurs.
Police may make an arrest after a request or on their own initiative.
Decisions made by a judge must be final unless challenged in a higher court.
Members of a jury must vote before their decisions are official.
Information rules
Prices of current offers to buy and sell must be made available.
No one is authorized to force information from others regarding preferences or costs.
In some jurisdictions, seller may be required to provide specific information on content of goods.
Payoff rules
Seller retains profit, if any, after payment for inputs, interest, and taxes.
Buyer retains consumer surplus, if any, after payment for goods.
Suspects pay fines, or spend time in jail, if judged guilty of criminal acts.
Buyers and/or sellers pay damages and costs to other parties if ordered to do so by judge.

The statement, however, can be thought of as broadly encompassing the type of rules listed in Table 5.1. Substantial changes in any one of these rules affect the structure of a market and the resulting inferences that can be made about equilibria and market performance.

Change of only a few key rules would generate a situation that one would not call a market at all. A change in the aggregation rule, for example, which allocates goods to participants by officials, transforms the resulting action situation into something other than a market. (This is the rule that the military uses to assign uniforms to recruits.) A change in boundary rules affects the number of participants who enter and exit and determines whether the resulting market is competitive, oligopolistic, or monopolistic. By affecting the structure of the market, changes in boundary rules affect predictions concerning the price at which goods will be sold, the quantity to be sold, and the relative distribution of producer and consumer surplus. Anderies (2002) uses these seven types of rules as initial steps in specifying a model of resource appropriation systems such as fisheries, rangelands, and forests. He then shows how changing position and boundary rules affects system dynamics over time.

Some colleagues are surprised at how many rules are needed to constitute a market. By simply manipulating the assumed components of action situations—e.g., the number of participants, the information they possess, the payoffs they face—a scholar using neoclassical economic theory is able to make predictions about probable interactions and outcomes that have considerable empirical support. Thus, for many aspects of theoretical and empirical research, we do not need to dig deeper than a set of assumptions about the structure of a particular type of situation.

When scholars, policy analysts, officials, and citizens try to change the structure of the action situations that exist in a particular firm, community, or region, however, they face a much more demanding task than simply “assuming law and order and an open, competitive market.” They are facing the problem of how to create such a market (or an innovative, efficient firm or a sustainably managed resource). As the cases in Shirley (2002) demonstrate, finding the rule configuration to enhance the performance of urban water systems in regard to efficiency and distribution to the poor residents of a city presents an immense challenge. Privatizing water supplies turns out not to be one policy option. Each of the six major cities studied developed different contractual arrangements as among producers, regulators, consumers, and citizens. The performance of the reformed systems varied substantially across measures as well as across cities.

Changing the rules of a planned economy so as to constitute effective open markets has also proved to be a major challenge since the fall of the Berlin Wall. Policies intended to create open markets in many “transitional economies” have constituted monopolies or oligopolies instead! Given the recent experiences of failed efforts to improve the economies of the developing countries and many of those of Eastern Europe, the importance of understanding the deeper levels of institutional analysis is becoming more obvious. Further, as empirical studies of the linkages between rule configurations and action situations have shown, the actual number of rules-in-use in many typical situations is immense. At least 27 different boundary rules are in use in different locations in the management of common-pool resources,

for example (Ostrom, 1999a, p. 511). Similarly, a large number of authority rules, information rules, and payoff rules have been identified. The set of identified rules is large enough that no analyst can perform a complete analysis of the structure of resulting situations devoid from all combinations of these rules.

Consequently, the next frontier in doing institutional analysis is a much more self-conscious study of how rules combine with one another, as well as with the physical and material conditions and the nature of a community to create the situations in which participants interact. Janssen (2002) has taken major steps in opening that frontier with his analysis of lessons learned from languages and from immune systems for the study of resilient human-ecological systems. The use of agent-based models will enable scholars to tackle the study of complex adaptive systems in a coherent and cumulative manner.

5.5.3 Physical and Material Conditions

While a rule configuration affects all of the elements of an action situation, some of the variables of an action situation (and thus the overall set of incentives facing individuals in a situation) are also affected by attributes of the physical and material world. What actions are physically possible, what outcomes can be produced, how actions are linked to outcomes, and what is contained in the actors' information sets are affected by the world being acted upon in a situation. The same set of rules may yield entirely different types of action situations depending upon the types of events in the world being acted upon by participants. These "events" are frequently referred to as the "goods and services" being produced, consumed, and allocated in a situation as well as the technology available for these processes.

The attributes of states of the world and their transformation are explicitly examined when the analyst self-consciously asks a series of questions about how the world being acted upon in a situation affects the outcome, action sets, action–outcome linkages, and information sets in that situation. The relative importance of the rule configuration and states of the world in structuring an action situation varies dramatically across different types of settings. The rule configuration almost totally constitutes some games, like chess, where physical attributes are relatively unimportant. The relative importance of working rules to attributes of the world also varies dramatically within action situations considered to be part of the public sector. Rules define and constrain voting behavior inside a legislature more than attributes of the world. Voting can be accomplished by raising hands, by paper ballots, by calling for the ayes and nays, by marching before an official counter, or by installing computer terminals for each legislator on which votes are registered. However, in regard to organizing communication within a legislature, attributes of the world strongly affect the available options. The principle that only one person can be heard and understood at a time in any one forum strongly affects the capacity of legislators to communicate effectively with one another (see Ostrom, 1987).

Considerable academic literature has focused on the effect of attributes of goods on the results obtained within an action situation. A key assumption made in the

analysis of a competitive market is that the outcomes of an exchange are highly excludable, easily divisible and transferable, and internalized by those who participate in the exchange. Markets are predicted to fail as effective decision mechanisms when they are the only arena available for producing, consuming, or allocating a wide variety of goods that fail to meet the criteria of excludability, divisibility, and transferability. Market failure means that the incentives facing individuals in a situation where the rules are those of a competitive market, but the goods are not “private goods,” do not motivate individuals to produce, allocate, and consume those goods close to an optimal level.

Let us briefly consider here several attributes that are frequently used to distinguish goods and services that are more effectively provided by a variety of non-market rules. A full elaboration of attributes of goods cannot be developed in this chapter given space limitations (see Ostrom et al., 1994). Goods that are generally considered to be “public goods” yield nonsubtractive benefits that can be enjoyed jointly and simultaneously by many people who are hard to exclude from obtaining these benefits. Common-pool resources yield benefits where beneficiaries are hard to exclude, but each person’s use of a resource system subtracts units of that resource from a finite total available for harvesting.

5.5.4 Excludability and the Free-Rider Problem

When it is difficult or costly to exclude beneficiaries from a good once it is produced, it is frequently assumed that such a good must be provided publicly, rather than privately. When the benefits of a good are available to a group, whether or not members of the group contribute to the provision of the good, that good is characterized by problems with excludability. Where exclusion is costly, those wishing to provide a good or service face a potential free-rider or collective-action problem (Olson, 1965). Individuals who gain from the maintenance of an irrigation system, for example, may not wish to contribute labor or taxes to maintenance activities, hoping that others will bear the burden. This is not to say that all individuals will free-ride whenever they can. A strong incentive exists to be a free-rider in all situations where potential beneficiaries cannot easily be excluded for failing to contribute to the provision of a good or service.

When it is costly to exclude individuals from enjoying benefits from a common-pool resource or an infrastructure facility, private, profit-seeking entrepreneurs, who must recoup their investments through quid pro quo exchanges, have few incentives to provide such services on their own initiative. Excludability problems can thus lead to the problem of free-riding, which in turn leads to underinvestment in capital and its maintenance.

Public sector provision of common-pool resources or infrastructure facilities raises additional problems in determining preferences and organizing finances. When exclusion is low-cost to the supplier, preferences are revealed as a result of many quid pro quo transactions. Producers learn about preferences through the consumers’ willingness to pay for various goods offered for sale. Where exclusion

is difficult, designing mechanisms that honestly reflect beneficiaries' preferences and their willingness to pay is complex, regardless of whether the providing unit is organized in the public or the private sphere. In very small groups, those affected are usually able to discuss their preferences and constraints on a face-to-face basis and to reach a rough consensus. In larger groups, decisions about infrastructure are apt to be made through mechanisms such as voting or the delegation of authority to public officials. The extensive literature on voting systems demonstrates how difficult it is to translate individual preferences into collective choices that adequately reflect individual views (Arrow, 1951; Shepsle, 1979; Buchanan & Tullock, 1962).

Another attribute of some goods with excludability problems is that, once they are provided, consumers may have no choice whatsoever as to whether they will consume. An example is the public spraying of insects. If an individual does not want this public service to be provided, there are even stronger incentives not to comply with a general tax levy. Thus, compliance with a broad financing instrument may, in turn, depend upon the legitimacy of the public-choice mechanism used to make provision decisions.

5.5.5 Subtractability of the Flow

Jointly used infrastructure facilities can generate a flow of services that is entirely subtractable upon consumption by one user; in other instances, consumption by one does not subtract from the flow of services available to others. The withdrawal of a quantity of water from an irrigation canal by one farmer means that there is that much less water for anyone else to use. Most agricultural uses of water are fully subtractive, whereas many other uses of water—such as for power generation or navigation—are not. Most of the water that passes through a turbine to generate power, for instance, can be used again downstream. When the use of a flow of services by one individual subtracts from what is available to others and when the flow is scarce relative to demand, users will be tempted to try to obtain as much as they can of the flow for fear that it will not be available later.

Effective rules are required if scarce, fully subtractive service flows are to be allocated in a productive way. Charging prices for subtractive services obviously constitutes one such allocation mechanism. Sometimes, however, it is not feasible to price services. In these instances, some individuals will be able to grab considerably more of the subtractive services than others, thereby leading to noneconomic uses of the flow and high levels of conflict among users.

Allocation rules also affect the incentives of users to maintain a system. Farmers located at the tail end of an irrigation system that lacks effective allocation rules have little motivation to contribute to the maintenance of that system because they only occasionally receive their share of water. Similarly, farmers located at the head end of such a system are not motivated to provide maintenance services voluntarily because they will receive disproportionate shares of the water whether or not the system is well maintained (Ostrom, 1996).

Consequently, for common-pool resources whose flows are highly subtractive, institutional arrangements related to the allocation of the flow of services are intimately tied to the sustainability of the resource. It is highly unlikely that one can achieve sustainability without careful attention to the efficiency, fairness, and enforceability of the rules specifying who can appropriate how much of the service flow, at what times and places, and under what conditions. Furthermore, unless responsibilities are linked in a reasonable fashion to benefits obtained, the beneficiaries themselves will resist efforts to insist that they take responsibilities.

5.5.6 Additional Attributes

In addition to these general attributes of physical and material conditions that affect the incentives of participants, resource systems are also characterized by a diversity of other attributes that affect how rules combine with physical and material conditions to generate positive or negative incentives. Whether resource units are *mobile* or *stationary* and whether *storage* is available somewhere in a system affect the problems that individuals governing and managing common-pool resources face (Schlager et al., 1994). The problems of regulating a lobster fishery, for example, are much simpler than those of regulating a salmon fishery. Similarly, allocating water in a predictable and efficient manner is easier to achieve when there is some storage in the system than when it is a run-of-the-river system.

If a natural resource system is renewable, such as many groundwater basins, the relevant time horizon for sustaining use is very long, and achieving appropriate rules may mean the difference between creating a sustainable conjunctive use system and destroying a groundwater basin. Devising an effective set of rules for regulating the use of an oil pool, on the other hand, involves determining an optimal path for mining a resource. The cost of withdrawing the last units of oil will be much higher if producers have not coordinated their withdrawal patterns. Since the oil field will eventually be used up, the lack of a future may not generate the incentives needed to achieve adequate regulation early in the development phase (Libecap, 1978).

The size of a resource system can also have a major impact on the incentives facing participants. The length and slope of a main canal of an irrigation system not only affects the cost of its maintenance but also the strategic bargaining that exists between headenders and tailenders on an irrigation system (Ostrom, 1996). Increasing the number of participants is associated with increased transaction costs. How steeply the costs rise depends, to a large extent, on the rules-in-use and the heterogeneity of the users.

The productivity, predictability, and patchiness of a resource affect the likelihood that private property arrangements will be successful and enhances the likelihood that common property arrangements will be necessary (Netting, 1982). Similarly, the resilience of a multispecies ecosystem affects the sensitivity of the system to both the rules used to govern the particular system and to changes in economic or environmental conditions elsewhere (Holling, 1994). These additional attributes are slowly being integrated into a body of coherent theory about the impact of physical and

material conditions on the structure of the situations that individuals face and their resulting incentives and behavior. Analysts diagnosing resource problems need to be sensitive to the very large difference among resource settings and the need to tailor rules to diverse combinations of attributes rather than some assumed uniformity across all resources in a particular sector within a country.

5.5.7 Attributes of the Community

A third set of variables that affects the structure of an action arena relates to the community. The attributes of a community that are important in affecting the structure of an action arena include the norms of behavior generally accepted in the community, the level of common understanding potential participants share about the structure of particular types of action arenas, the extent of homogeneity in the preferences of those living in a community, and the distribution of resources among those affected. The term *culture* is frequently applied to this bundle of variables.

For example, when all appropriators from a common-pool resource share a common set of values and interact with one another in a multiplex set of arrangements, the probabilities of their developing adequate rules and norms to govern resources are much greater (Taylor, 1987). The importance of building a reputation for keeping one's word is important in such a community, and the cost of developing monitoring and sanctioning mechanisms is relatively low. If the appropriators from a resource come from many different communities and are distrustful of one another, the task of devising and sustaining effective rules is substantially increased.

Whether individuals use a written vernacular language to express their ideas, develop common understanding, share learning, and explain the foundation of their social order is also a crucial variable of relevance to institutional analysis (Ostrom, 1997). Without a written vernacular language, individuals face considerably more difficulties in accumulating their own learning in a usable form to transmit from one generation to the next.

5.6 Linking Action Arenas

In addition to analysis that digs deeper into the factors affecting individual action arenas, an important development in institutional analysis is the examination of linked arenas (see McGinnis & Williams, 2000). Whereas the concept of a "single" arena may include large numbers of participants and complex chains of action, most of social reality is composed of multiple arenas linked sequentially or simultaneously (Hooghe & Marks, 2001).

When individuals wish to intervene to change the structure of incentives and deterrents faced by participants in socially constructed realities to guide (or control) participants toward a different pattern of results, they do so by attempting to change the rules individuals use to order their interactions within particular types of action arenas. Some interesting and important institutional arrangements for coordinating

complex chains of actions among large numbers of actors involve multiple organizations competing with one another according to a set of rules. Markets are the most frequently studied institutional arrangements that achieve coordination by relying primarily on rule-governed, competitive relationships among organizations. Rule-governed competition among two or more political parties is considered by many analysts to be an important requisite for a democratic polity. Less studied, but potentially as important a means for achieving responsiveness and efficiency in producing public goods and services, are arrangements that allow rule-ordered competition among two or more potential *producers* of public goods and services.

All rules are nested in another set of rules that define how the first set of rules can be changed. The nesting of rules within rules at several levels is similar to the nesting of computer languages at several levels. What can be done at a higher level will depend on the capabilities and limits of the rules at that level and at a deeper level. Whenever one addresses questions about *institutional change*, as contrasted to action within institutional constraints, it is necessary to recognize the following:

1. Changes in the rules used to order action at one level occur within a currently “fixed” set of rules at a deeper level.
2. Changes in deeper-level rules usually are more difficult and more costly to accomplish, thus increasing the stability of mutual expectations among individuals interacting according to a set of rules.

As discussed briefly above, it is useful to distinguish three levels of rules that cumulatively affect the actions taken and outcomes obtained in any operational setting (Kiser & Ostrom, 1982). *Operational rules* directly affect day-to-day decisions made by the participants in any setting. *Collective-choice rules* affect operational activities and results through their effects in determining who is eligible and the specific rules to be used in changing operational rules. *Constitutional-choice rules* affect operational activities and their effects in determining who is eligible and the rules to be used in crafting the set of collective-choice rules that in turn affect the set of operational rules. One can even think about a “meta-constitutional” level underlying all the others that is not frequently analyzed. One can think of the linkages among these rules and related level of analysis as shown in Fig. 5.3.

For most practical applications, three or four levels are enough, but there is no theoretical justification for any specific number of levels. For the purposes of formal theory, we may need to assume as long a series of layers as is needed until we hit rock bottom—the physical world. Very deep layering—even infinite layering—turns out to be needed in many aspects of formal theory. Game theorists, for example, have had to assume that the common knowledge needed for one to assume that there is a game is nested infinitely. “Information is *common knowledge* if it is known to all players, each player knows that all of them know it, and each of them knows that all of them know that all of them know it, and so forth ad infinitum” (Rasmusen, 1989, p. 50). Thus, one can always assume that there are even more primitive rules underlying those that one is analyzing at any one level—thus our positing of a very general meta-constitutional level—until one gets to the constraints of a physical world.

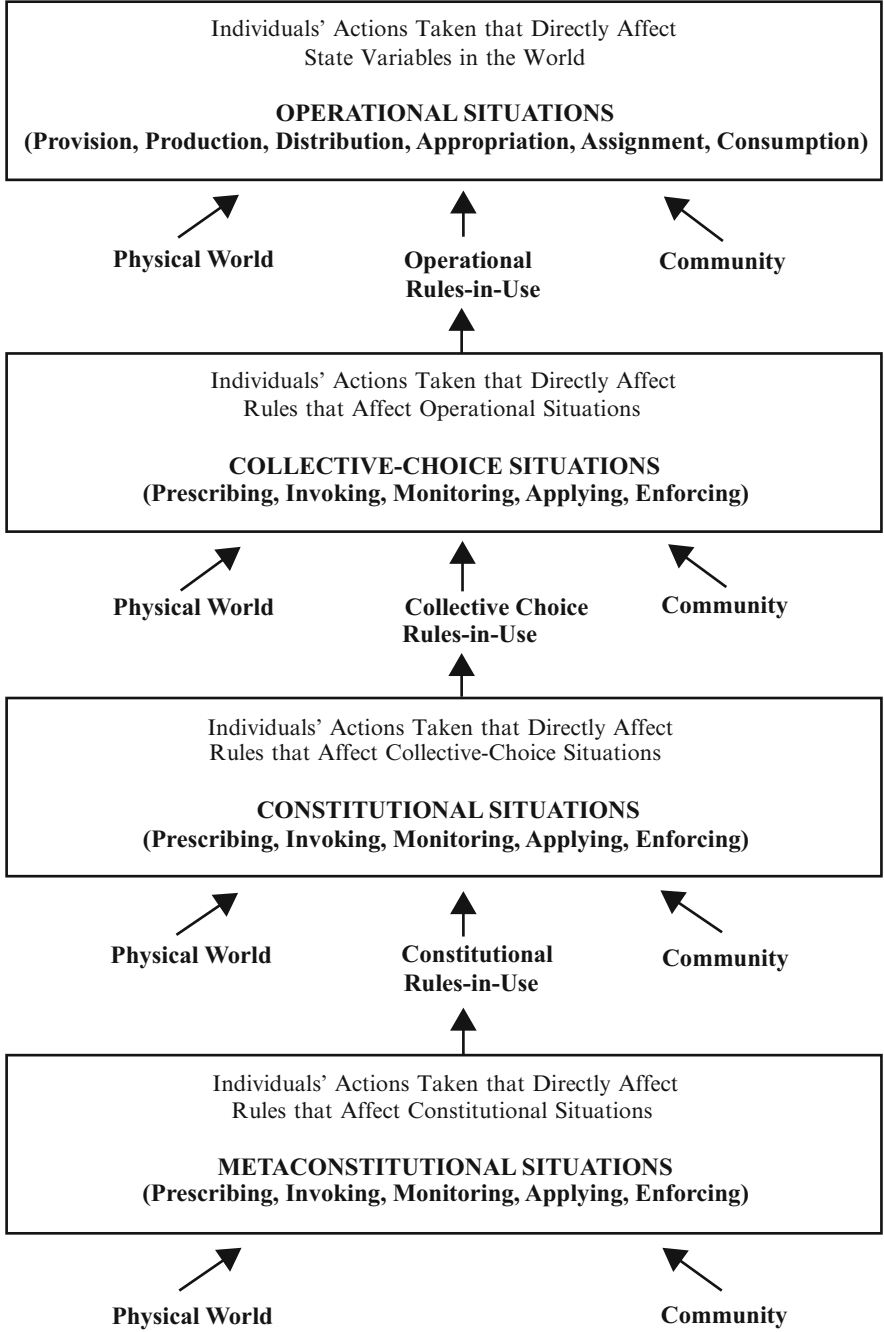


Fig. 5.3 Levels of analysis and outcomes

The participants in collective-choice games may be the same participants as in linked operational-choice games (as when all firms in an industry agree upon a particular industry standard that they will all use in manufacturing goods). Or, participants in collective-choice games may differ from those in operational games. They may, for example, be legislative representatives selected in electoral games (themselves part of the collective-choice level of action) to be the agents of a set of principals—the citizens engaged in a wide diversity of operational games that will be affected over time by legislative policies. Participants in the third level can, again, either be participants in the other two levels or not. And participants in constitutional choices may not recognize that they are making a constitutional rule—they may be simply trying to fix a problem with the way that they have been making policy choices over the last several years.

At each level of analysis, there may be one or more arenas in which the types of decisions made at that level will occur. In the collective-choice, constitutional, and meta-constitutional situations, activities involve prescribing, invoking, monitoring, applying, and enforcing rules (Lasswell & Kaplan, 1950; Oakerson, 1994). The concept of an “arena” as described earlier does not imply a formal setting but can include such formal settings as legislatures and courts. Policymaking (or governance) regarding the rules that will be used to regulate operational-level choices is usually carried out in one or more collective-choice arenas as shown in Fig. 5.3.

5.7 Conclusion

When a theorist chooses to analyze a situation at any particular level, he or she must assume that the institutional rules at that level are temporarily fixed for the purpose of analysis. These rules form a part of the *structure* of the situation rather than the *solution* to the game created by that structure. When the purpose of analysis is to understand the origin of the rules at one level, knowing the structure of the situation at the next higher level is essential for that enterprise. The equilibria achieved at one level are thus supported by equilibria that have been achieved at higher levels.⁸ Understanding the role of these nested levels does not, however, require that the analyst specify the full supporting infrastructure in elucidating how individuals are expected to behave at one level. Thus, assuming that there are multiple levels where decisions are made that affect actions at other levels actually greatly simplifies analysis rather than complicating it.

Some situations within any one of these levels may be simple enough that one can generate a clear and empirically supported prediction about outcomes. One can do

⁸The decisions reached by members of a legislature are not stable if opponents of these decisions win a majority of seats in future elections on promises that they will reverse earlier decisions or change the constitutions so as to make the decisions unconstitutional. The stability of decisions in complex modern institutions is dependent not only upon the preferences and procedures used to organize decision-making in one arena but upon the entire nested set of arenas (Shepsle, 1989).

this, for example, in a highly competitive market producing goods characterized by low costs of exclusion and subtractability. Here, one can rely on well-tested results from prior theoretical and empirical work. It is usually much more difficult to predict results when one is analyzing a collective-choice or constitutional-choice situation as it impacts on an operational-level setting. Existing theoretical results are frequently not available for predicting results in an operational situation whose rules are being changed at a collective-choice level. When new and unanalyzed situations are created by the process of changing parts of a rule configuration, institutional analysis needs to proceed to undertake a deeper analysis of how participants view the new rules, how they come to understand them, how they will be monitored and enforced, and what types of individual actions and collective outcomes are produced. This is frequently a challenging, difficult, and complex theoretical and empirical task.

Acknowledgments The author is appreciative of the support of the “Institutions Project” of the Resilience Alliance and the comments of two reviewers of an earlier draft of this manuscript. Patty Lezotte has once again done a wonderful job of editing the manuscript.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Agrawal, A. (1998). *Greener pastures: Politics, markets, and community among a migrant pastoral people*. Duke University Press.
- Anderies, J. M. (2002). The transition from local to global dynamics: A proposed framework for agent-based thinking in socio-ecological systems. In M. A. Janssen (Ed.), *Complexity and ecosystem management: The theory and practice of multi-agent systems* (pp. 13–34). Cheltenham.
- Andersson, K. (2002). *Can decentralization save Bolivia's forests? An institutional analysis of municipal forest governance*. Ph.D. dissertation, Indiana University, Bloomington.
- Arrow, K. (1951). *Social choice and individual values* (2nd ed.). Wiley.
- Black, M. (1962). *Models and metaphors*. Cornell University Press.
- Blomquist, W. (1992). *Dividing the waters: Governing groundwater in Southern California*. ICS Press.
- Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge University Press.
- Buchanan, J. M., & Tullock, G. (1962). *The calculus of consent*. University of Michigan Press.
- Commons, J. R. (1957). *Legal foundations of capitalism*. University of Wisconsin Press.
- Crawford, S. E. S., & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review*, 89(3), 582–600.
- Dietz, T., Ostrom, E., & Stern, P. (2003). The struggle to govern the Commons. *Science*, 302(5652), 1907–1912.
- DiMaggio, P., & Powell, W. (1983). The iron cage revisited. *American Sociological Review*, 48(2), 147–160.
- Firmin-Sellers, K. (1996). *The transformation of property rights in the Gold Coast: An empirical study applying rational choice theory*. Cambridge University Press.
- Fligstein, N. (1998). *Fields, power, and social skill: A critical analysis of the new institutionalisms* (Working paper). University of California.
- Ganz, J. S. (1971). *Rules: A systematic study*. Mouton.

- Gibson, C. (1999). *Politicians and poachers: The political economy of wildlife policy in Africa*. Cambridge University Press.
- Gibson, C., Williams, J., & Ostrom, E. (2005). Local enforcement and better forests. *World Development*, 33, 273–284.
- Harré, R. (1974). Some remarks on ‘rule’ as a scientific concept. In T. Mischel (Ed.), *Understanding other persons* (pp. 143–184). Basil Blackwell.
- Holling, C. S. (1994). An ecologist view of the Malthusian conflict. In K. Lindahl-Kiessling & H. Landberg (Eds.), *Population, economic development, and the environment* (pp. 79–103). Oxford University Press.
- Hooghe, L., & Marks, G. (2001). *Multi-level governance and European integration*. Rowman & Littlefield.
- Janssen, M. A. (2002). Changing the rules of the game: Lessons from immunology and linguistics for self-organization of institutions. In M. A. Janssen (Ed.), *Complexity and ecosystem management: The theory and practice of multi-agent systems* (pp. 35–47). Cheltenham.
- Kaminski, A. (1992). *An institutional theory of communist regimes: Design, function, and breakdown*. ICS Press.
- Kiser, L. L., & Ostrom, E. (1982). The three worlds of action: A metatheoretical synthesis of institutional approaches. In E. Ostrom (Ed.), *Strategies of political inquiry* (pp. 179–222). Sage.
- Lam, W. F. (1998). *Governing irrigation systems in Nepal: Institutions, infrastructure, and collective action*. ICS Press.
- Lasswell, H., & Kaplan, A. (1950). *Power and society: A framework for political inquiry*. Yale University Press.
- Libecap, G. D. (1978). Economic variables and the development of law: The case of Western mineral rights. *Journal of Economic History*, 38(2), 338–362.
- Loveman, B. (1993). *The constitution of tyranny: Regimes of exception in Spanish America*. University of Pittsburgh Press.
- Mayntz, R., & Scharpf, F. W. (1995). *Steuerung und Selbstorganisation in staatsnahen Sektoren*. Campus-Verlag.
- McGinnis, M. (Ed.). (1999). *Polycentricity and local public economies: Readings from the workshop in political theory and policy analysis*. University of Michigan Press.
- McGinnis, M., & Williams, J. (2000). Policy uncertainty in two-level games: Examples of correlated equilibria. In M. McGinnis (Ed.), *Polycentric games and institutions: Readings from the workshop in political theory and policy analysis* (pp. 202–236). University of Michigan Press.
- Mwangi, E. (2003). *Institutional change and politics: The transformation of property rights in Kenya's Maasailand*. Ph.D. dissertation, Indiana University, Bloomington.
- National Research Council. (2002). Committee on the human dimensions of global change. In E. Ostrom, T. Dietz, N. Dolšák, P. C. Stern, S. Stonich, & E. Weber (Eds.), *The drama of the commons*. National Research Council, National Academy Press.
- Netting, R. M. C. (1982). Territory, property, and tenure. In R. M. Adams, N. J. Smelser, & D. J. Treiman (Eds.), *Behavioral and social science research: A national resource* (pp. 446–501). National Academy Press.
- North, D. C. (1990). *Institutions, institutional change, and economic performance*. Cambridge University Press.
- Oakerson, R. J. (1992). Analyzing the commons: A framework. In D. W. Bromley et al. (Eds.), *Making the commons work: Theory, practice, and policy* (pp. 41–59). ICS Press.
- Oakerson, R. J. (1994). *The logic of multi-level institutional analysis*. Paper presented at the Workshop on the Workshop conference, Indiana University, Workshop in political theory and policy analysis, Bloomington, 15–19 June 1994.
- Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Harvard University Press.
- Ostrom, V. (1964). Editorial comment: Developments in the ‘no-name’ fields of public administration. *Public Administration Review*, 24(1), 62–63.
- Ostrom, V. (1980). Artisanship and artifact. *Public Administration Review*, 40(4), 309–317.

- Ostrom, E. (1986). An agenda for the study of institutions. *Public Choice*, 48, 3–25.
- Ostrom, V. (1987). *The political theory of a compound republic: Designing the American experiment* (2nd ed.). University of Nebraska Press.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
- Ostrom, V. (1991). *The meaning of American federalism: Constituting a self-governing society*. ICS Press.
- Ostrom, E. (1992). The rudiments of a theory of the origins, survival, and performance of common-property institutions. In D. W. Bromley et al. (Eds.), *Making the commons work: Theory, practice, and policy* (pp. 41–59). ICS Press.
- Ostrom, E. (1996). Incentives, rules of the game, and development. In *Proceedings of the Annual World Bank conference on development economics 1995*. The World Bank, Washington, DC, pp. 207–234.
- Ostrom, V. (1997). *The meaning of democracy and the vulnerability of democracies: A response to Tocqueville's challenge*. University of Michigan Press.
- Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action. *American Political Science Review*, 92(1), 1–22.
- Ostrom, E. (1999a). Coping with tragedies of the commons. *Annual Review in Political Science*, 2, 493–535.
- Ostrom, E. (1999b). Institutional rational choice: An assessment of the institutional analysis and development framework. In P. A. Sabatier (Ed.), *Theories of the policy process* (pp. 35–71). Westview Press.
- Ostrom, E., Gardner, R., & Walker, J. (1994). *Rules, games, and common-pool resources*. University of Michigan Press.
- Polski, M. M. (2003). *The invisible hands of U.S. commercial banking reform: Private action and public guarantees*. Kluwer.
- Rapoport, A. (1966). *Two-person game theory: The essential ideas*. University of Michigan Press.
- Rasmusen, E. (1989). *Games and information: An introduction to game theory*. Basil Blackwell.
- Sawyer, A. (1992). *The emergence of autocracy in Liberia: Tragedy and challenge*. ICS Press.
- Scharpf, F. W. (1997). *Games real actors play: Actor-centered institutionalism in policy research*. Westview Press.
- Schlager, E. (1990). *Model specification and policy analysis: The governance of coastal fisheries*. Ph.D. dissertation, Indiana University, Bloomington.
- Schlager, E., Blomquist, W., & Tang, S. Y. (1994). Mobile flows, storage, and self-organized institutions for governing common-pool resources. *Land Economics*, 70(3), 294–317.
- Scott, W. R. (2001). *Institutions and organizations* (2nd ed.). Sage.
- Seo, M.-G., & Creed, W. E. D. (2002). Institutional contradictions, praxis, and institutional change: A dialectical perspective. *Academy of Management Review*, 27(2), 222–247.
- Shepsle, K. A. (1979). Institutional arrangements and equilibrium in multidimensional voting models. *American Journal of Political Science*, 23, 27–59.
- Shepsle, K. A. (1989). Studying institutions: Some lessons from the rational choice approach. *Journal of Theoretical Politics*, 1(2), 131–147.
- Shimanoff, S. B. (1980). *Communication rules: Theory and research*. Sage.
- Shirley, M. M. (2002). *Thirsting for efficiency: The economics and politics of urban water system reform*. Pergamon.
- Sproule-Jones, M. (1993). *Governments at work: Canadian parliamentary federalism and its public policy effects*. University of Toronto Press.
- Tang, S. Y. (1992). *Institutions and collective action: Self-governance in irrigation*. ICS Press.
- Taylor, M. (1987). *The possibility of cooperation*. Cambridge University Press.
- Toulmin, S. (1974). Rules and their relevance for understanding human behavior. In T. Mischel (Ed.), *Understanding other persons* (pp. 185–215). Basil Blackwell.
- Walker, S. T. (1998). *Both pretense and promise: The political economy of privatization in Africa*. Ph.D. dissertation, Indiana University, Bloomington.

- Weissing, F. J., & Ostrom, E. (1991). Irrigation institutions and the games irrigators play: Rule enforcement without guards. In R. Selten (Ed.), *Game equilibrium models II: Methods, morals, and markets* (pp. 188–262). Springer.
- Williamson, O. E. (1991). Comparative economic organization: The analysis of discrete structural alternatives. *Administrative Science Quarterly*, 36(2), 269–296.
- Wynne, S. G. (1989). *The land boards of Botswana: A problem in institutional design*. Ph.D. dissertation, Indiana University, Bloomington.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part II

Political Institutions

Electoral Institutions and Political Competition: Coordination, Persuasion, and Mobilization

6

Gary W. Cox

Contents

6.1	Introduction	106
6.2	Electoral Rules	107
6.2.1	Translating Votes into Seats	108
6.2.2	The Number of Seats per District, Tier, and Segment	109
6.2.3	The Number of Votes per Voter	110
6.2.4	Summary	111
6.3	Electoral Coordination	112
6.3.1	Coordination of Votes Within Districts	112
6.3.2	The “M + 1 Rule”: Theory	112
6.3.3	The “M + 1 Rule”: Evidence	113
6.3.4	Summary	115
6.4	Coordination of Elites Within Districts	115
6.4.1	Strategic Contributions	115
6.4.2	Optimal Nomination	116
6.5	Coordinating Across Districts	117
6.6	Electoral Persuasion	119
6.6.1	Office-Seeking Competitors with Credible Promises	119
6.6.2	Office-Seeking Competitors Without Credible Promises	120
6.6.3	Rent-Seeking Competitors with Credible Promises	120
6.7	Mobilization	121
6.8	Conclusion	122
6.8.1	Coordination	122
6.8.2	Persuasion	123
6.8.3	Mobilization	123
	References	123

Abstract

How does changing the rules of electoral competition change parties’ and candidates’ strategies? This chapter considers three kinds of rule changes—those that

G. W. Cox (✉)

Department of Political Science, Stanford University, Stanford, CA, USA

e-mail: gwcox@stanford.edu

increase the effective number of votes per voter, the effective number of seats per district, and the proportionality of the votes-to-seats mapping(s)—and focuses on three types of campaign strategy: those that address the challenges of coordinating, persuading and mobilizing voters.

Keywords

Electoral coordination · Spatial competition · Strategic mobilization ·
M + 1 rule · Cross-district coordination

6.1 Introduction

In the Schumpeterian conception, democracy consists of regular and nonviolent competition for control of government between alternative teams of elites (Schumpeter, 1942). The question that much scholarship in electoral studies addresses, and on which this chapter will focus, is: how does changing the rules of the electoral game change the strategies of parties and candidates, hence the outcome of elections?

Figure 6.1 illustrates both the sequence of events in a stylized democracy and some of the topics to be covered. In the beginning, there is a set of potential electoral competitors. These agents decide (at stage 1 of the diagram) whether to enter a particular electoral competition—that is, to formally nominate candidates for one or more elective offices. Since winning office requires amassing a sufficient number of votes, the nature of the entry game between potential competitors has a coordination game flavor to it. For example, if 15 right-of-center parties all enter the race as separate competitors, while the left unites behind a single option, the right is likely to do poorly (under most extant electoral systems). The right can do better if some potential competitors withdraw, but each potential competitor will likely prefer that *it* remain and *the others* withdraw.

n potential competitors



(1) coordination of entry.

$m \leq n$ actual competitors in the election(s)



(2.1) persuasion; (2.2) coordination of votes; (2.3) mobilization.

vote distribution across competitors



(3) mechanical translation of votes into seats.

seat distribution across competitors



(4) government formation process.

portfolio distribution across competitors

Fig. 6.1 The office-seeking sequence in a hypothetical democracy

After a given set of competitors have entered the race, each decides to allocate effort to one or more of three vote-producing activities: (2.1) electoral persuasion (providing voters with reasons, such as better policy positions or larger bribes, to prefer it to the other competitors); (2.2) coordination of votes (convincing supporters of other parties that the expected utility of their vote, in terms of affecting the allocation of seats across competitors, will be higher if they support it than if they support their most-preferred competitor); and (2.3) strategic mobilization (boosting the probability that its known supporters will actually participate in the election). Each of these vote-producing activities is cost-effective under somewhat different conditions.

After the election has been held, an allocation of votes across the available competitors is determined. This allocation of votes is translated into an allocation of seats by a series of deterministic mathematical operations mandated by the relevant electoral rules (in particular, the electoral “formulas”) of a system. Finally, after the allocation of offices has been determined, those competitors who hold seats in the national legislature can bargain among themselves over the distribution of portfolios (cabinet ministerial posts).

In this chapter, I focus on the first three stages of Fig. 6.1, leaving government formation to others. In order to simplify the exposition, I do not deal much with the detailed electoral rules, instead categorizing electoral systems by three broad architectonic features: the number of votes per voter, the number of seats per district, and the proportionality of the votes-to-seats translation. These features play the role of independent variables, with candidates’ and parties’ strategies of coordination, persuasion, and mobilization in the role of dependent (or sometimes intermediate) variables.¹

Although not the focus of this chapter, it may be worth suggesting how parties’ strategies in turn affect policy choices. A short answer is that parties’ strategies help define the sort of actor that will be in government. In particular, the larger are the electoral aggregates that form (the greater is the equilibrium level of coordination), the broader are the interests those aggregates will represent, and the more that parties choose to persuade via promises to provide differing packages of public goods, rather than differing packages of private goods, the greater the pressure will be on them to deliver such goods when in office.²

6.2 Electoral Rules

There is an immense range of different electoral rules that can be combined into an even larger number of possible electoral systems. Here, I shall focus on the three main components of any electoral system: votes, seats, and the rules translating votes

¹ Throughout, I focus (albeit not exclusively) on formally derived institutional comparative statics results. For other reviews similar in spirit, see Myerson (1995, 1999).

² More elaborate answers along these lines can be found in Shugart (2001), Cox and McCubbins (2001), and Tsebelis (2002).

into seats. Each of these components can be characterized in detail, by describing all the features of the particular electoral system at hand. Each can also be characterized in a more summary fashion—the approach taken here.

6.2.1 Translating Votes into Seats

One can characterize a given electoral system in terms of *where* and *how* votes are converted into seats. As far as *where* is concerned, votes can be translated into seats within electoral districts, within upper tiers and within electoral segments. Let's consider each of these in turn.

An *electoral district* is a geographical area within which votes are counted and seats awarded. For example, the US House of Representatives has 435 districts, each returning a single member. Votes are counted at the district level, and seats are awarded at the district level. There is no process by which votes from an election in California's fourth district can affect the outcome of the election in New York's third or California's fifth.

Other electoral systems have provisions by which votes and sometimes seats are translated from their district of origin to a secondary district (a larger geographic area encompassing two or more districts), with an additional round of votes-to-seats conversions occurring at that level. For example, one could modify the US system by stipulating that all votes not cast for a winning candidate in their district of origin transfer to the national level, where they can translate into seats according to stipulated rules.

Many electoral systems have *upper tiers* of secondary districts placed over their primary districts. In Belgium as of 1960, for example, votes were cast within 30 *arrondissements* (primary districts) which were in turn grouped into 9 provinces (secondary districts). A party's votes in a given *arrondissement* might suffice to "buy" one or more seats at that level, where the "price" in votes per seat was the so-called Hare quota (equal roughly to the total number of votes cast divided by M). Any unused votes cast for a party in a given *arrondissement* transferred to the provincial level, where they combined with the party's unused votes from the other *arrondissements* in the province, and might then suffice to "buy" some seats at the provincial level. The purpose of layering an upper tier of provinces over a lower tier of *arrondissements* was to increase the proportionality of the overall system (on which more below).

In addition to layering secondary districts over primary districts, another technique to affect overall proportionality that has become increasingly popular is to have two parallel segments of districts. An *electoral segment* consists of a set of electoral districts that together cover the entire nation, along with any associated upper tiers. The typical two-segment electoral designs combine one segment of single-member districts with one segment of multimember districts operating under proportional representation (see Shugart and Wattenberg (2001)). In Japan, for example, the 1993 electoral reform created 1 segment of 300 single-member

districts (which together covered the entire nation) and 1 segment of 11 multimember districts (which also collectively covered the nation) (see Reed and Thies (2001)).

A full description of where votes are converted into seats thus involves a stipulation first of the number of segments (one or two) and then a characterization of each segment. In addition, one would wish to specify the procedures by which seats are apportioned to districts and tiers, as well as the procedures (if any) by which district boundaries can be redrawn. Here, however, the structure of districts, tiers, and segments will be reduced to a single number (the effective magnitude; see the next section) and the procedures by which reapportionment and redistricting occur will be left in the background.

In addition to knowing *where* votes are converted into seats, one would also wish to know *how*. Particular mathematical formulas are used at each stage of seat allocation in an electoral system. In Belgium, for example, the electoral code stipulates a “price” for each seat (the Hare quota), and each party in a given *arrondissement* wins as many seats as its votes can buy at the stipulated price. A somewhat different formula (a different way of deciding the price per seat) is then used at the provincial level to award seats at that stage.

Individual electoral formulas can be arrayed along a continuum indicating their *proportionality*. At one extreme are the winner-take-all formulas: the candidate or list with the most votes wins all the seats available to be won in the particular district. Such formulas provide strong incentives to form “large” coalitions, in order to be able to compete seriously for the seats at stake. Opposed to these winner-take-all systems are those in which a list receives seats in strict proportion to its votes. Here, the larger parties receive no bonus seats (seats in excess of what they would get on a strictly proportional allocation).³

Because an electoral system may have more than one electoral formula operating at different levels (as in Belgium or Japan), it is not always easy to characterize the overall proportionality of a system. Putting this technical difficulty aside (as does most of the literature), one can denote the overall proportionality of the votes-to-seats translation in a system by P , with higher values of P denoting more proportional systems.

6.2.2 The Number of Seats per District, Tier, and Segment

For a given structure of districts, tiers, and segments, the question arises as to how many seats will be at stake at each point in the system. For simple (one-tier, one-segment) electoral systems, the number of seats elected from a given district (known as the district magnitude) is a key parameter, because it determines the

³Of course, there are many different ways to define the gap between vote and seat shares, leading to different algorithmic embodiments of the proportional ideal (Cox & Shugart, 1991). Conceptually, I prefer to characterize systems in terms of responsiveness or *big-party bias*, defined as the degree to which parties with larger vote shares tend also to receive higher-than-proportional seat shares (Cox & Shugart, 1991). Most of the literature, however, has focused on proportionality, and an array of workable summary measures of it exist (see Gallagher (1991)).

minimum vote share that guarantees that a list or candidate will win a seat, hence the minimum viable size of electoral alliances. For example, in a single-member district in the United States, a candidate must win over 50% of the vote to guarantee a seat, whereas in a three-seat district in pre-1993 Japan, over 25% sufficed.⁴

In multitiered systems, it is harder to define the minimum vote share that guarantees that a party will win a seat. Electoral scholars have come up with approximate formulas, under the rubric of the *effective magnitude* or *effective threshold*, that translate complex systems to their equivalent one-tier systems (Taagepera & Shugart, 1989; Lijphart, 1994). The effective magnitude, in other words, is an attempt to put all systems on a single metric, reflecting the minimum or threshold size which office-seeking politicians might aim to exceed. In what follows, I denote the effective magnitude of a system by M . The larger the effective magnitude, the lower the threshold an alliance must surpass to guarantee a seat.⁵

6.2.3 The Number of Votes per Voter

The method of voting in a system regulates the translation of voters' preferences into votes. The systems most widely used in regional and national elections give voters a *fixed number* of votes which they must then *allocate* across candidates and/or lists. Voters thus have a limited ability to rank candidates (or lists), which in turn limits the ability of a voting method to select "good" candidates. For example, only voting methods in which voters can rank-order the candidates can ensure that Condorcet winners are selected, when they exist.

Proposals to offer voters more complex options have long existed and have been increasingly promoted in the last half century. One proposal is approval voting, in which voters have a *variable number* of votes which they must *allocate* across candidates (Brams & Fishburn, 1978). A second proposal is to allow voters to rank all candidates on the ballot from most to least preferred, with varying ideas about how to translate ranks into seats—e.g., the single transferable vote (Mill, 1861) or total vote runoff (Foley, 2023).

One problem with such systems is that voters who think electoral maladministration, manipulation, and fraud are serious problems may be especially worried about complex voting methods. Bolstering public trust in electoral administration—an important issue since the January 6, 2021, break-in at the US Capitol—may require simplifying voters' choices.

The remainder of the paper focuses mostly on the fixed-vote allocation systems. These systems can be characterized as enabling voters either to reward the best competitor(s) or to punish the worst competitor(s) (Cox, 1987). To understand this characterization, note that different systems allow different numbers of votes per

⁴To clarify the meaning of "guarantee" here, note that it is possible to lose a single-member district with a vote percentage of 49.9% but not with 50.1%.

⁵See Lijphart and Gibberd (1977) for a study of thresholds.

voter. Those that give voters only one vote enable each of them to reward the candidate they judge best (by casting their single vote for that candidate). However, voters have no ability to single out for special opprobrium the worst among the candidates not voted for. In contrast, consider systems that give voters $M - 1$ votes and force them to cast all of them for separate competitors, as used to be the case in many local elections in the United States (Cox, 1984). Such systems enable each voter to punish the candidate they judge worst (by withholding a vote). However, voters have no ability to single out for special favor the best among those voted for. Yet, other voting methods fall along a continuum running from the pure best-rewarding to the pure worst-punishing cases.

A summary measure of the degree to which a system is best-rewarding has been devised by Cox (1987). I shall call a transformation of this measure the *effective number of votes per voter*, denoted V . The logic is that, after one clears away the details of different voting rules, one is left with a simple contrast between systems that give few votes (and thus enable voters only to distinguish between the best and the rest) and systems that give many votes (and thus, assuming all votes must be cast, enable voters only to distinguish between the worst and the rest, assuming all must be cast).⁶

Empirically, most national assemblies are elected using one-vote ($V = 1$) best-rewarding methods. A certain number do use methods intermediate between the best-rewarding and worst-punishing extremes, however. For example, Spain's Senate is elected using a limited vote (Lijphart et al., 1986), as was part of Great Britain's House of Commons in the nineteenth century. There are even fewer countries using a pure worst-punishing method—one being Mauritius (Cox, 1997, pp. 146–147).

Another distinction worth noting at this point is that between systems in which votes are cast directly for individual *candidates* and those in which votes are cast for sets or *lists* of candidates, with each list endorsed by one or more parties. When votes are cast for lists (or aggregate to the list level), it is possible to use various proportional representation formulas in translating votes into seats. When votes aggregate only to the level of the individual candidate, in contrast, it is not possible to use proportional representation formulas (except in a mathematically trivial sense). Thus, the distinction drawn below between more and less proportional formulas is in part a function of the voting options given to voters.

6.2.4 Summary

In summary, electoral systems can be characterized by how many votes voters cast (by extension, the effective number of votes per voter, V), how many seats are awarded in the typical district (by extension, the effective magnitude, M), and how proportionally votes are converted into seats (P). The sections that follow review what we know about how electoral competitors' strategies of coordination, persuasion, and mobilization change, in response to changes in V , M , and P .

⁶Multiple vote systems that allow partial abstention are similar to approval voting (Cox, 1984).

6.3 Electoral Coordination

Modern representative democracy presents at its core a series of coordination problems that arise as natural consequences of electoral competition for governmental offices. A group with enough votes to elect some number of candidates in a given (legislative or executive) race will in fact elect that number only if it can make its votes count by concentrating them appropriately. One way to avoid spreading votes too thinly is to limit the number of candidates vying for the group's support. But which potential candidates, representing what shades of opinion, will withdraw in favor of which others? If attempts to limit the number of candidates fail, another chance to make votes count arises on polling day, when voters can concentrate their votes on a subset of the available candidates. But which candidates will bear the brunt of strategic voting and which will be its beneficiaries?

Electoral coordination—whether the coordination of entry, resources, or votes—can occur via a variety of processes. These processes can be classified in terms of the main actors involved (voters or elites) and the level at which their interaction is pitched (within individual electoral districts or across districts), yielding four categories. Let's briefly discuss each of these.

6.3.1 Coordination of Votes Within Districts

Traditionally, strategic voting (voting so as to secure the best possible outcome rather than to support the most-preferred competitor) has been thought to concentrate votes. Consider, for example, a three-candidate contest for a single seat in which candidates A and B are tied in the polls at 40%, with C trailing at 20%. Those who most prefer the trailing candidate, C, may decide to vote for whichever of the leading candidates they prefer. Such strategic voting would have the consequence of concentrating the actual vote, relative to what would result if every voter simply voted for their most-preferred candidate. Indeed, in the extreme, all C's supporters might vote strategically and what had been a three-way race would reduce to a two-way race.

Does strategic voting have similar vote-concentrating consequences in other electoral contexts? As it turns out, the extent to which strategic voting concentrates votes depends on the electoral system in place, in two main ways (cf. Cox, 1997). First, only in best-rewarding systems does strategic voting concentrate votes; in worst-punishing systems, it typically does just the opposite (Cox, 1999a). Second, within the class of best-rewarding systems, smaller effective magnitudes lead to a greater concentration of votes, according to the “M + 1 rule.” Let's consider this second proposition in greater detail.

6.3.2 The “M + 1 Rule”: Theory

Best-rewarding systems can be thought of as each having a maximum “carrying capacity” of parties, call it C. When the number of parties falls short of C, it is theoretically possible that every party in the system can be in serious competition to

win a seat—expected either to win one or more seats or to finish as the runner-up for the last-allocated seat. Strategic voting should be minimal in such situations. When the number of parties exceeds C , in contrast, it is increasingly unlikely that all parties can be seriously in contention for a seat. Instrumentally rational voters will avoid voting for parties that are unlikely to contend for a seat, however. Thus, if voters' initial priors concerning the distribution of preferences in the electorate are sufficiently precise, so that the identity of the trailing competitors is clear enough, one expects the vote share of trailing parties to fall short of what their vote share would be were all voters to vote sincerely.⁷ That is, weak parties find some of their supporters voting for the best of the candidates still in contention for the last-allocated seat, rather than for them.

For systems using M -seat districts in which each voter has a single vote, cast either for a candidate or a party list, we know that the “carrying capacity” $C = M + 1$ (Cox, 1997). That is, in an M -seat district, one expects no more than $M + 1$ viable competitors.⁸

6.3.3 The “ $M + 1$ Rule”: Evidence

Empirical evidence supporting the $M + 1$ rule comes in a variety of forms. Here, I shall briefly discuss evidence from mass surveys documenting the existence of strategic voting and evidence from cross-national aggregate analyses documenting the relationship between district magnitude (M) and the number of viable parties.

Rather than examine survey evidence from a variety of countries, I focus on the case of Great Britain, for which perhaps the largest literature on strategic voting exists. During the elections of the 1980s, when the Alliance surged to near-parity in votes with the Labour Party, estimates of the percentage of voters who voted for their

⁷The models of Palfrey (1989), Myerson and Weber (1993), and Cox (1997) all assume that voters' know the *expected* constituency-wide breakdown of preferences with certainty. Myatt (2007), in contrast, assumes that voters have diffuse priors over this expected breakdown of preferences. These two differing formal assumptions correspond to two polar substantive assumptions. To take the case of three-party competition in a single-member district, one might assume that voters know, from previous elections, that the Conservatives generally get between 40% and 42% of the vote, with Labour at 38–42% and the Liberals at 16–22%. Pushing this substantive idea—that the voters know a lot about the *expected* breakdown of the vote between the three parties—to its logical extreme, one arrives at the models of Palfrey, Myerson and Weber, and Cox. On the other hand, what if the race is between three new parties in a new democracy or a realignment of forces has made past results a poor guide to the future? Pushing this idea—that voters know little about the expected breakdown of preferences in a constituency—to its logical extreme, one arrives at the model of Myatt.

⁸The results stated in this paragraph follow from the Palfrey/Myerson-Weber/Cox model. In Myatt's model (which is fully developed for the $M = 1$ case), there is a *tendency* to concentrate votes on two competitors, but it stops short of the extreme identified in the earlier models. Thus, the version of the “ $M + 1$ rule” that Myatt's model would support would probably read something like: “When there are more than $M + 1$ competitors, votes will concentrate on $M + 1$ of them (but only to a limited extent).”

second- or third- rather than first-preference candidate range from 5.1% to 17% (see, e.g., Johnston and Pattie (1991), Heath and Jowell (1991, p. 54), Lanoue and Bowler (1992), Niemi et al. (1992), and Crewe (1987, p. 55)). Estimates of the percentage of voters who would “consider” voting tactically vary from an average Gallup figure in 1986–1987 of 15% to an average BBC Newsnight figure of 41% (Catt, 1989).

These estimates of the proportion of the British population who voted strategically all count a respondent as voting strategically only if they reported voting for a candidate who was not their most preferred but was “in the running” to win the seat. If one looks only at voters who had an opportunity to vote strategically—supporters of trailing candidates—one naturally finds much higher rates of strategic voting (Blais & Nadeau, 1996; Alvarez et al., 2006; Abramson et al., 2018). A few studies have begun to examine the characteristics of voters prone to vote strategically, finding that older and richer (Eggers & Vivyan, 2020) and more risk-acceptant (Martin, 2021) voters are more likely to do so.

A handful of studies have begun to compare what the outcome would have been, had there been no strategic voting, to the actual outcome. The earliest approach analyzed vote swings in consecutive elections where strategic voting incentives had shifted (e.g., Curtice & Steed, 1997 (Appendix 2), 2001 (Appendix 2)). Myatt and Fisher (2002) used game-theoretic modeling to construct counterfactual outcomes. Kiewiet (2013) estimated a sincere vote outcome for each constituency, based on its demographics, then compared the post-estimation counterfactual election results to those actually observed. Kim and Fording (2001) and Herrmann, Munzert, and Selb (2016) used survey data to construct counterfactual outcomes. These various estimation approaches reached broadly similar conclusions. Estimates of the total number of UK constituencies in which strategic voting was pivotal to the observed outcome range from 13 to 46 in the general election of 1997 and from 7 to 15 in the general election of 2001 (Herrman et al., 2016, p. 601).

This excursion into the literature on strategic voting suffices to demonstrate that part of the causal mechanism underlying the $M + 1$ rule appears to operate as envisioned by the theory. Another sort of study that bears more directly on the $M + 1$ rule relates characteristics of electoral systems to the size of the party system. Many works investigate this relationship cross-nationally, including Rae (1971), Taagepera and Shugart (1989), and Lijphart (1994) (for further citations, consult these works). All report a relationship between the number of parties in a country and the average or median district magnitude that is broadly consistent with the $M + 1$ rule.

For a limited number of systems, there are more direct tests of the $M + 1$ rule. These can be roughly divided into systems where the $M + 1$ rule is predictively accurate—such as the United States, Japan (Reed, 1990; Cox, 1997; Niemi & Hsieh, 2002), Taiwan (Hsieh & Niemi, 1999), and India (Chhibber & Kollman, 1998)—and systems where it is not, such as Canada (Gaines, 1999; Blais, 2002) and Papua New Guinea (Cox, 1997). For both cases of apparent success and apparent failure, the interesting question is whether the theoretical preconditions of the $M + 1$ rule (e.g., relatively precise public knowledge of the candidates’ likely order of finish, prior to

polling day) are met or not. Thus far, none of the exceptions to the rule occur where the theoretical preconditions are met, and none of the successes occur where the preconditions clearly fail. Nonetheless, it remains unclear how much the rule's variable success is driven by variations in strategic voting as opposed to variations in strategic coordination at the elite level.

6.3.4 Summary

To reframe the two results just noted, variations in the voting method (V) affect whether strategic voting leads to a deconcentration of votes (in worst-punishing systems) or to a concentration of votes (in best-rewarding systems). Variations in the effective magnitude (M) affect the carrying capacity of best-rewarding systems, with higher values allowing more competitors in equilibrium.⁹

6.4 Coordination of Elites Within Districts

There are several species of coordination problem that elites face within electoral districts, of which I consider two. One deals with the provision of campaign finance by contributors. Another concerns optimal nomination.

6.4.1 Strategic Contributions

Suppose that contributors of campaign finance are primarily interested in legislative services that an elected representative can provide once in office (e.g., lobbying party leaders on behalf of the contributor, introducing bills for the contributor). Suppose also that potential entrants seek financing for a campaign by essentially selling access to their future labor (cf. Denzau & Munger, 1986). If elected to office, candidates pay off their financial backers. Otherwise, their contributors get nothing. Given these assumptions, there will be a tendency for contributors to coordinate their contributions, because only winning candidates pay off and those with more contributions are more likely to win. At an informal level, what can be conjectured is that no more than $M + 1$ competitors in an M -seat district will attract significant financial backing from contributors seeking legislative services (although contributions from ideologically motivated sources are a different story).

Such a conjecture seems to fit the facts in the US congressional case (cf. Jacobson, 1980). Often, there is only one well-financed candidate: the

⁹The third feature, P (proportionality), also plays a role—though it is harder to disentangle from that of V and M . For either $V > 1$ or $M > 1$, the possibility of significant variation in P arises, and the general rule is this: the more proportional the system, the less coordination is demanded, and the more parties there can be in equilibrium.

incumbent. Occasionally, there are two well-financed candidates, in which case the actual outcome tends to be much closer. Finding three or more well-financed candidates in the general election is extremely rare.

6.4.2 Optimal Nomination

Parties face three recurring problems when nominating candidates for office: overnomination, undernomination, and factional cheating on nomination deals. In this section, I briefly consider each of these problems.

Overnomination means nominating too many candidates, who then split the party's vote too thinly and end up winning fewer seats than would have been possible had the party nominated fewer candidates.¹⁰ Undernomination means nominating fewer candidates than the party has votes to elect. Optimal nomination means fielding a number of nominees that maximizes the party's expected seat share in the district.

In some systems, figuring out the optimal number of nominees is easy. For example, in the United States, each major party runs either zero or one candidate in each district. The zeroes typically arise in districts that are so "safe" for the other party that the party decides to save itself the cost of running a candidate there. In most districts, both parties decide to run a candidate, although their commitment of resources is highly sensitive to the closeness of the race.

In systems with larger-magnitude districts, figuring out the optimal number of nominees is not as easy. For example, in a three-seat district in the Japanese election of 1980, some Liberal Democrats might have thought the party had enough votes to sweep the three seats, while others believed that only two seats could be won. If there were currently two Liberal Democratic incumbents in the district, they would naturally not wish to have a third colleague nominated, as this might reduce their probability of winning a seat from essentially 1 to roughly 2/3 (if they are correct that the party will win only two seats and if additionally each of the party's nominees has an equal chance of winning a seat).

Studies of nomination in the Japanese system over the period of Liberal Democratic dominance (1955–1993) find a steady improvement in the party's ability to arbitrate internal disagreements over nominations of the sort described above. In particular, early in the period, the party fairly often overnominated, with the consequence that two or more of its candidates "fell down together," as the Japanese put it, in a version of the game of *Chicken*. Experiences of this sort led fairly quickly to

¹⁰In single-member districts, overnomination simply means nominating more than one candidate. It is an easy mistake to spot, and all parties operating in single-member districts go to considerable (and generally successful) lengths to avoid dual candidacies. In M-seat districts, it is harder to say what constitutes overnomination. Nominating more than M candidates will usually be overnominating—but not necessarily if votes transfer or pool. In systems where votes do not transfer or pool, the party will have beliefs about its likely vote share. If it has votes enough to win M-2 seats, then it should nominate M-2 candidates and no more.

improved deal-making among the factions and fewer overnominations (Reed, 1990; Cox & Rosenbluth, 1994; Niemi & Hsieh, 2002).

Similar coordination problems arise under other systems with multimember districts, such as the limited vote in Spain (Lijphart et al., 1986), the single non-transferable vote in Taiwan (Cox & Niou, 1994), and the cumulative vote system in Illinois (Goldburg, 1994). When district magnitudes exceed five, there is an increasing probability that elections will be based on lists, which essentially solve the coordination problem mechanically by forcing candidates to share their votes with one another (or, alternatively, forcing voters to vote for indivisible groups rather than individuals).

6.5 Coordinating Across Districts

A distinctly different problem of coordination arises at the cross-district level. If there are multiple districts, each with its own population of parties, will those parties cooperate across district lines or not?

One advantage of cross-district coordination—or linkage—can be suggested by considering two districts in which two leftist parties each run a candidate against a single rightist candidate. Let us suppose that the leftists can win both districts, if they combine their votes, but neither, if they split. One solution to their problem is to negotiate cooperation in each district separately. Another, potentially easier, solution is to *trade* withdrawals: party A withdraws in district 1 in exchange for party B withdrawing in district 2. Part of the deal, of course, is that each party's supporters will be encouraged to vote for the other's nominee.¹¹

Cross-district trading of nominations and withdrawals can occur at various levels of “intensity.” At the low end, there can be a few scattered deals affecting only a small number of districts (e.g., early cooperation between the *Komeito* and other parties in Japan; cf. Christensen, 2000). Then, there can be comprehensive but nonetheless election-specific deals, in which the whole pattern of nominations is decided centrally (e.g., the alliance between the Social Democrats and Liberals in the United Kingdom). A set of parties' (or factions') relations can be even further deepened if they *regularly* negotiate nominations centrally—as has occurred, for example, in Chile within the *Concertación*, in post-reform Italy within Berlusconi's group (Di Virgilio, 1998), and in pre-reform Japan within the LDP (Cox & Rosenbluth, 1996). Finally, the various parties might fuse, formally abandoning their separate labels and organizations.

In addition to trading nominations, parties in some systems have a clear incentive to legally unite for purposes of seat allocations in upper tiers. In Belgium, a party in a given *arrondissement* can participate in the allocation of seats at the provincial level only if it formally affiliates with parties from other *arrondissements* in the province.

¹¹ Yet, another solution along the same lines is for party A actually to jointly nominate B's candidate and vice versa. This is legally permissible in some but not all systems.

In Hungary, parties are eligible to run national-level lists only if they field at least seven regional lists. In Japan, a candidate running in one of the single-member districts is eligible to win a seat in the encompassing PR district only if she formally affiliates with a party running such a list. Similar incentives to affiliate with larger electoral forces arise in a number of other multitier and multi-segment systems (cf. Shugart & Wattenberg, 2001).

A third incentive to form broad national parties arises in presidential systems. The presidential election so greatly dominates the mass media that hanging on to the president's coattails is in some systems a natural electoral strategy for legislators. At the same time, cultivating the support of local politicians is a natural strategy for presidential candidates. This may help explain why the resuscitation of real competition for the presidency in the United States brought with it the emergence of our second party system (McCormick, 1975), why the creation of presidential elections in France led quickly to a bipolarization of legislative elections (Wilson, 1980), and why nominations in Uruguay are dominated by the presidential candidates (Morgenstern, 2001). The same story can play out in parliamentary systems, to the extent that parliamentary elections revolve around the prospective prime ministers, as in the United Kingdom, Israel, or Germany.

The extent of cross-district coordination (in its various forms) depends on what such coordination can win. Two obvious prizes are legislative seats and executive portfolios. As noted above, each electoral system presents different incentives to coordinate across districts in pursuit of legislative seats—to use votes in primary districts more efficiently, to pool votes in secondary districts, and to take advantage of executive coattail effects. Each system also presents different incentives to coordinate in pursuit of portfolios, although here, the key conditions pertain not so much to the electoral system proper as to various constitutional and legislative features.

Hicken (2002) points out that incentives to coordinate are greater the more concentrated is power in a single post whose election depends on winning a majority in the assembly. From this perspective, federalism can diminish such incentives, as can bicameralism, presidentialism, and the dispersion of executive power among coequal ministers. Chhibber and Kollman (1998) and Samuels (1998) both explore a more specific version of this hypothesis—that greater fiscal centralization in a state will lead to greater linkage.

Another line of studies looks not at the short-term office benefits of forming larger electoral aggregates but instead at the long-term policy benefits of refusing to coordinate. Studies of third-party movements (e.g., Rosenstone et al., 1996 and Hug, 2001) typically find strong policy preferences motivating the formation of separate vehicles, even when going it alone sacrifices seats that could be won with a more pragmatic stance (and the consequent alliance that would then become feasible). A variant on this theme concerns regional parties, especially those based on ethnic identities. Also related are studies of coordination in newly established democracies, where uncertainty about who the players are and who can outlast whom delays coordination (cf. Moser, 1999; Zielinski, 2002).

6.6 Electoral Persuasion

One way to win office is, *given* a particular distribution of preferences, to coordinate campaign finance, candidate entry, and voters' choices in such a way as to maximize the number of seats the "socialists" or "liberals" or "Christian Democrats" win. The focus is on translating preferences efficiently into seats (via votes).

Another way to win office is to persuade. Rather than take preferences as given, one influences those preferences as best one can.

The models of persuasion considered in this section all take parties or candidates as the main actors. Some assume that competitors seek only office, while others assume they seek to maximize the rents they can extract from office. Some assume that competitors can promise only public goods, while others assume that they can promise private goods (only or as well).

6.6.1 Office-Seeking Competitors with Credible Promises

Suppose that competitors seek only to win office and can make credible promises during the campaign concerning the policies they will pursue if elected. One possibility is that candidates promise packages of public goods which can be characterized as falling somewhere along the left-right spectrum ("position taking"). Another possibility is that candidates promise packages of transfers ("bribing").

When competitors take positions (i.e., promise only public goods), the consequences of changing electoral rules on their behavior can be stated as follows: higher values of V or M , along with more proportional values of P , lead competitors to disperse across the left-right spectrum (Cox, 1987, 1990). The intuition behind the last two results is roughly as follows. When there are many seats to win and they are allocated proportionally, small shares of the vote suffice to win seats. Thus, electoral competitors can carve out narrow ideological niches and still be successful. In contrast, when there is only one seat to win and it is allocated to the plurality winner, electoral competitors can win seats only if they can amass the largest share of votes. This means that appealing to a narrow ideological niche is insufficient to win seats and a broader appeal must be fashioned.¹²

When competitors distribute (only) private goods, similar results obtain. Higher values of V or M , along with more proportional values of P , lead competitors to concentrate their "bribes" on smaller subsets of voters, giving the rest almost nothing (Myerson, 1993). The logic is again driven by the minimum share of votes that will suffice to win a seat. In electoral systems where this minimum is lower, more concentrated appeals are more effective. The main difference between the models is simply in the tools that competitors are assumed to have available.

A related theoretical effort is that of Carey and Shugart (1995), who attempt to rank-order a wide range of electoral systems in terms of the incentives they provide

¹²The first result is somewhat more subtle. See Cox, 1987, 1990).

politicians to cultivate a personal vote, rather than rely on their party's overall image. As the means by which one might cultivate a personal reputation is usually the distribution of particularistic goods, Carey and Shugart's ranking reflects the insight offered above. In addition, however, Carey and Shugart identify a number of other features of electoral systems that promote personal vote seeking, including a range of provisions that essentially force members of the same party and district to compete against one another (e.g., the single non-transferable vote, open lists, and preference votes). The degree of forced intraparty competition is an important feature of electoral systems that is not captured by the (V, M, P) coding suggested here.

6.6.2 Office-Seeking Competitors Without Credible Promises

What if voters do not believe the competitors' promises, because they suspect that, once elected, politicians will have incentives to do the bidding of special interests, rather than fulfill campaign pledges?¹³ The credibility of competitors' promises is an important analytical issue for some purposes. However, this problem does not appear to affect the main comparative statics result derived above—that increases in V, M, or P induce greater niche-seeking by competitors. Thus, I ignore it here.

6.6.3 Rent-Seeking Competitors with Credible Promises

In the office-seeking models discussed above, winning an office confers a fixed amount of utility on the victor. Conceptually, this utility might consist of the salary of the office, its prestige, and any other "ego rents" attached thereto.

Another class of models assumes that office-holders can extract a variable amount from their office. In particular, several models envision one or more "predatory parties" whose maximand is their *expected rent*: the probability of their winning control of government times the fiscal residuum they can extract once in control. Each party promises a certain tax rate, a certain expenditure on public goods, and a certain bundle of transfers to the groups in the electorate. If elected, the party collects

¹³ There are several approaches to dealing with this credibility problem. Ferejohn (1986) and Alesina (1988), for example, consider repeated games of various sorts, attempting to identify when credibility will emerge endogenously. Bernhardt and Ingberman (1985) simply assume it is costly to state positions distant from those taken in the past. Osborne and Slivinski (1996) and Besley and Coate (1997) go even further, stipulating that the only credible promise that a candidate can make is to implement his or her ideal point. The Besley-Coate and Osborne-Slivinski models can be construed as replacing one sort of credibility problem with another. There is no longer the problem of voters believing anything the candidates say, because the candidates only state their ideals, which they then have an incentive to implement if elected. However, the models depend heavily on common knowledge of ideal points of all citizens. If a particular citizen's ideal point is not common knowledge, then that citizen faces the issue of how credibly to communicate what his or her ideal really is.

taxes, provides the level of public goods and transfers promised, and keeps any residual for itself.

Several comparative static results drawn from this model hinge on electoral rules. I consider just one here: Persson and Tabellini (2000) show that changing from a single nationwide district operating under proportional representation (high M and high P) to a set of single-member districts operating under plurality rule (low M and low P) induces the following changes: parties promise more transfers to swing districts, financing the increased expenditure by reducing both the provision of public goods and their own rents. The logic of this result is as follows. If control of government comes down to who wins in a particular handful of districts, as it can in a district-based electoral system, competition will focus on those districts. But this means that both parties will offer more transfers to the voters in the swing districts than they would offer to the same voters were there were no districts (just a nationwide PR election).¹⁴ Studies by Lizzeri and Persico (2001) and Milesi-Ferretti et al. (2002) also conclude that, because targetable goods are electorally more valuable in district-based systems, the equilibrium level of transfers (pork barrel projects and other geographically targetable benefits) will be higher. As evidence in favor of the idea that district-based electoral systems promote transfers at the expense of general public goods, Persson and Tabellini (2000) show that expenditures on welfare are higher in more proportional systems, controlling for the age structure of the population, per capita income, trade openness, and federalism, *inter alia*.

6.7 Mobilization

The literature offers several decision-theoretic analyses of how electoral mobilization varies with electoral rules, which typically argue that turnout will be higher under multi-seat PR than single-seat plurality rules (see Cox (1999b) for a review). Game-theoretic models (Shachar & Nalebuff, 1999; Herrera et al., 2014, 2016) have explored how equilibrium turnout levels vary across systems that vary in proportionality (via variations in M and P). For example, Herrera et al. (2014) show that district-level turnout will be highest in (sufficiently) competitive single-seat plurality-rule districts, intermediate in multi-seat PR districts, and lowest in uncompetitive single-seat plurality-rule districts. Thus, aggregate turnout in a single-seat plurality system may be higher or lower than in a multi-seat PR system, depending on how many single-seat districts are competitive. Moreover, changing from single-seat plurality to multi-seat PR should result in a contraction of district-level turnout rates, with competitive districts declining toward the PR level and uncompetitive districts increasing (Cox et al., 2016).

¹⁴ That these increased transfers are financed by reducing *both* public goods *and* private rents simply reflects that neither the level of public goods nor the fiscal residuum was pegged at corner solutions prior to the lower of M and P .

Early empirical evidence in favor of this model—all cross-national—can be found in Blais and Carty (1990), Blais and Dobrzynska (1998), Jackman (1987), Jackman and Miller (1995), and Powell (1980, 1982, 1986). More recent empirical support is surveyed by Cancela and Geys (2016). While they find that most papers continue to examine cross-national variation, ten surveyed papers considered subnational variation in electoral rules. As Smith (2018, p. 198) reports, “70 percent of these subnational studies (and 76 percent of the individual tests) produce positive findings” (i.e., that more proportional electoral rules are associated with higher turnout rates).

Higher levels of mobilization are also more likely as V decreases. The logic is simply that, if voters cast lots of votes, then a single competitor may not internalize the full benefits of mobilizing any particular voter or segment of voters (Cox, 2015). Of course, parties may mitigate this free rider problem by forming pre-election coalitions (Tillman, 2015).

6.8 Conclusion

The vast majority of models of electoral competition have not been concerned with institutional comparative statics. Only a few take the electoral rules themselves as the primary independent variables. And, thus far, there is little work that alters non-institutional features of the model then considers whether those alterations condition the effect of changing the rules. In this sense, the formal literature analyzing electoral systems is relatively thin.

In this chapter, I have characterized the independent variable—the electoral system—in terms of three features: the effective number of votes per voter, the effective number of seats per district, and the proportionality of the votes-to-seats mapping(s). The dependent variables—the strategies adopted by electoral competitors within a given electoral system—have been parsed into strategies of coordination, persuasion, and mobilization.

6.8.1 Coordination

Taking voters’ preferences and turnout probabilities as fixed, political competitors face problems in coordinating endorsements, entry, campaign finance, and votes to maximize their respective seat shares. The severity of the coordination problem (s) that competitors face depends on the electoral rules governing their competition. The most prevalent voting methods give citizens just one vote to cast (for either a candidate or a list) and fall into the “best-rewarding” category. For such voting methods, strategic coordination leads to a concentration of votes upon a subset of viable candidates. The size of this subset is determined by the number of seats available to be won, with the most general rule of thumb encapsulated in the “ $M + 1$ rule”: the number of viable competitors cannot exceed $M + 1$ in equilibrium (when voters are primarily interested in who wins the current election and have sufficiently precise information concerning the likely order of finish of the competitors).

6.8.2 Persuasion

Rather than taking preferences as fixed, political competitors can also engage in a variety of persuasive activities. They can promise to deliver either broadly targeted goods (in the extreme, Samuelsonian public goods) or narrowly targeted goods (in the extreme, private goods) or something in between. The results on persuasion I have reviewed here all make essentially the same point: that electoral systems differ in the extent to which they encourage competitors to fashion narrow appeals.

At the within-district level, fewer votes per voter, more seats per district, and greater proportionality lead competitors to cater to narrower clienteles within the electorate. If competitors are constrained to promise only public goods, they “cater to narrower clienteles” by spreading out over the ideological spectrum, rather than bunching at the median (Cox, 1987, 1990). If they are constrained to promise only private goods, they “cater to narrower clienteles” by targeting their bribes, rather than diffusing them (Myerson, 1993).

At the across-district level, elections based on small-magnitude districts lead competitors to prefer geographically targetable goods, rather than national public goods. The reason is, roughly, that geographically targetable goods can be promised specifically to districts where the competition is close, whereas public goods cannot.

6.8.3 Mobilization

Holding voters’ preferences and their estimates of candidates’ viability constant, competitors can seek to affect their decisions to vote or abstain. The general findings in the literature are that fewer votes per voter, more seats per district, and greater proportionality increase the mean level of turnout and decrease the variance in turnout. This relationship suggests an alternative causal mechanism to explain the observation (Persson & Tabellini, 2000) that more proportional electoral systems foster greater expenditures on welfare: more proportional systems (higher M and P) lead to higher turnout rates among the poorest citizens (those who are the first to “drop out” under more majoritarian systems with their lower turnouts); more consistent turnout among the poorest citizens leads to more consistent policies serving their social insurance desires (cf. Lijphart, 1997).

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Abramson, P. R., Aldrich, J. H., Diskin, A., Houck, A. M., Levine, R., Scotto, T. J., & Sparks, D. B. (2018). The effect of national and constituency expectations on tactical voting in the British general election of 2010. In L. B. Stephenson, J. H. Aldrich, & A. Blais (Eds.), *The many faces of strategic voting* (pp. 28–61). University of Michigan Press.

- Alesina, A. (1988). Credibility and policy convergence in a two-party system with rational voters. *American Economic Review*, 78(4), 796–805.
- Alvarez, M., Boehmke, F., & Nagler, J. (2006). Strategic voting in British elections. *Electoral Studies*, 25(1), 1–19.
- Bernhardt, M. D., & Ingberman, D. E. (1985). Candidate reputations and the ‘incumbency effect’. *Journal of Public Economics*, 27(1), 47–67.
- Besley, T., & Coate, S. (1997). An economic model of representative democracy. *Quarterly Journal of Economics*, 112(1), 85–114.
- Blais, A. (2002). Why is there so little strategic voting in Canadian plurality elections? *Political Studies*, 50(3), 445–454.
- Blais, A., & Dobrzynska, A. (1998). Turnout in electoral democracies. *European Journal of Political Research*, 33, 239–261.
- Blais, A., & Ken Carty, R. (1990). Does proportional representation Foster voter turnout? *European Journal of Political Research*, 18, 167–182.
- Blais, A., & Nadeau, R. (1996). Measuring strategic voting: A two-step procedure. *Electoral Studies*, 15(1), 39–52.
- Brams, S. J., & Fishburn, P. C. (1978). Approval Voting. *American Political Science Review*, 72(3), 831–847.
- Cancela, J., & Geys, B. (2016). Explaining voter turnout: A meta-analysis of national and sub-national elections. *Electoral Studies*, 42(1), 264–275.
- Carey, J., & Shugart, M. S. (1995). Incentives to cultivate a personal vote: A rank ordering of electoral formulas. *Electoral Studies*, 14(4), 417–439.
- Catt, H. (1989). Tactical voting in Britain. *Parliamentary Affairs*, 42, 548–559.
- Chhibber, P., & Kollman, K. (1998). Party aggregation and the number of parties in India and the United States. *American Political Science Review*, 92(2), 329–342.
- Christensen, R. (2000). *Ending the LDP hegemony: Party cooperation in Japan*. University of Hawaii Press.
- Cox, G. W. (1984). Strategic electoral choice in multi-member districts: Approval voting in practice? *American Journal of Political Science*, 28(4), 722–738.
- Cox, G. W. (1987). Electoral equilibrium under alternative voting institutions. *American Journal of Political Science*, 31(1), 82–108.
- Cox, G. W. (1990). Centripetal and centrifugal incentives in electoral systems. *American Journal of Political Science*, 34(4), 903–935.
- Cox, G. W. (1997). *Making votes count: Strategic coordination in the World's electoral system*. Cambridge University Press.
- Cox, G. W. (1999a). Electoral rules and electoral coordination. *Annual Review of Political Science*, 2, 145–161.
- Cox, G. W. (1999b). Electoral rules and the calculus of mobilization. *Legislative Studies Quarterly*, 24(3), 387–419.
- Cox, G. W. (2015). Electoral rules, mobilization and turnout. *Annual Review of Political Science*, 18, 49–68.
- Cox, G. W., & McCubbins, M. D. (2001). The institutional determinants of economic policy outcomes. In M. D. McCubbins & S. Haggard (Eds.), *Presidents, parliaments and policy* (pp. 21–63). Cambridge University Press.
- Cox, G. W., & Rosenbluth, F. (1994). Reducing nomination errors: Factional competition and party strategy in Japan. *Electoral Studies*, 13(1), 4–16.
- Cox, G. W., & Rosenbluth, F. (1996). Factional competition for the party endorsement: The case of Japan's Liberal democratic party. *British Journal of Political Science*, 26(2), 259–297.
- Cox, G. W., & Shugart, M. (1991). Comment on Gallagher's ‘proportionality, disproportionality and electoral systems’. *Electoral Studies*, 10(4), 348–352.
- Cox, G. W. & Niou, E. (1994). Seat bonuses under the single nontransferable vote system: evidence from Japan and Taiwan. *Comparative Politics*, 26(2), 221–236.

- Cox, G. W., Fiva, J., & Smith, D. M. (2016). The contraction effect: How proportional representation affects mobilization and turnout. *Journal of Politics*, 74(4), 1249–1263.
- Crowe, I. (1987). What's left for labour: An analysis of Thatcher's victory. *Public Opinion*, 10, 52–56.
- Curtice, J., & Steed, M. (1997). *The British general election of 1997*. Macmillan.
- Curtice, J., & Steed, M. (2001). *The British general election of 2001*. Palgrave.
- Denzau, A. T., & Munger, M. C. (1986). Legislators and interest groups: How unorganized interests get represented. *American Political Science Review*, 80(1), 89–106.
- Di Virgilio, A. (1998). Electoral alliances: Party identities and coalition games. *European Journal of Political Research*, 34(1), 5–33.
- Eggers, A., & Vivyan, N. (2020). Who votes more strategically? *American Political Science Review*, 114(2), 470–485.
- Foley, E. B. (2023). Total vote runoff: A majority-maximizing form of ranked choice voting. *The University of New Hampshire Law Review*, 22(1), 1–33.
- Ferejohn, J. (1986). Incumbent performance and electoral control. *Public Choice*, 50(1/3), 5–25.
- Gaines, B. (1999). Duverger's law and the meaning of Canadian exceptionalism. *Comparative Political Studies*, 32(7), 835–861.
- Gallagher, M. (1991). Proportionality, disproportionality, and electoral systems. *Electoral Studies*, 10(1), 33–51.
- Goldburg, C. (1994). The accuracy of game theory predictions for political behavior: Cumulative voting in Illinois revisited. *Journal of Politics*, 56(4), 885–900.
- Heath, A., & Jowell, R. (Eds.). (1991). *Understanding political change: The British voter 1964–87*. Pergamon Press.
- Herrera, H., Morelli, M., & Palfrey, T. (2014). Turnout and power sharing. *Economic Journal*, 124(574), F131–F162.
- Herrera, H., Morelli, M., & Nunnari, S. (2016). Turnout across democracies. *American Journal of Political Science*, 60(3), 607–624.
- Herrman, M., Munzert, S., & Selb, P. (2016). Determining the effect of strategic voting on election results. *Journal of the Royal Statistical Society Series A*, 179(2), 583–605.
- Hicken, A. (2002). *Party systems, political institutions and policy: Policymaking in developing democracies*. Unpublished Ph.D. dissertation, University of California, San Diego.
- Hsieh, J., & Niemi, R. (1999). Can Duverger's law be extended to SNTV? The case of Taiwan's legislative yuan elections. *Electoral Studies*, 18(1), 101–116.
- Hug, S. (2001). *Altering party systems: Strategic behavior and the emergence of new political parties in Western democracies*. University of Michigan Press.
- Jackman, R. W. (1987). Political institutions and voter turnout in the industrial democracies. *American Political Science Review*, 81(2), 405–424.
- Jackman, R. W., & Miller, R. A. (1995). Voter turnout in the industrial democracies during the 1980s. *Comparative Political Studies*, 27(4), 467–492.
- Jacobson, G. C. (1980). *Money in congressional elections*. Yale University Press.
- Johnston, R. J., & Pattie, C. J. (1991). Tactical voting in Great Britain in 1983 and 1987: An alternative approach. *British Journal of Political Science*, 21(1), 95–108.
- Kiewiet, R. D. (2013). The ecology of tactical voting in Britain. *Journal of Elections, Public Opinion, and Parties*, 23(1), 86–110.
- Kim, H. M., & Fording, R. C. (2001). Does tactical voting matter? The political impact of tactical voting in recent British elections. *Comparative Political Studies*, 34(3), 294–311.
- Lanoue, D. J., & Bowler, S. (1992). The sources of tactical voting in British parliamentary elections, 1983–1987. *Political Behavior*, 14(2), 141–157.
- Lijphart, A. (1994). *Electoral systems and party systems: A study of twenty-seven democracies, 1945–1990*. Oxford University Press.
- Lijphart, A. (1997). Unequal participation: Democracy's unresolved dilemma. *American Political Science Review*, 91(1), 1–14.

- Lijphart, A., & Gibberd, R.W. (1977). Thresholds and payoffs in list systems of proportional representation. *European Journal of Political Research*, 5, 219–244.
- Lijphart, A., Pintor, R. L., & Stone, Y. (1986). The limited vote and the single nontransferable vote: Lessons from the Japanese and Spanish examples. In B. Grofman & A. Lijphart (Eds.), *Electoral laws and their political consequences* (pp. 154–169). Agathon Press.
- Lizzeri, A., & Persico, N. (2001). The provision of public goods under alternative electoral incentives. *American Economic Review*, 91(1), 225–239.
- Martin, D. (2021). Risk aversion and strategic voting. *International Journal of Public Opinion Research*, 33(3), 532–550.
- McCormick, R. P. (1975). Political development and the second American party system. In W. N. Chambers & W. D. Burnham (Eds.), *The American party systems* (2nd ed., pp. 90–116). Oxford University Press.
- Milesi-Ferretti, G. M., Perotti, R., & Rostagno, M. (2002). Electoral systems and public spending. *Quarterly Journal of Economics*, 117(2), 609–657.
- Mill, J. S. (1861). *Considerations on representative government*. Parker, Son, and Bourn.
- Morgenstern, S. (2001). Organized factions and disorganized parties: Electoral incentives in Uruguay. *Party Politics*, 7(2), 235–256.
- Moser, R. G. (1999). Electoral systems and the number of parties in postcommunist states. *World Politics*, 51(3), 359–384.
- Myatt, D. (2007). On the theory of strategic voting. *Review of Economic Studies*, 74(1), 255–281.
- Myatt, D., & Fisher, S. (2002). Tactical coordination in plurality electoral systems. *Oxford Review of Economic Policy*, 18(4), 504–522.
- Myerson, R. (1993). Incentives to cultivate favored minorities under alternative electoral systems. *American Political Science Review*, 87(4), 856–869.
- Myerson, R. (1995). Analysis of democratic institutions: Structure, conduct and performance. *Journal of Economic Perspectives*, 9(1), 77–89.
- Myerson, R. (1999). Theoretical comparisons of electoral systems. *European Economic Review*, 43(4–6), 671–697.
- Myerson, R., & Weber, R. (1993). A theory of voting equilibria. *American Political Science Review*, 87(1), 102–114.
- Niemi, R., & Hsieh, J. (2002). Counting candidates: An alternative to the effective N (with an application to the M + 1 rule in Japan). *Party Politics*, 8(1), 75–99.
- Niemi, R., Whitten, G., & Franklin, M. (1992). Constituency characteristics, individual characteristics and tactical voting in the 1987 British general election. *British Journal of Political Science*, 23(4), 549–563.
- Osborne, M., & Slivinski, A. (1996). A model of political competition with citizen candidates. *Quarterly Journal of Economics*, 111(1), 65–96.
- Palfrey, T. (1989). A mathematical proof of Duverger's law. In P. C. Ordeshook (Ed.), *Models of strategic choice in politics* (pp. 69–91). University of Michigan Press.
- Persson, T., & Tabellini, G. (2000). *Political economics: Explaining economic policy*. MIT Press.
- Powell, G. B., Jr. (1980). Voting turnout in thirty democracies: Partisan, legal, and socio-economic influences. In R. Rose (Ed.), *Electoral participation: A comparative analysis* (pp. 5–34). Sage Press.
- Powell, G. B., Jr. (1982). *Contemporary democracies: Participation, stability, and violence*. Harvard University Press.
- Powell, G. B., Jr. (1986). American voter turnout in comparative perspective. *American Political Science Review*, 80(1), 17–43.
- Rae, D. (1971). *The political consequences of electoral Laws* (rev. ed.). Yale University Press.
- Reed, S. (1990). Structure and behavior: Extending Duverger's law to the Japanese case. *British Journal of Political Science*, 20(3), 335–356.
- Reed, S., & Thies, M. (2001). The causes of electoral reform in Japan. In M. S. Shugart & M. P. Wattenberg (Eds.), *Mixed-member electoral systems: The best of both worlds?* (pp. 152–172). Oxford University Press.

- Rosenstone, S. J., Behr, R., & Lazarus, E. (1996). *Third parties in America: Citizen response to major party failure* (2nd ed.). Princeton University Press.
- Samuels, D. J. (1998). *Careerism and its consequences: Federalism, elections, and policy-making in Brazil*. Unpublished Ph.D. dissertation, University of California at San Diego.
- Schumpeter, J. (1942). *Capitalism, socialism, and democracy*. Harper.
- Shachar, R., & Nalebuff, B. (1999). Follow the leader: Theory and evidence on political participation. *American Economic Review*, 89(3), 525–547.
- Shugart, M. S. (2001). ‘Extreme’ electoral systems and the appeal of the mixed-member alternative. In M. S. Shugart & M. P. Wattenberg (Eds.), *Mixed-member electoral systems: The best of both worlds?* (pp. 25–54). Oxford University Press.
- Shugart, M. S., & Wattenberg, M. P. (Eds.). (2001). *Mixed-member electoral systems: The best of both worlds?* Oxford University Press.
- Smith, D. M. (2018). Electoral systems and turnout. In E. Herron, R. Pekkanen, & M. Shugart (Eds.), *The Oxford handbook of electoral systems* (pp. 193–212). Oxford University Press.
- Taagepera, R., & Shugart, M. S. (1989). *Seats and votes: The effects and determinants of electoral systems*. Yale University Press.
- Tillman, E. R. (2015). Pre-electoral coalitions and voter turnout. *Party Politics*, 21(5), 726–737.
- Tsebelis, G. (2002). *Veto players: How political institutions work*. Princeton University Press.
- Wilson, F. L. (1980). Sources of party transformation: The case of France. In P. H. Merkl (Ed.), *Western European party systems* (pp. 526–551). Free Press.
- Zielinski, J. (2002). Translating social cleavages into party systems: The significance of new democracies. *World Politics*, 54(2), 184–211.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Presidential, Parliamentary, and Hybrid Forms of Government

7

John M. Carey and Juan Pablo Micozzi

Contents

7.1	Introduction	130
7.2	The Academic Consensus Against Presidentialism	133
7.3	Theoretical and Empirical Challenges: Neither the Worst, Nor the Best, Nor the Only Choice?	134
7.4	Research on Institutional Design in Parliamentary and Presidential Systems	137
7.4.1	Institutions Under Parliamentarism	137
7.4.2	Partisan Compatibility of Presidents and Assemblies	138
7.4.3	Methods of Electing Presidents	139
7.4.4	Electoral Cycles and Legislative Rules	140
7.4.5	Policymaking Powers of Presidents	141
7.4.6	Decree	142
7.4.7	Agenda Powers	144
7.4.8	Veto	145
7.5	Used Presidents and Government Crises: “Should I Stay or Should I Go Now?”	147
7.5.1	Reelection	148
7.5.2	Government Crises and the Parliamentarization of Presidential Systems	150
7.6	Conclusion	153
	References	155

Abstract

Foundational scholarship on constitutional regime type favors parliamentarism over presidentialism on grounds that the former fosters democratic stability by

We thank Gabriel Negretto for generous data sharing.

J. M. Carey (✉)

Department of Government, Dartmouth College, Hanover, NH, USA

e-mail: john.m.carey@dartmouth.edu

J. P. Micozzi

Instituto Tecnológico Autónomo de México (ITAM), Mexico City, Mexico

e-mail: juan.micozzi@itam.mx

better aligning the interests of legislatures and executives. Yet, presidentialism and constitutional hybrids that combine a popularly elected president with a cabinet subject to parliamentary confidence remain abundant even among regimes that engaged in constitutional engineering in recent decades. This chapter describes the key features that distinguish these regime types and how they vary empirically among democracies. We focus especially on how presidents are elected and on the policymaking powers they wield, such as decree authority, control of the legislative agenda, and the veto. We review two trends from recent decades—loosening restrictions on presidential reelection and legislative replacement of presidents during government crises—that suggest convergence with parliamentarism in practice.

Keywords

Parliamentarism · Presidentialism · Semi-presidential · Hybrid regime · Democracy · Constitution · Reelection · Decree · Veto

7.1 Introduction

As the number of democracies increased during the late twentieth century and then as threats to the resilience and performance of those democracies presented themselves in the first decades of the twenty-first, scholarly interest has focused on the design of formal institutional rules by which citizens select political leaders and by which elected officials interact to form governments and policy. Central to this scholarship was the study of variation in the institutions ruling the origin, interactions, and termination of the branches of government and how these arrangements affect governance, long-term policy implementation, the management of political conflict, and chances of regime consolidation. Among the central debates is on the relative advantages of a presidential or a parliamentary system. During this period, hybrid constitutional formats that combine some aspects of both presidentialism and parliamentarism also became more common.

We start our discussion establishing some key terminology regarding how the popular branches of government—the legislature and the executive—are selected and terminated and how they interact to form policy and administer the government. Assemblies—variously known as congresses, parliaments, legislatures, or a host of country-specific names—are popularly elected in all democracies, but executives are not. The general characteristics of parliamentary and presidential regimes are as follows.

Parliamentarism

- The executive is dual (Head of State/Head of Government).
- The Head of Government is selected by the assembly.
- The Head of Government remains in office subject to legislative confidence.
- The Head of Government can terminate the mandate of the assembly, subject to case-specific variations.

Presidentialism

- The chief executive (usually one single figure) is popularly elected.
- The terms of the chief executive and of the assembly are fixed and not subject to mutual confidence.
- The elected executive appoints and leads the government (cabinet) and has some constitutionally granted lawmaking authority.

The key principles that distinguish parliamentary and presidential government entail the *origin* and the *survival* of the popular branches. Under parliamentarism, only the assembly is elected, and cabinet members are selected from within the Parliament by the floor members. Thus, the origin of the executive is derivative to that of the assembly and will remain in place as long as the confidence of a majority is granted. In most parliamentary systems, moreover, this dependence is mutual, as the branches have proactive (no-confidence votes, parliamentary dissolution, *guillotine*) and reactive (confidence requests, parliamentary denial of this request) powers that may incite the early end of mandates and the call to new elections. Thus, parliamentarism is frequently distinguished from presidentialism on the grounds that powers are fused, rather than separated, and have flexible, instead of rigid, mandates.

Under presidentialism, the origins of the two branches are electorally distinct, with the chief executive (always the president(s) and sometimes one or more vice presidents as well) elected separately from the assembly, for a fixed term.¹ The last element in the definition of presidentialism is simply that this elected president exerts substantial powers over the rest of the executive branch—the ministries—and over the lawmaking process. This distinguishes presidential regimes from those republics that elect a ceremonial Head of State who may be called a president but who lacks considerable constitutional authority (e.g., Ireland, Germany, or Italy).²

If the principles by which the executive and assembly are founded and operated are distinct under presidentialism and parliamentarism, there are also several constitutional regimes that combine elements from both ideal types. Given there is no unanimous consensus in the literature about how to label them (i.e., Duverger, 1980 considers them semi-presidential, while Shugart & Carey, 1992 & Elgie, 2011 distinguish premier-presidential and president-parliamentarism), we opt to keep the hybrid denomination. At the broadest level, these hybrid regimes have the following characteristics:

- The president is popularly elected and is endowed with meaningful powers.
- There also exists a prime minister and cabinet, subject to assembly confidence.

¹There are historical experiences of collegiate presidencies, such as Uruguay or Cyprus. See Altman, (2020).

²There is variation in the selection mechanisms, ranging from direct popular choice and an ad hoc election of a share of delegates, to the decision of the legislative assembly.

Table 7.1 Constitutional regime type among democracies, 2022

	Parliamentary	Presidential	Hybrid
Americas	Barbados, Canada, Guyana, Jamaica, Suriname, Trinidad and Tobago	Argentina, Bolivia, Brazil, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, Honduras, Mexico, Panama, Paraguay, Peru, Uruguay, United States	
Postcommunist	Armenia, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Estonia, Georgia, Kosovo, Latvia, Lithuania, Moldova, Mongolia, Montenegro, North Macedonia, Romania, Slovakia, Slovenia		Poland
Western Europe and Oceania	Australia, Austria, Belgium, Denmark, Finland, Germany, Greece, Iceland, Ireland, Italy, Luxembourg, Malta, Netherlands, Norway, Spain, Sweden, Switzerland, United Kingdom	Cyprus	Portugal, France
Asia	Bhutan, Israel, Japan, Nepal, Solomon Islands, Vanuatu	Indonesia, Maldives, South Korea, Sri Lanka	Taiwan, Timor-Leste
Africa	Botswana, Cape Verde, Lesotho, Mauritius, Sao Tome and Principe, South Africa	Gambia, Ghana, Kenya, Liberia, Malawi, Namibia, Seychelles, Sierra Leone, Zambia	Niger, Senegal

Countries ranked above 0.5 in V-Dem's electoral democracy index

This broad definition describes well a wide range of hybrids where the specific powers of the elected president, and their relationship to the prime minister and cabinet, vary considerably. Table 7.1 charts the systems of government for the 90 political systems that reach or exceed the 0.5 threshold of V-Dem's electoral democracy index.³ Classification was not an easy endeavor. For the sake of consistency and transparency, we followed a series of steps that can be easily replicated. Based on V-Dem's scores for 2022, we first distinguished those cases where the variable Head of Government was null and recognized them as presidential, subject to further validation. After realizing that 5 of these 34 coded observations did not have an elected president, we removed them from the sample and recoded back as

³The V-Dem (or Varieties of Democracy) Institute provides the most comprehensive and thoroughly documented annual ratings of democratic performance worldwide. The index ranges from 0 to 1.0, and countries rated above 0.5 "tend to be considered electoral democracies" (Coppedge et al., 2021, p.6). Conventional empirical pieces (see Lindberg, 2016; Luhrmann et al., 2018) have adopted the same threshold.

null. Then, those 16 cases where the Head of State is a King, a Queen, an Emperor, a Grand Duke, or a Governor-General and also have a separate Head of Government were considered parliamentary. Using a similar logic, those remaining countries where the Head of State was not popularly elected (i.e., by the legislative assembly or indirect procedures) were also coded as parliamentary. The hardest task at this point was the distinction among systems using confidence requirements for their Head of Government which include popular elections for their Head of State. As mentioned, aside from electoral legitimacy, presidents need substantial powers for a system to be considered hybrid. For that reason, we relied on V-Dem's variables measuring the most relevant attributes in the hands of presidents. We created a variable, *presidential power*, equaling the mean of the indicators capturing whether the Head of State has dissolution powers, appoints the cabinet in practice, holds veto powers, and can get rid of ministers at discretion. Based on those scores, we considered *hybrid* these countries in which the cabinet is subject to parliamentary confidence and where Heads of State have an equal or higher value than the lowest presidential country (Liberia with 1.5 in our sample).

This chapter reviews the debate over the relative advantages and disadvantages of various constitutional frameworks, the characteristics of systems that have been of particular interest to academics and reformers, and some recent trends in the design and performance of regimes. Most of the focus is on presidentialism and its contrast with the other systems, for a couple of reasons. First, whereas pure parliamentary democracies are clustered among the relatively prosperous and politically stable Organisation for Economic Co-operation and Development (OECD) countries, presidential designs are more common among newer democracies and among countries that have experienced more political and constitutional instability. Thus, most of the empirical action in constitutional design in recent decades took place in cases with elected presidencies. Second, in line with this empirical trend, academic debate over systems of government has focused on the design and performance of systems with presidencies. Among systems with consequential presidencies, we devote more attention to Latin American cases than to presidencies elsewhere, again for a couple of reasons. First, the presidential tradition is oldest in the Americas, so the wealth of empirical material is there. Second, our own research experience is primarily in the Latin American area.

7.2 The Academic Consensus Against Presidentialism

The debate over systems of government was largely triggered by the transitions to democracy in Latin America after the lengthy experience of many countries in the region with military authoritarian regimes from the 1960s through the 1980s. The process of reestablishing civilian government raised the question of whether *dysfunctional* institutions could promote systematic failures of democratic governments. In this sense, Latin America's presidential heritage was a potential source of democratic failure. Juan Linz (1994), who had already witnessed the Spanish and Portuguese democratic transitions in the 1970s, was the pioneer in arguing that presidentialism was inherently more prone to democratic breakdown than

parliamentarism and that Latin American elites should channel their novel democracies through parliamentary constitutions, in order to avoid several by-design-induced complications. His reasons were manifold.

First, Linz contended that presidentialism fosters zero-sum politics and winner-takes-all dynamics. As politicians just care about the major, indivisible prize (the presidency), the best strategy in equilibrium for $N-1$ actors is the failure of the current ruler. Thus, cooperative behavior is discouraged. This could become especially risky in moments of political crisis and presidents' low popularity, boosted by Linz's second main criticism: dual legitimacy and the lack of an institutional safety valve. As mentioned above, both the presidency and the legislative branch have popular legitimacy for a fixed time period, and there is no clear basis to declare one branch's mandate as more legitimate than the other's. In other words, both branches could claim to be depositories of the popular will. Given the absence of institutional mechanisms to resolve interbranch conflict, according to Linz, military coups, self-coups, or presidential initiatives to bypass Congress are frequent outcomes. Unlike parliamentarism, where the Prime Minister could dissolve the Parliament or the legislators could censor the cabinet or deny a confidence vote, presidentialism just has a more blunt and cumbersome tool, impeachment, which generally demands supermajorities and a reason other than mere policy opposition to remove the executive from office.

Finally, Linz contends that popular election fosters a presidential self-aggrandizement that could encourage executives to challenge other branches' checks and balances, opening the door to democratic backsliding and eventually authoritarianism. According to the classic Linzian critique, then, presidentialism is bad and dysfunctional for democracies.

Empirically, the breakdown of several Latin American democracies in the 1960s and 1970s supported Linz's case for the failure of presidential government. In Brazil in 1964, Peru in 1968, Chile in 1973, Uruguay in 1974, and Argentina in 1976, episodes of legislative-executive conflict had preceded military intervention in politics that displaced civilian leaders, imposing long periods of authoritarian rule. Scholars endorsed Linz's arguments with country case studies (Giuseppe DiPalma, 1990; Bolivar Lamounier, 1993; Arend Lijphart, 1990, 1999; Arturo Valenzuela, 1994; Giovanni Sartori, 1994). And broad cross-national studies based on quantitative data supported the proposition that, even controlling for factors such as economic development and colonial history, presidential regimes are more inclined toward democratic breakdown than are parliamentary ones (Alfred Stepan & Cindy Skach, 1993; Adam Przeworski & Fernando Limongi, 1997).

7.3 Theoretical and Empirical Challenges: Neither the Worst, Nor the Best, Nor the Only Choice?

By the time a near consensus in favor of parliamentarism had formed among scholars of comparative politics, real-world events exposed political scientists to different realities and forced theoretical challenges to assertions made by Linz. Increases in the number of cases implied more data and strengthened the needs of utilizing longer time series to test conventional propositions about the effects of

institutional design. Shugart & Carey (1992) and Mainwaring & Shugart (1997) suggested that some of Linz's assertions were based on a selection bias (pre-World War II parliamentary breakdowns were not considered), that there was high collinearity between systems of government's success and wealth, and that it was not true that every presidential system banned executive reelection. Linz's characterization of presidentialism as a zero-sum game (actors just care about the presidency) also sat uneasily with his emphasis on the risks of gridlock and deadlock (which suggests that legislatures matter!). Subsequent scholarship put a central idea at the core: interbranch dynamics are likely to be contentious in presidential (and hybrid) democracies, but the fine print might reveal what institutional arrangements allow presidentialism to work.

Following this line, after decades of presidentialism being scrutinized, scholars started working at lower levels of analysis, assessing whether *within* system variations could prompt different outcomes at the regime level. Specifically, studies tried to answer if specific institutional combinations might make this system less stable. Mainwaring (1993) and Mainwaring & Shugart (1997) highlighted the role of party system fragmentation and levels of discipline to explain government and (likely) regime stability. Mark Jones (1995, 1999) centered his studies in how electoral rules affect multi-partism, cohesion, and democratic survival. Matthew Shugart (1995) added the consideration of the electoral cycle to understand patterns of divided government and therefore government's gridlock. Barry Ames (1995, 2002) raised the interaction between electoral rules and the vertical distribution of power (i.e., federalism) to explain variation in governance and stability. Cheibub (2002) analyzed the impact of rules over fragmentation and minority status of executives and found no significant relationship with regime maintenance.

The most recent decades show some changing patterns in institutional design. On the one hand, Latin America witnessed some modest constitutional moves toward the principle of legislative confidence after the 1990s. The requirement of assembly confidence for ministers is not unprecedented in the region. Although presidents have never been subject to confidence, cabinets were during Chile's period of "parliamentary republic" in the late nineteenth and early twentieth centuries. In a 1994 reform, Argentina created a new office called Chief of the Cabinet of Ministers, subject to parliamentary confidence, although with limited practical influence. Venezuela's 1999 constitution established a similar position, called Prime Minister, and the current rules let the assembly dismiss any cabinet member by a 3/5 majority. Peru's 1993 constitution empowered a Prime Minister named by the president who can be censored by a congressional majority, and the president can dissolve the legislature if its members refuse to confirm the cabinet after three attempts. Ecuador's 2008 constitution included a mechanism called *muerte cruzada*, by which the president can dissolve the legislature (as president Lasso did in 2023) at the expense of his own removal and subsequent call for general elections (Polga-Hecimovich, 2020). However, other reforms put strong formal powers in the hands of executives throughout the region, such as emergency decrees (Carey & Shugart, 1998), preferential legislative agenda setting (Magar et al., 2021), or direct democracy devices (Altman, 2010).

Outside Latin America, the trend during the 1990s was toward the creation of powerful presidencies—a pattern reinforced by the sudden proliferation of new, postcommunist regimes in central Europe and the former Soviet Union. The greatest concentrations of formal presidential power are found among the most autocratic post-Soviet systems, like Russia and Belarus, as well as whose credentials as democracies are most dubious, particularly in the so-called Stans of Central Asia.⁴ Elected presidents, however, are central to politics in many of the more democratic postcommunist systems as well, including Poland, Bulgaria, Romania, Lithuania, Croatia, Serbia, Georgia, Ukraine, and Moldova (Timothy Frye, 1997; Ishiyama and Kennedy, 2001). Finally, during the late 1980s and 1990s, transitions from authoritarian rule to democracy in South Korea, Taiwan, and the Philippines produced a crop of new, or renewed, regimes with powerful presidencies in East Asia, along with brand new cases from other regions such as Botswana, Zambia, Senegal, Niger or Suriname. As a reference, although the office crafted for Charles de Gaulle has been considered a landmark powerful presidency, presidential powers in the aforementioned cases rank higher than those of the French presidency in our V-Dem-based index.

Many postcommunist regimes (especially those out of our sample for their democratic levels), as well as the Taiwanese and several African cases, are hybrids, combining elected presidents endowed with substantial constitutional authority with prime ministers and cabinets that are subject to parliamentary confidence (Frye, 1997; Robert Elgie, 2011). Such regimes are characterized in some accounts as *semi-presidential*—a term originally coined by Maurice Duverger (1980) to describe the 1958 constitution of the French Fifth Republic. None of the new hybrids, however, as nearly approximates the parliamentary ideal as does France, where the president acts as *formateur*, and retains the authority to dissolve the assembly, but exercises unilateral control over neither the composition of the cabinet nor the legislative agenda nor wields a legislative veto. The new hybrids of the 1990s are constitutionally distinct from the French system in important ways. First, in most cases, presidents can name *and* remove the prime minister, rendering this post primarily a creature of the president rather than an agent of assembly majorities. Second, the hybrid regimes of the 1990s endow presidents with formidable legislative powers—including vetoes that require extraordinary legislative majorities for override and in many cases the authority to issue decrees with the force of law. These powers make presidents, whose origin and survival in office remain independent from assembly confidence, central players in the lawmaking process in the new hybrid systems. Contemporary hybrid regimes do consistently vary in each Head of Government's prerogatives, and their results are also very dissimilar.

While scholarly and policy specialists in the early 1990s praised the advantages of hybrid systems, for their potential to provide flexibility with stability, in practice, many of these polities' political performance has been as chaotic as the most fragmented presidential and parliamentary systems. Scholars showed that, beyond formal

⁴Note that regimes with low levels of democracy (V-Dem indices below 0.5) as of 2022 are not listed in Table 7.1.

prerogatives in presidents' hands, hybrid systems exhibit variation in the atomization of party systems, the median survival of governments, the frequency of cohabitations (i.e., presidents and Premiers belonging to different parties), duration of bargaining for cabinet formation, and frequency of dissolution and call to early parliamentary elections, among others (Jose Cheibub & Svetlana Chemykh, 2008, Jorge Fernandes & Pedro Magalhaes, 2016). Anecdotal evidence provides colorful insights to some of the features of these hybrid regimes. Making use of the strong formal powers endowed to the president by the Russian constitution and his popularity, Vladimir Putin could shape policy at almost his own discretion. Facing his constitutional term limit in 2008, Putin vacated the office for Dmitry Medvedev but continued to exercise *de facto* executive authority while serving formally as Prime Minister for a single term. Putin then returned to the presidency in 2012 and has not relinquished that office since. In neighboring Ukraine, one of the many political conflicts between the president and the premier was solved through the impeachment of President Viktor Yanukovich, accused of abandoning his country, by the parliament in 2014. Paradoxically, decades after Linz's *dire* forecast, the competing legitimacies problem was resolved, in this hybrid regime, by the safety valve of presidential removal.

To sum up, across the various regimes that engaged in constitutional engineering in the latter decades of the twentieth century, there was a strong inclination to create and to retain powerful, elected presidencies. There was, at the same time, a trend toward acknowledging some role for assemblies in exercising confidence authority over at least some part of the executive. These reforms do not appear to have put to rest the fundamental challenges of conflict between the elected branches, however, and we continue to observe instances of constitutional improvisation.

7.4 Research on Institutional Design in Parliamentary and Presidential Systems

7.4.1 Institutions Under Parliamentarism

During the past decades, the academic literature on democratic institutions has expanded enormously. Among scholars focusing primarily on parliamentary systems, electoral systems have been the element of institutional design attracting the most scrutiny. Rein Taagepera and Shugart (1989), Lijphart (1994), and Gary Cox (1997) all provide landmark investigations of the effects of electoral rules on representation. The main focus within the electoral systems literature, moreover, is on the effects of different rules for selecting multimember bodies—primarily parliaments—the rules for which exhibit more variance than those for selecting presidents. We discuss the interaction of presidential and legislative electoral rules below, but the electoral systems literature more generally is discussed independently in Cox's contribution to this volume. The effects of varying electoral institutions over the composition of plural bodies raised scholars' interest for their implications on government formation, especially the environment for coalition-building. Social choice theory was a key motivation for studies on parliamentary dynamics, centered on dilemmas of instability of decisions under majority rule (Condorcet, 1789;

Kenneth Arrow, 1951; Richard McKelvey, 1976; William Riker, 1982). The suspicions raised by social choice theorists about the coherence of majority rule in the presence of multiple issue dimensions apply in a straightforward manner to parliamentary government, particularly where majorities in a single legislative chamber are sovereign over national governments and where party system fragmentation and multiple social cleavages create the potential for fluid majority coalitions. Much of the initial literature on coalition formation in parliamentary systems motivated by social choice theory is reviewed in Michael Laver and Norman Schofield (1990). In the subsequent decades, advances in the study of parliamentary government have examined in greater detail the rules and practices that determine how authority within governments is allocated, in an effort to understand not only which governments form but the implications of this for the policies they produce (Terry Moe & Michael Caldwell, 1994; Kathleen Bawn, 1999; John Huber, 1998). Scholars have developed agendas centered on the determinants of coalition formation (Lanny Martin & Randy Stevenson, 2001; Martin & Vanberg, 2003; Craig Volden & Cliff Carrubba, 2004; Hanna Back et al., 2011), policymaking under coalition governments (Martin and Vanberg, 2003, 2004), portfolio allocation and government stability (Royce Carroll & Gary Cox, 2007, John Huber & Cecilia Martinez-Gallardo, 2008), governance under minority governments (Jose Cheibub et al., 2021; Bonnie Field & Shane Martin, 2022), and the determinants of government termination (Daniel Diermeier & Randy Stevenson, 2000; Kaare Strøm & Stephen Swindle, 2002; Petra Schleiter & Edward Morgan-Jones, 2009), among many relevant topics. During this same period, scholars recognized that partisan coalition-building within presidentialism was driven by many of the same factors as under parliamentarism, incorporating insights from research that initially focused on parliamentary democracy, especially regarding the increased frequency in which presidents could not count on disciplined majority support in Congress. As scholarship identified, with increasing precision, the conditions for cooperative power sharing within parliamentary coalitions, studies of presidential government investigated whether those conditions applied more generally.

7.4.2 Partisan Compatibility of Presidents and Assemblies

Following the previous line, empirical work on presidential and hybrid regimes has focused on the mechanics of relations between the executive and legislative branches. Pursuant to the battery of criticisms directed at presidentialism, research focused on whether presidentialism inherently undermines bargaining and cooperation between the branches and, if so, to what effect. An initial reaction was to point out that there is tremendous variance among systems with presidents in terms of institutional design and to examine whether these details could help explain regime performance (Shugart & Carey, 1992; Scott Mainwaring & Shugart, 1997). One of the central themes in this research is the interaction between presidential and legislative elections and the implications for party systems. Understanding the sources of divided government in the United States is, of course, a longstanding cottage industry (Morris Fiorina, 1996; Walter Mebane & Jasjeet Sekhon, 2002).

But the diversity of electoral arrangements across presidential systems provides a richer environment for examining institutional explanations for partisan conflict or compatibility across branches. Moreover, one of the central questions raised by presidentialism critics is whether regime crises are frequent when presidents lack majority support in Congress (Mainwaring, 1993; Cheibub, 2002).

A critical turning point for comparative analyses of presidential systems was Scott Morgenstern and Benito Nacif's (2002) *Legislative Politics in Latin America*. Its epilogue, developed by Gary Cox and Morgenstern, offered an original analytical frame that encourages simultaneous analyses of presidents' and assemblies' strategies. Executives know in advance what kind of legislature they are interacting with and thus need to anticipate its members' preferences to advance with the right strategy for policymaking. If a president counts with a legislative majority that will agree with her initiatives without major hesitations, she will be *dominant* and the assembly subservient. On the other extreme, recalcitrant legislatures that systematically oppose the executive's proposals force presidents to behave as *imperial*, relying on decrees or measures of exception to overcome the obstruction. In the middle, if legislatures' cooperation depends on different preferences/payoffs, executives need to calculate their offers. Whereas policy-oriented legislators demand influence in public outcomes, they can be seduced by *coalitional* presidents who open their cabinets and offer influential portfolios. Instead, if legislators demand local, targeted policies (i.e., pork, public works at the subnational level, or budget influence) in exchange for their support to the executive's proposals, then a *nationally oriented* president would bargain with a parochial assembly. Thus, consideration of the law of anticipated reactions (Friedrich, 1963) might be the key to refine Linz's diagnoses, and presidential systems might work appropriately if the right interbranch strategies are enacted.

Of course, incentives for each set of actions/reactions heavily rely on the available resources to each of the elected branches. This realization drives us back to the seminal Shugart and Carey's (1992) work, where the main sources of power are distinguished between partisan and constitutional prerogatives. While the former depend on the distribution, cohesion, and loyalty of legislators (a function of electoral systems, itself), the latter are determined by written rules. Therefore, analyses of both sources of power have been present in the literature. Regarding constitutional rules, two key factors associated with the system for electing presidents are central to this research: the formula for electing presidents and the relative timing of presidential and legislative elections.

7.4.3 Methods of Electing Presidents

The two most common formulas for electing presidents are the plurality and the majority runoff systems. Under plurality rule, there is one election, and the candidate with the most votes is elected. Under the majority runoff system, if no candidate wins an outright majority of the vote in the first round, there is a second round pitting the top first-round candidates against each other. In addition, we can see in Latin America examples of the so-called Double Complement Rule (Shugart & Taagepera, 1989),

where lower thresholds are required to avoid a runoff. For example, Costa Rica declares a candidate as the winner if she gets 40% of valid votes. In Argentina, a candidate who reaches 45%, with a margin of at least 10% above the next most popular candidate wins in the first round. Ecuador uses a similar system to that of Argentina but lowers the threshold to 40% of the votes. Research on two-round methods for electing presidents suggests that the format contributes to the proliferation of presidential candidates and the fragmentation of the first-round presidential vote (Shugart & Carey, 1992; Mark Jones, 1995; Mainwaring & Shugart, 1997; Cox, 1997).

The rationale rests on both the differing incentives for strategic voting under the two systems (Maurice Duverger, 1954; Cox, 1997) and the opportunities for between-rounds bargaining unique to runoff elections for presidential office. Under plurality elections, where the threshold for success is high, the best strategy for a presidential aspirant who cannot reasonably expect to win the most votes is to enter a pre-election coalition with a viable candidate. Under a runoff system, on the other hand, the initial threshold is lower, as the second-place, first-round candidate survives. Moreover, given that electoral coalitions can be renegotiated after the first round in anticipation of the runoff, even nonviable candidates are induced to compete in the first round to establish the value of their second-round endorsement. The more marginal candidates enter in the first round, moreover, the more fragmented the expected vote, and the more unpredictable the results become.

By encouraging fragmentation and unpredictability, majority runoff elections are associated with a favorable environment for outsider candidates and minority presidencies. Prominent cases in which outsider candidates built surprising first-round performances into second-round victories include Fernando Collor in Brazil in 1989 and Jorge Serrano in Guatemala, along with Alberto Fujimori in 1990 and Pedro P. Kuczynski in 2016, both in Peru. Each of these candidates ran against the traditional party system in his country, survived the first round with less than one-third of the vote, and won the presidency in a runoff. Despite their second-round “mandates,” moreover, each of these presidents quickly found themselves mired in interbranch conflicts which culminated in the premature deaths of Collor’s, Serrano’s, and Kuczynski’s presidencies and of Peru’s Congress with Fujimori. In this sense, it could also be argued that runoff systems, fostering fragmentation, could increase the chances of one of the controversial outcomes of presidentialism: the impeachment of presidents endowed with poor legislative support (without a *legislative shield*) (Aníbal Pérez Liñán, 2007). Of course, this is also influenced by the electoral cycle and the relative timing of executive and legislative elections.

7.4.4 Electoral Cycles and Legislative Rules

A second key element of institutional design is the relative timing of executive and legislative elections—the electoral cycle—which interacts with the presidential electoral formula to affect the prospects for partisan alignment between the branches. Where elections are held on dyssynchronous schedules, campaigns are conducted independently, and the contours of the legislative party system are not bound to

presidential elections. Where elections for the presidency and those for the assembly are concurrent, presidential candidates' electoral coattails can affect fragmentation in assembly elections and therefore the partisan distribution of seats (Jones, 1995; Shugart, 1995). Whereas plurality elections for the president can reduce fragmentation of legislative party systems, this effect is absent under the two-round format. As a result, the likelihood of divided government—that is, of presidents who lack the support of partisan majorities in the assembly—is higher under nonconcurrent than under concurrent electoral cycles and, among systems with concurrent cycles, is higher when presidents are elected by majority runoff than by plurality (Mainwaring & Shugart, 1997). Empirically speaking, most elections in Latin America tend to be concurrent (Mark Payne et al., 2007), subject to specific variations triggered by the kind of legislative renewal (full or partial) and constitutional provisions. Often, legislative elections are held concurrently with the executive election (or with the first round in two-round systems). If runoff mechanisms foster fragmentation, these trends tend to be mirrored in the legislative branch and thus exacerbate fragmentation. As a curious exception, Colombia elects its legislators several months before the presidential poll.

The effects of the electoral cycle and the executive rules are also filtered by the legislative electoral systems. While this chapter does not get involved in depth with electoral rules' details, it is important to note that most countries in the region use some kind proportional representation.⁵ As stated in the literature (Mainwaring & Shugart, 1997; Jones, 1995), fragmented congresses exacerbate presidents' problems to move policy from the status quo, giving rise to what Mainwaring (1993) characterized as the *difficult combination* of presidentialism with a fragmented legislative party system.

7.4.5 Policymaking Powers of Presidents

Partisan support in the legislature is critical to the effectiveness of presidents in realizing their agenda. Indeed, dominance over policymaking is frequently attributed to presidents with relatively limited formal authority over standard lawmaking procedures, provided that their party or coalition controls the legislature and that the president is the effective leader of his party (Lynn Kelley, 1973; Daniel Levine, 1973; Jeffrey Weldon, 1997). In such cases, it is not the formal authority of the presidency that accounts for executive influence but rather the coincidence of the presidency with party leadership in the same individual. In other cases, however, formal authorities are enshrined in the office of the presidency, independent of partisan support. The specific nature of these powers varies considerably across presidencies, but here, we briefly discuss three common types of presidential power over legislation: decree, agenda setting, and veto authority.

⁵Even in the region's mixed-member systems (for example, Bolivia, Venezuela, and Mexico), the proportional tier allows multi-partism in the assembly.

Table 7.2 Examples of presidential decree authority

		Decree becomes permanent law?	
		Yes	No
Decree in effect immediately?	Yes	Russia (Art. 90) Peru (Art. 118(19)) Colombia (Art. 215)	Brazil (Art. 62) Colombia (Art. 213) Argentina (Art. 99.3)
	No	Ecuador (Art. 65) <i>France (Art. 49.3)^a</i>	

^aRefers to power vested in the premier, not the president, to make proposals under the *guillotine* procedure

7.4.6 Decree

Decree is the authority of the executive to establish law in lieu of action by the assembly. This may include executive policy initiatives that eventually require legislative ratification, provided the initiatives go into effect without prior legislative action (Carey & Shugart, 1998). Empirically, then, constitutional decree authority of executives varies according to whether the initiatives

- Are effective as policy immediately (yes/no).
- Become permanent law even without legislative action (yes/no).

The four possible combinations form a 2×2 matrix, shown in Table 7.2, with empirical examples in each box. The formal procedures attached to decree authority determine whether there is an opportunity for assembly debate on a measure before it becomes law and whether the assembly must take explicit action to rescind the measure.

At the top left of Table 7.2 is the prototypical decree authority whereby the executive issues a proposal that becomes permanent law immediately and without any legislative action. Executives can effectively present policy initiatives under this format as *fait accompli*, and only through the passage of new legislation (or a new decree) can the policy be altered. Very few democratic constitutions grant presidents such power, and those that do generally include some constraints on the policy jurisdictions in which executives may exercise decree.

The Colombian president may use decree to “restore economic order” and the Peruvian “on economic and financial matters, when so required by the national interest.” Presidents in both countries have, in practice, interpreted these powers expansively in setting economic policy.

The top right box in Table 7.2 represents *provisional* decree authority in which executive proposals take effect immediately but lapse after some designated period unless ratified by the legislature.⁶ In Brazil, presidential decrees lapse after 30 days; in Colombia, decrees other than those “to restore economic order” lapse after a

⁶This type of decree authority is not unique to presidential government. For example, the Italian cabinet can issue decrees with immediate force but which lapse after 60 days if not ratified by parliament (Art. 77)

maximum of 180 days. Whether such a provision is effective for initiating long-term changes even against legislative opposition depends in part on whether the decrees may be reissued at the end of the period.

The bottom left of Table 7.2 describes *delayed* decree, whereby executive proposals do not take effect immediately but become law *unless* the legislature acts to reject them. In Ecuador, for example, the president can propose legislation, declaring it “urgent,” and if Congress fails to act within 15 days, the proposal becomes law. The format is similar to France’s *guillotine* (Art. 49.3), whereby if parliament rejects the executive’s proposal, then the government falls; but if parliament takes no action, the proposal becomes law. It is important to note, however, that the authority over the *guillotine* is vested in the French premier, not the president, consistent with France’s relatively parliamentary hybrid regime.

Table 7.3 replicates Gabriel Negretto’s (2013) categorization of decree powers in Latin America according to the scope of their effect (none, restricted to specific

Table 7.3 Scope and conditions for reversion of decrees in Latin America

Decree authority and restrictions in content	No decree	Restricted scope	No restrictions
	Bolivia Costa Rica Chile Dominican Republic Ecuador El Salvador Guatemala Honduras Mexico Nicaragua Panama Paraguay Uruguay Venezuela	Argentina Brazil Colombia Peru	
Decree authority and reversion	No decree	Vanishes without congressional approval	Stands without congressional approval
	Bolivia Costa Rica Chile Dominican Republic Ecuador El Salvador Guatemala Honduras Mexico Nicaragua Panama Paraguay Uruguay Venezuela	Brazil	Argentina Colombia Peru

Source: Negretto (2013)

areas, and minimum restrictions) and their validity vis-à-vis congressional action (none, decree vanishes without legislative ratification, or it holds unless it is vetoed by the other branch). The table underscores that most presidents lack strong autonomous decree authority, and whenever they have it, it is constrained. This may sound counterintuitive to observers of the many reforms implemented by decree in the 1990s and the 2000s. But many such presidential initiatives were undertaken via Delegated Decree Authority (Carey & Shugart, 1998), which involves the explicit delegation of authority, by statute, from a congressional majority to the president and therefore is not the product of the president's constitutional authority.

Its frequency aside, the use of decree authority has been central to conflicts between legislatures and executives that have generated constitutional crises in a number of countries. In the first years of Fujimori's administration in Peru, the president's increasing reliance on decree in the face of legislative opposition to his policy proposals prompted Congress to pass legislation clarifying and constraining the scope of executive decree authority (Schmidt, 1998). Before the bill could be passed over an expected presidential veto, Fujimori called out the tanks and closed Congress (Maxwell Cameron, 1997; Philip Mauceri, 1997).

In sum, where constitutions provide presidents with decree, the use of this authority to avoid negotiation with legislative opponents has been the subject of conflict between the branches, which in some cases has evolved into regime crisis.

7.4.7 Agenda Powers

Short of outright decree authority, presidents are frequently endowed with the authority to make policy proposals that have privileged procedural status before the legislature, either because

- They must be considered within a limited time period.
- There are limitations on the manner in which they can be amended.
- The proposal determines the set of policy alternatives among which the assembly must choose.
- Some combination of these.

In most presidential systems, presidents are endowed with the ability to introduce legislation and in many cases with the authority to require legislative action on proposals within a limited time. When executive agenda authority includes limitations on amendments or influence over what policy pertains even if the legislature does not support the executive proposal, then executives can wield enormous influence over policy outcomes. As documented by Negretto (2013), almost every Latin American country has some kind of urgency powers that vary in their scope and effects, such as floor arrest (legislative activity is paralyzed until this bill has been considered), automatic adoption (decree becomes law unless overthrown), or indeterminacy (no penalties attached if priority is not honored) (Magar et al., 2021). Magar et al.'s piece, indeed, is a rich example of how the use of fast-track authorities

Table 7.4 Legislative powers in the hands of presidents, by region

Region	Presidents submit legislation, including some exclusive domains	Power shared with legislature	Presidents cannot submit legislation
Americas	Argentina, Brazil, Chile, Colombia, Uruguay	Bolivia, Costa Rica, Dominican Republic, Ecuador, Honduras, Mexico, Panama, Paraguay, Peru	United States
Postcommunist		Poland	
Africa	Ghana, Liberia, Malawi, Namibia	Kenya, Niger, Senegal, Seychelles, Sierra Leone, Gambia, Zambia	
Asia		Indonesia, Maldives, South Korea, Sri Lanka	Taiwan, Timor-Leste
Western Europe		Cyprus, France	Portugal

Source: V-Dem (2022)

gives executives an advantage for the consideration of their preferred policies, applied to the case of Chile. The edited volume compiled by Eduardo Aleman and George Tsebelis (2016) provides additional insights about existing variation in Argentina, Chile, Colombia, Mexico, and Peru.

Table 7.4 summarizes systematic information provided by V-Dem for 2022, where the powers of presidents to submit legislation are compared across countries and regions. Most cases in Latin America and Africa endow executives with some powers to introduce legislation. With the exception of the United States, the countries where those powers are the lowest tend to be hybrids and on the edge for consideration as semi-presidential or parliamentary republics.

One critical difference between agenda and decree authority is worth noting. Agenda power entails presidential control over the policy alternatives among which legislatures debate and select, whereas decree allows presidents to implement policies without legislative debate or assent. Democratic theory commonly holds that debate in itself is a valuable political good, even apart from its effects on policy choices (Lijphart, 1999; David Miller, 1993). The experience of many presidential systems with decree authority supports this intuition. In terms of leverage over policy outcomes, agenda powers may provide presidents influence on par with decree authority, but legislative procedure constrained by agenda powers still involves debate and negotiation. That procedural distinction appears to mitigate conflict between the branches, whereas policymaking by decree can contribute to intractable conflict between the branches.

7.4.8 Vetoes

The most common presidential power over legislation is the veto. The structure of presidential vetoes, if allowed, varies along two critical dimensions. The first is the requirement for the assembly to override a veto, which varies from

- Simple majority (e.g., Colombia)⁷
- Absolute majority of the assembly's membership
- Absolute majority in joint session in a bicameral system
- Qualified majorities such as 3/5 or 2/3

The second dimension is whether the veto may be partial (also known as the item veto) or must apply to an entire piece of legislation (package veto).⁸

The veto is generally considered a reactive authority—one that allows presidents to hold up legislative initiatives, requiring further deliberation and even the construction of supermajorities—rather than a proactive authority such as decree or agenda power that allows presidents to initiate policy changes. Although this characterization is accurate in the case of the package veto format, however, it is not for the partial veto. The package veto allows legislatures to offer logroll-type proposals to presidents, enticing presidents to accept some policies the legislature wants in exchange for securing some preferred by the president. Advocates of the partial veto contend it allows presidents to unpack such legislative logrolls, removing wasteful or inefficient programs supported by individual legislators or factions. This account, however, overlooks the strategic impact the partial veto has on bargaining over legislation between the branches. Presidents can unpack logrolls only if such compromise legislation is passed by the legislature, but the very existence of the partial veto discourages compromise by allowing presidents to alter policies approved by the legislature unilaterally before implementation.

The point of veto authority is the same as that made above regarding other legislative powers entrusted to presidents. When executives can alter policy and then proceed to implement those changes without an intervening step of legislative debate and assent, incentives for compromise between the branches are undermined. Thus, arguments on behalf of the partial veto made on the grounds of budgetary efficiency, for example, must be weighed against the extent to which the partial veto weakens the executive's ability to commit to compromise agreements with the legislature.⁹

Pioneering work by Eduardo Aleman and Thomas Schwartz (2006) shed light on veto procedures in Latin America, where case-based variation depicted dissimilar interbranch games. In this line, Valeria Palanza and Gisela Sin's (2013, 2014) work analyze in depth the challenges of veto authority in the Argentine multiparty context. Table 7.5 distinguishes different variations of veto/override dynamics in the non-parliamentary systems that are considered democratic by V-Dem. Most cases are in intermediate cells but more prone to making it difficult for a legislature to override a presidential decision.

⁷This is effectively a presidential request to reconsider the legislation, given that the same majority that passed legislation initially can override the veto.

⁸For the configurations of veto authorities across presidencies, see Shugart & Carey, 1992; Lucky, 1993; Carey et al., 1997.

⁹For an analogous discussion of how sequential budget procedures can discourage compromise by undermining the ability of the last mover to commit, see Persson et al., 1997.

Table 7.5 Presidential veto/legislative override procedures, by country

Type Region	No veto	Overriden by simple majority	Overriden by members' majority	Overriden by qualified majority	No override
Americas		Colombia	Brazil Honduras Paraguay Peru	Argentina Bolivia Chile Costa Rica Dominican Republic Mexico Panamá Uruguay United States	Ecuador
Postcommunist				Poland	
Africa		Seychelles	Kenya Liberia	Ghana Malawi Niger Senegal	Namibia Sierra Leone The Gambia Zambia
Asia	Taiwan	Indonesia Sri Lanka		South Korea Timor-Leste	
Western Europe	France		Portugal		Cyprus

7.5 Used Presidents and Government Crises: “Should I Stay or Should I Go Now?”

The wave of democratization around the world spanning the late 1970s through the early 2000s prompted debates over the relative merits of presidentialism versus parliamentarism and whether regime type can affect democratic stability. That subsequent research homed in on increasingly fine distinctions in institutional design and examined their effects on regime performance is, in part, a happy by-product of the fact that breakdowns of democracy, across all regime types, were rare in the last two decades of the twentieth century.¹⁰ Yet crises of individual governments persist, and in systems where presidents are chief executives, these present the question of presidential survival. Fixed presidential terms, and the absence of a safety valve procedure allowing for the constitutional removal of a feckless president, remain

¹⁰ Among the presidential systems of Latin America, this is in part due to the fact that Latin American militaries were unable, or unwilling, to intervene in politics for the long haul. The reasons vary from country to country, but one effect is that when presidents and legislatures find themselves at a standoff, neither necessarily holds the option of knocking on the barracks door to ask for assistance. Another change is at the international level. The rest of the Western hemisphere is less tolerant of nondemocratic neighbors than it was in the mid-twentieth century.

problematic. Indeed, fixed terms can generate tensions both when presidents face extraordinary opposition and when they enjoy strong support. This section reviews the constitutional constraints on presidential reelection and the outcomes of government crises in Latin America that have pitted presidents against intransigent legislative opposition.

7.5.1 Reelection

Parliamentary systems do not place restrictions on the reelection of chief executives. In this sense, there is no constraint other than sustained assembly support (and, indirectly, voter support) to the stability in office of a popular prime minister. Presidential and hybrid systems, by contrast, not only place greater obstacles to the early removal of the president; they almost uniformly place constitutional limitations on presidential reelection. This was one of the sources of instability identified by Linz—that countries needed to find a new leader every 4 or 5 years, inhibiting thereby the implementation of long-term policies. Four decades later, to what extent are those fears borne out?

Table 7.6 shows the status of constitutional restrictions on presidential reelection in those presidential or hybrid polities that are considered democratic by V-Dem's electoral index. In this sample, we find just six cases that place no restrictions on reelection, without any regional pattern. The African presidential systems allow for two consecutive terms with lifetime limits thereafter; the only exceptions are Niger and Gambia, which do not impose any restriction. The starkest variation lies in the group of presidential regimes where almost every combination of number of terms, timing, and length of the pause is implemented. Most of the presidential systems of the Americas employ some kind of restriction. Two of them, Mexico and Paraguay, maintain the classic commitment to no presidential reelection that used to be the norm in Latin America.¹¹ The most conspicuous changes in electoral institutions in Latin America during the 1990s were, in fact, constitutional amendments to reverse longstanding prohibitions on consecutive presidential reelection. Chile, Peru, and Uruguay don't restrict the number of terms served but forbid them to be consecutive. In recent decades, presidents Michelle Bachelet (2006–2010 and 2014–2018) and Sebastian Pinera (2010–2014, 2018–2022) in Chile, Alan Garcia (1985–1990 and 2006–2011) in Peru, and Tabare Vazquez (2005–2010 and 2015–2020) in Uruguay waited at least one period to rerun and win second mandates. Their colleagues from Costa Rica and Panama face a two-term embargo before they can qualify for another mandate. Facing this more stringent restriction, only the *tico* Oscar Arias (1986–1990 and 2006–2010) won repeated mandates, starting his second term

¹¹ Interestingly, Mexico recently modified its traditional rule of nationwide no consecutive reelection, making it possible for legislators and mayors (12 straight years in the Chamber and the Senate, varying municipal rules state by state), but keeping the lifetime ban after one period for the president and governors.

Table 7.6 Executive term limits in presidential and hybrid systems^a

Region	<i>One term, then...</i>			<i>Two terms, then...</i>		<i>No limits</i>
	No reelection	<i>Eligible after one interim term</i>	<i>Eligible after two interim terms</i>	<i>No reelection</i>	<i>Eligible after one interim term</i>	
Americas	Mexico Paraguay	Chile Perú Uruguay	Costa Rica Panama	Colombia Ecuador United States	Brazil Argentina Dominican Republic	Bolivia ^b Honduras
Postcommunist				Poland		
Africa				Ghana Liberia Malawi Maldives Namibia Kenya Senegal Seychelles Sierra Leone Sri Lanka Zambia		Niger Gambia
East Asia	South Korea			Taiwan Timor-Leste		Indonesia
Western Europe		Portugal		Cyprus France		

^aArmenia, Austria, Botswana, Bulgaria, Cape Verde, Croatia, Estonia, Finland, Georgia, Iceland, Ireland, Lithuania, Macedonia, Romania, Slovakia, and Ukraine all select presidents by direct popular election, but the constitutional authorities of those presidencies are sufficiently limited that they are characterized as parliamentary, not hybrid, by our coding system

^bConstitution says two terms then eligible after one interim term, but 2017 decision by Supreme Tribunal of Justice held that any restriction on reelection would violate the American Convention on Human Rights

16 years after the end of his first. The American two-term, never again model is followed in Colombia and Ecuador a rule that ended the presidencies of such towering political figures as Alvaro Uribe (2002–2010) in Colombia and Rafael Correa (2007–2017) in Ecuador, each of whom tried, unsuccessfully, to cultivate loyal successors through whom they could continue to wield influence beyond the expiration of their presidencies.¹² Argentina, Brazil, and Dominican Republic do not forbid lifetime presidential careers but force presidents to wait one period after two in a row to run again. There is no clear pattern of success among two-term presidents who have attempted comebacks after their constitutionally mandated interregna.

¹²Ecuador voted to reinstate term limits in 2015, modifying Correa's former amendment that got rid of bans to reelection.

Argentine President Carlos Menem (1989–1999) failed in his effort to recapture the presidency in 2003 and aborted an announced campaign in 2011. Another two-term Argentine president Cristina Fernandez de Kirchner (2007–2015) was elected vice president in 2019. The most politically successful such former president to date is Brazil’s Luiz Ignacio da Silva (Lula) (2003–2011) who was reelected in 2022 and as of this writing is serving his third term (2023–[2027]).

Historically, strict no-reelection provisions in Latin America represented reactions to the abuse of power by presidents seeking to ensure their perpetuation in office (Carey, 2003). Popular incumbent presidents’ intent on securing two or more terms in office advocated relaxation of these provisions from the 1990s on, based on the grounds that the possibility of consecutive reelection removes constraints on the electoral choices allowed to voters, enhancing the quality of democracy. In fact, Gabriel Negretto (2022) shows that 11 out of 125 presidents between 1978 and 2015 were able to change the rules and overcome restrictions to their reelection rights. Charlotte Heyl and Mariana Llanos (2022) count 93 successful presidential term limit reforms in Latin America and sub-Saharan Africa. While some of these reforms triggered protests and struggles and were the seeds of governmental crises, other attempts to change rules failed, pursuant to the popular mobilizations and interbranch struggles they triggered. While it is not clear that these attempts have systematically threatened regime stability, it is likely that changes in such central rules could have been a source of crisis.

7.5.2 Government Crises and the Parliamentarization of Presidential Systems

Attempts by incumbents to extend their control over the executive are not the only sources of government crises in presidential systems. In Latin America, there have been multiple cases over three decades of government crises in which one or the other branch has been supplanted before its constitutional term expired. While some explanations highlight Latin America’s reputation for *caudillismo* (rule by political strongmen) and presidential dominance as sources of instability, institutional explanations would argue that the problem lies with presidentialism. With just a couple of exceptions, Latin American presidential systems do not have in place a constitutional provision analogous to a no-confidence procedure for the president, by which the chief executive can be removed simply for having lost the political confidence of her original selectors.¹³

¹³The Venezuelan Constitution of 1999 establishes a provision for referenda to recall any elected officials, including the president, upon the collection of signatures of 10% of registered voters (Arts. 72, 74). The threshold required for recall is the number of votes won in the original popular vote, thus rendering presidents elected with broad support relatively secure, but leaving a president elected by a narrow plurality—perhaps in a divided field of candidates—vulnerable to recall. As mentioned above, Ecuador enacted the *muerte cruzada* mechanism, which operates as a mutual dissolution of powers by presidential decision.

In this sense, recalling Linz's warnings, crises of government can end up threatening presidential regimes, potentially prompting military coups or attempts of self-coup (*autogolpe*). Impeachments, it was argued, were hard to attain due to the large majorities required and the limited reasons to start the process. When we contrast expectations with real-world events, however, we observe many legislative replacements justified on idiosyncratic constitutional grounds. The 1996 Ecuadorian Constitution, in place when President Abdala Bucaram was replaced in 1997, allowed Congress to remove the president for "physical or mental incompetence" (Art. 100) but did not spell out how such a condition would be established. President Bucaram disagreed with the legislative diagnosis but was unable to prevent his own removal. More recently, Brazilian President Dilma Rousseff was separated from her position in 2016 for alleged failure to comply with budget rules. Similarly, Paraguayan President Fernando Lugo was impeached in Paraguay for his alleged responsibility in the assassination of 17 peasants.

No country has experienced more legislative removals in the modern era than Peru. In 2000, Alberto Fujimori's abdication to Japan left the presidency vacant, but Congress's subsequent declaration of "moral incapacity" formalized his removal according to the constitution's Article 113. Although Fujimori's entire administration had been tainted by the breaking corruption scandal, the constitutional line of succession (Art. 115) formally ran through the first and second Vice Presidents before reaching the President of Congress, Valentin Paniagua. Nevertheless, Congress installed Paniagua as Fujimori's successor. In the years since, the Peruvian Congress has returned repeatedly to Article 113. In 2018, President Kuczynski resigned just before facing near-certain removal on allegations of corruption during an earlier period as Finance Minister. Kuczynski's vice president, Martin Vizcarra, was allowed to replace him but lasted just 2 years before he was removed by Congress in 2020, on the basis of alleged corruption during his prior term as a regional governor. Just 1 year later, in 2021, President Pedro Castillo sought to preempt a vote to end his short-lived presidency, again under Art. 113 and again on various corruption allegations. Castillo attempted to counter the threat by calling on the military to support a self-coup against Congress. The military arrested Castillo instead, however, and Congress formalized his removal with a vote of moral incapacity under Art. 113.

These cases would suggest that, even while Latin American constitutions remain presidential, the replacement of presidents in practice increasingly displays a more parliamentary flavor, giving priority to legislative discretion. Table 7.7 underscores the pattern. We report 25 cases of early termination of the executive or the legislative branch between 1992 and 2023 in the region.¹⁴ Congress prevails and removes (or forces the resignation of) the president in 20 of these cases. Presidents survive in only two—Fujimori's first self-coup in 1992 and Hugo Chavez's reconstitution of the Venezuelan political system in 1999–2000. Despite its institutional rigidities,

¹⁴Numbers differ as a function of definition, measurement, and criteria. As examples, Pérez Liñán and Polga-Hecimovich (2017) identify 4 coups and 14 executive removals between 1977 and 2014, while Hochstetler and Edwards (2009) recognize 11 failed presidencies between 1978 and 2005.

Table 7.7 Early terminations of presidents or legislatures in Latin America, 1992-2023

Year	Country	President	Winner	Outcome
1992	Peru	Alberto Fujimori	President	Successful <i>autogolpe</i> . New legislative elections, 1993
1992	Brazil	Fernando Collor De Mello	Congress	Impeachment, Vice President Itamar Franco completes the term
1993	Guatemala	Jorge Serrano	Congress	Congress selects an interim replacement for Serrano then calls new elections
1993	Venezuela	Carlos Andres Perez	Congress	Congress selects Ramon Jose Velasquez interim president for duration of Perez's term
1997	Ecuador	Abdala Bucaram	Congress	Destitution, Congress nominates its then-Speaker, Fabian Alarcon, as president for duration of Bucaram's term
1999	Paraguay	Raul Cubas Grau	Congress	Resignation. Vice President had been assassinated; Senate Leader Luis Gonzalez Macchi Grau completes the term
1999	Venezuela	Hugo Chavez	President	Interim legislative council appointed by president
2000	Ecuador	Jamil Mahuad	Congress	Vice President Jaime Noboa finishes the term
2000	Peru	Alberto Fujimori	Congress	Resignation. Congress selects a legislator, Valentin Paniagua, as interim president until elections are held, 6 months later
2001	Argentina	Fernando de la Rúa	Congress	Resignation. Congress selects San Luis state Governor Adolfo Rodriguez Saa as interim president
2001	Argentina	Adolfo Rodriguez Saa	Congress	Resignation. Congress selects former Senator and Buenos Aires state Governor Eduardo Duhalde as interim president
2003	Bolivia	Gonzalo Sanchez de Lozada	Congress	Resignation. Vice President Carlos Mesa empowered to complete the term
2003	Argentina	Eduardo Duhalde	Neither	Resignation. Commits to call to early elections
2004	Haiti	Jean Bertrand Aristide	Neither	Coup. Supreme Court president Boniface Alexandre completed the term
2005	Ecuador	Lucio Gutierrez	Congress	Remotion. Vice President Alfredo Palacio completes the term
2005	Bolivia	Carlos Mesa	Congress	Resignation. President of the Supreme Court Eduardo Rodriguez empowered to call to elections
2009	Honduras	Manuel Zelaya	Congress	Resignation/removal. Congress nominates Congressional leader Roberto Micheletti to complete the term
2012	Paraguay	Fernando Lugo	Congress	Impeachment. Vice President Federico Franco completes the term

(continued)

Table 7.7 (continued)

Year	Country	President	Winner	Outcome
2015	Guatemala	Otto Perez Molina	Congress	Resignation. Vice President Alejandro Maldonado completes the term
2016	Brazil	Dilma Rousseff	Congress	Impeachment. Vice President Michel Temer completes the term
2018	Peru	Pedro Pablo Kuczynski	Congress	Resignation. Vice President Martin Vizcarra empowered to complete the term
2019	Bolivia	Evo Morales	Congress	Resignation. Congress nominates Senator Jeanine Anez to complete the mandate
2020	Peru	Martin Vizcarra	Congress	Impeachment. Congress nominates Deputy Manuel Merino, who resigns. Term completed by Deputy Francisco Sagasti
2021	Peru	Pedro Castillo	Congress	Impeachment. Vice President Dina Boluarte succeeds Castillo
2023	Ecuador	Guillermo Lasso	Neither	“Muerte cruzada”: dissolution of both bodies and call to new elections

Source: Carey and Shugart (1998) and further updates by the authors

presidential democracies seem to have found ways to solve interbranch conflicts, most often without prompting a complete destruction of the existing constitutional order. In this sense, aside from a violation of legal standard, the early termination of an executive might approximate the safety valve favored by Linz. In the authoritative work on this subject, Aníbal Pérez Liñán (2007) highlights rising popular protests and legislative support falling below $\frac{1}{3}$ of Congress as the factors that make incumbent presidents vulnerable to legislative removal.

7.6 Conclusion

Parliamentary government fuses the selection of the executive to the popular vote for the assembly and the survival of the executive to assembly confidence. Academic research on parliamentary systems has focused extensively on how the selection and maintenance of governments drawn from assembly coalitions affects which parties and interests are represented in the executive (Laver & Schofield, 1998; Laver & Shepsle, 1996; Lanny Martin & Randolph Stevenson, 2001), for what duration (Paul Warwick, 1994; Arthur Lupia & Kaare Strøm, 1995; Laver & Shepsle, 1996; Diermeier & Stevenson, 2000), and with what effects on the distribution of policymaking authority (Laver and Shepsle, 1996; Huber, 1998). Although there are important differences in institutional design among parliamentary regimes, with consequences for representation and policy (Huber, 1998; Huber & Charles Shipan, 2002; Bingham Powell, 2000; Lijphart, 1999), the differences with presidential and hybrid regimes, and the institutional variation among the latter, are substantially more dramatic. Moreover, the level of institutional instability and change in presidential and hybrid regimes has outstripped that in parliamentary systems in recent decades, particularly as the newly democratic, and redemocratizing, regimes

associated with democracy's Third Wave, beginning in the late twentieth century, have overwhelmingly adopted directly elected chief executives.

Presidential systems allow voters maximum discretion over the composition of the executive and legislative branches of government. Hybrid regime formats represent efforts to provide a direct vote on the chief executive while retaining some mechanisms to ensure that the politicians administering the day-to-day operation of government maintain assembly support.

In all systems with consequential presidencies, and thus with parallel elections that determine the shape of government at the national level, the specific rules of competition affect whether those branches will support complementary objectives. Most directly, concurrent elections for the two branches, and electoral rules that encourage the construction of broad first-round coalitions, maximize the prospects for unified government.

To the extent that the preferences of the branches differ, the rules of legislative procedure affect the manner in which presidents and assemblies resolve policy differences through negotiation and compromise or whether such conflict leads to government crisis. The specific legislative powers of presidents vary enormously, with some configurations encouraging negotiation with legislative opponents prior to the adoption of policy and others allowing executive action to implement policies—perhaps provisionally—even in lieu of prior legislative ratification.

Several trends in institutional design and performance characterize the past three decades. In presidential regimes, we recognize loosening restrictions on reelection and the rise of legislative replacement of presidents during government crises in Latin America. Both imply convergence with parliamentary systems in that they relax the fixed nature of presidential terms and provide safety valves that decompress political struggles without threatening democracy. In this sense, both phenomena are consistent with broader arguments, particularly prominent among Brazilian scholars, that presidential systems do not necessarily operate as differently from parliamentary regimes as pure institutional analyses often suggest (Cheibub, 2002; Argelina Figueiredo & Fernando Limongi, 2000; Octavio Amorim Neto, 2002a, b) and that focus on the incentives for accommodation even between branches of government whose origins and survival are, constitutionally, separated. In parallel, most new democracies in Eastern Europe chose the promised *best of the world's* advantages offered by hybrid systems: the anchor offered by a fixed-termed president along with a prime minister and a cabinet designated by the parliament. These new hybrid regimes tend to provide their presidencies with extensive constitutional powers over the formation of governments and of legislation. We also observe that, in line with Table 7.1, Poland is the only one of the whole sample of former Communist countries with a hybrid regime that remains above the democratic threshold. We are not in a position to draw a causal inference between regime characteristics and democratic performance based on the cross-national comparison, but it is noteworthy that 1990s-era optimism regarding hybrid regimes has not borne out (Sartori, 1994, Elgie, 2011).

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Aleman, E., & Schwartz, T. (2006). Presidential vetoes in Latin American constitutions. *Journal of Theoretical Politics*, 18(1), 98–120.
- Alemán, E., & Tsebelis, G. (Eds.). (2016). *Legislative institutions and lawmaking in Latin America*. Oxford University Press.
- Altman, D. (2010). *Direct democracy worldwide*. Cambridge University Press.
- Altman, D. (2020). Checking executive Personalism: Collegial governments and the level of democracy. *Swiss Political Science Review*, 26(3), 316–338.
- Ames, B. (1995). Electoral rules, constituency pressures, and pork barrel: Bases of voting in the Brazilian Congress. *The Journal of Politics*, 57(2), 324–343.
- Ames, B. (2002). *The deadlock of democracy in Brazil*. University of Michigan Press.
- Amorim Neto, O. (2002a). The puzzle of party discipline in Brazil. *Latin American Politics and Society*, 44(1), 127–144.
- Amorim Neto, O. (2002b). Presidential cabinets, electoral cycles, and coalition discipline in Brazil. In S. Morgenstern & B. Nacif (Eds.), *Legislative politics in Latin America* (pp. 48–78). Cambridge University Press.
- Arrow, K. J. (1951). *Social choice and individual values*. Yale University Press.
- Back, H., Debus, M., & Dumont, P. (2011). Who gets what in coalition governments? Predictors of portfolio allocation in parliamentary democracies. *European Journal of Political Research*, 50(4), 441–478.
- Bawn, K. (1999). Money and majorities in the Federal Republic of Germany: Evidence for a veto players model of government spending. *American Journal of Political Science*, 43(3), 707–736.
- Cameron, M. A. (1997). Political and economic origins of regime change in Peru: The Eighteenth Brumaire of Alberto Fujimori. In M. A. Cameron & P. Mauceri (Eds.), *The Peruvian Labyrinth: Polity, society, economy* (pp. 37–69). Penn State Press.
- Carey, J. M. (2003). The reelection debate in Latin America. *Latin American Politics and Society*, 45(1), 119–133.
- Carey, J. M., & Shugart, M. S. (1998). *Executive decree authority*. Cambridge University Press.
- Carey, J. M., Neto, O. A., & Shugart, M. S. (1997). Appendix: Outlines of constitutional powers in Latin America. In S. P. Mainwaring & M. S. Shugart (Eds.), *Presidentialism and political parties in Latin America* (pp. 440–460). Cambridge University Press.
- Carroll, R., & Cox, G. W. (2007). The logic of Gamson's Law: Pre-election coalitions and portfolio allocations. *American Journal of Political Science*, 51(2), 300–313.
- Chaisty, P., Cheeseman, N., & Power, T. J. (2018). *Coalitional presidentialism in comparative perspective: Minority presidents in multiparty systems*. Oxford University Press.
- Cheibub, J. A. (2002). Minority governments, deadlock situations, and the survival of presidential democracies. *Comparative Political Studies*, 35(3), 284–312.
- Cheibub, J. A., & Chernykh, S. (2008). Constitutions and democratic performance in semi-presidential democracies. *Japanese Journal of Political Science*, 9(3), 269–303.
- Cheibub, J. A., Martin, S., & Rasch, B. E. (2021). Investiture rules and formation of minority governments in European parliamentary democracies. *Party Politics*, 27(2), 351–362.
- Coppedge, M., et al. (2023). V-Dem [Country-Year/Country-Date] Dataset v13. Varieties of Democracy (V-Dem) Project. <https://doi.org/10.23696/vdemds23>
- Cox, G. W. (1997). *Making votes count: Strategic coordination in the world's electoral systems*. Cambridge University Press.
- Cox, G., & Morgenstern, S. (2002). Epilogue: Latin America's reactive assemblies and proactive presidents. In S. Morgenstern & B. Nacif (Eds.), *Legislative politics in Latin America*. Cambridge University Press.
- Di Palma, G. (1990). *To craft democracies: An essay on democratic transitions*. University of California Press.
- Diermeier, D., & Stevenson, R. T. (2000). Cabinet terminations and critical events. *American Political Science Review*, 94(3), 627–640.
- Duverger, M. (1954). *Political parties*. Wiley.

- Duverger, M. (1980). A new political system model: Semi-presidential government. *European Journal of Political Research*, 8(2), 165–187.
- Elgie, R. (2011). *Semi-presidentialism: Sub-types and democratic performance*. Oxford University Press.
- Fernandes, J. M., & Magalhaes, P. C. (2016). Government survival in semi-presidential regimes. *European Journal of Political Research*, 55(1), 61–80.
- Field, B. N., & Martin, S. (2022). *Minority governments in comparative perspective*. Oxford University Press.
- Figueiredo, A. C., & Limongi, F. (2000). Constitutional change, legislative performance and institutional consolidation. *Revista Brasileira de Ciencias Sociais*, 29, 73–94.
- Figueiredo, A. C., & Limongi, F. (2001). Presidential power, legislative organization, and party behavior in Brazil. In *Political parties* (pp. 133–152). Routledge.
- Fiorina, M. P. (1996). *Divided government*. Allyn and Bacon.
- Friedrich, C. (1963). *Man and his government: An empirical theory of politics*. McGraw-Hill.
- Frye, T. (1997). A politics of institutional choice: Post-communist presidencies. *Comparative Political Studies*, 30(5), 523–552.
- Heyl, C., & Llanos, M. (Eds.). (2022). *Contested, violated but persistent: Presidential term limits in Latin America and sub-Saharan Africa*. Routledge.
- Hochstetler, K., & Edwards, M. E. (2009). Failed presidencies: Identifying and explaining a South American anomaly. *Journal of Politics in Latin America*, 1(2), 31–57.
- Huber, J. D. (1998). How does cabinet instability affect political performance? Portfolio volatility and health care cost containment in parliamentary democracies. *American Political Science Review*, 92(3), 577–591.
- Huber, J. D., & Martinez-Gallardo, C. (2008). Replacing cabinet ministers: Patterns of ministerial stability in parliamentary democracies. *American Political Science Review*, 102(2), 169–180.
- Huber, J. D., & Shipan, C. R. (2002). *Deliberate discretion?: The institutional foundations of bureaucratic autonomy*. Cambridge University Press.
- Ishiyama, J. T., & Kennedy, R. (2001). Superpresidentialism and political party development in Russia, Ukraine, Armenia and Kyrgyzstan. *EuropeAsia Studies*, 53(8), 1177–1191.
- Jones, M. P. (1995). *Electoral laws and the survival of presidential democracies*. University of Notre Dame Press.
- Kelley, R. L. (1973). Venezuelan constitutional forms and realities. In J. D. Martz & D. J. Myers (Eds.), *Venezuela: The democratic experience*. Praeger Publishers.
- Lamounier, B. (1993). Institutional structure and governability in the 1990s. In M. D.' A. Kinzo (Ed.), *Brazil: The challenges of the 1990s* (pp. 117–137). British Academic Press.
- Laver, M., & Schofield, N. (1998). *Multiparty government: The politics of coalition in Europe*. University of Michigan Press.
- Laver, M., & Shepsle, K. A. (1996). *Making and breaking governments: Cabinets and legislatures in parliamentary democracies*. Cambridge University Press.
- Levine, D. (1973). *Conflict and political change in Venezuela*. Princeton University Press.
- Lijphart, A. (1990). Presidencialismo y democracia de mayoría. In O. G. Arcaya (Ed.), *Hacia una democracia moderna: la opción parlamentaria* (pp. 109–128). Ediciones Universidad Católica de Chile.
- Lijphart, A. (1994). *Electoral systems and party systems: A study of twenty-seven democracies, 1945–1990*. Oxford Academic.
- Lijphart, A. (1999). *Patterns of democracy: Government forms and performance in thirty-six countries*. Yale University Press.
- Linz, J. J. (1994). Presidentialism or parliamentarism: Does it make a difference? In J. J. Linz & A. Valenzuela (Eds.), *The failure of presidential democracy* (pp. 3–90). Johns Hopkins University Press.
- Lucky, C. (1993). A comparative chart of presidential powers in Eastern Europe. *East European Constitutional Review*, 2(4), 81–94.
- Lupia, A., & Strom, K. (1995). Coalition termination and the strategic timing of parliamentary elections. *American Political Science Review*, 89(3), 648–665.

- Magar, E., Palanza, V., & Sin, G. (2021). Presidents on the fast track: Fighting floor amendments with restrictive rules. *The Journal of Politics*, 83(2), 633–646.
- Mainwaring, S. (1993). Presidentialism, multipartism, and democracy: The difficult combination. *Comparative Political Studies*, 26(2), 198–228.
- Mainwaring, S., & Shugart, M. S. (1997a). Juan Linz, presidentialism, and democracy: A critical appraisal. *Comparative Politics*, 19(4), 449–471.
- Mainwaring, S., & Shugart, M. S. (1997b). *Presidentialism and democracy in Latin America*. Cambridge University Press.
- Martin, L. W., & Stevenson, R. T. (2001). Government formation in parliamentary democracies. *American Journal of Political Science*, 45(1), 33–50.
- Martin, L. W., & Vanberg, G. (2003). Wasting time? The impact of ideology and size on delay in coalition formation. *British Journal of Political Science*, 33(2), 323–332.
- Martin, L. W., & Vanberg, G. (2004). Policing the bargain: Coalition government and parliamentary scrutiny. *American Journal of Political Science*, 48(1), 13–27.
- Mauceri, P. (1997). The transition to ‘Democracy’ and the failures of institution building. In M. A. Cameron & P. Mauceri (Eds.), *The Peruvian Labyrinth: Polity, society, economy* (pp. 13–36). Penn State Press.
- McKelvey, R. (1976). Intransitivities in multidimensional voting models and some implications for agenda control. *Journal of Economic Theory*, 12(3), 472–482.
- Mebane, W. R., & Sekhon, J. S. (2002). Coordination and policy moderation at midterm. *American Political Science Review*, 96(1), 141–157.
- Miller, D. (1993). Deliberative democracy and social choice. In D. Held (Ed.), *Prospects for democracy*. Stanford University Press.
- Moe, T. M., & Caldwell, M. (1994). The institutional foundations of democratic government: A comparison of presidential and parliamentary systems. *Journal of Institutional and Theoretical Economics*, 150(1), 171–195.
- Morgenstern, S., & Nacif, B. (2002). *Legislative politics in Latin America*. Cambridge University Press.
- Negretto, G. L. (2013). *Making constitutions: Presidents, parties, and institutional choice in Latin America*. Cambridge University Press.
- Negretto, G. L. (2022). Tinkering with executive term limits: Partisan imbalances and institutional legacies in Latin America. *Democratization*, 29(1), 38–56.
- Palanza, V., & Sin, G. (2013). Item vetoes and attempts to override them in multiparty legislatures. *Journal of Politics in Latin America*, 5(1), 37–66.
- Palanza, V., & Sin, G. (2014). Veto bargaining and the legislative process in multiparty presidential systems. *Comparative Political Studies*, 47(5), 766–792.
- Payne, M., et al. (Eds.). (2007). *Democracies in development: Politics and reform in Latin America*. IDB.
- Pérez-Liñán, A. (2007). *Presidential impeachment and the new political instability in Latin America*. Cambridge University Press.
- Pérez-Liñán, A., & Polga-Hecimovich, J. (2017). Explaining military coups and impeachments in Latin America. *Democratization*, 24(5), 839–858.
- Polga-Hecimovich, J. (2020). Reshaping the state: The unitary executive presidency of Rafael Correa. In S. Pachano & F. Sánchez (Eds.), *Assessing the left turn in Ecuador* (pp. 15–39). Palgrave Macmillan.
- Powell, G. B., Jr. (2000). *Elections as instruments of democracy: Majoritarian and proportional visions*. Yale University Press.
- Przeworski, A., & Limongi, F. (1997). Modernization: Theories and facts. *World Politics*, 49(2), 155–183.
- Riker, W. H. (1982). *Liberalism against populism*. Waveland Press.
- Sartori, G. (1994). *Comparative constitutional engineering: An inquiry into structures, incentives and outcomes*. New York University Press.
- Schleiter, P., & Morgan-Jones, E. (2009). Constitutional power and competing risks: Monarchs, presidents, prime ministers, and the termination of East and West European cabinets. *American Political Science Review*, 103(3), 496–512.

- Schmidt, G. (1998). Presidential usurpation or congressional preference?: The evolution of executive decree authority in Peru. In J. M. Carey & M. S. Shugart (Eds.), *Executive decree authority: Calling out the tanks or just filling out the forms?* (pp. 104–141). Cambridge University Press.
- Shugart, M. S. (1995). The electoral cycle and institutional sources of divided presidential government. *American Political Science Review*, 89(2), 327–343.
- Shugart, M. S., & Carey, J. M. (1992). *Presidents and assemblies: Constitutional design and electoral dynamics*. Cambridge University Press.
- Stepan, A., & Skach, C. (1993). Constitutional frameworks and democratic consolidation: Parliamentarianism versus presidentialism. *World Politics*, 46(1), 1–22.
- Strøm, K., & Swindle, S. M. (2002). Strategic parliamentary dissolution. *American Political Science Review*, 96(3), 575–591.
- Taagepera, R., & Shugart, M. S. (1989). *Seats and votes: The effects and determinants of electoral systems*. Yale University Press.
- Valenzuela, A. (1994). Party politics and the crisis of presidentialism in Chile: A proposal for a parliamentary form of government. In J. Linz & A. Valenzuela (Eds.), *The failure of presidential democracy* (pp. 165–224). The Johns Hopkins University Press.
- Volden, C., & Carrubba, C. J. (2004). The formation of oversized coalitions in parliamentary democracies. *American Journal of Political Science*, 48(3), 521–537.
- Warwick, P. (1994). *Government survival in parliamentary democracies*. Cambridge University Press.
- Weldon, J. (1997). Political sources of Presidentialismo in Mexico. In M. S. Shugart & S. P. Mainwaring (Eds.), *Presidentialism and democracy in Latin America* (pp. 225–258). Cambridge University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Performance and Stability of Federalism: An Institutional Perspective

8

Barry R. Weingast

Contents

8.1	Introduction	160
8.2	Economic Theories of Federalism	162
8.2.1	The Economic Classics of Federalism	162
8.2.2	Modern Economics of Federalism	164
8.2.3	Market-Preserving Federalism	165
8.3	Political Incentives	166
8.3.1	The Fiscal Interest Model	167
8.4	Endogenous Federalism	170
8.4.1	Further Insights	173
8.4.2	Conclusions	174
8.5	Federalism in the United States	175
8.5.1	The United States Under the Articles of Confederation, 1781–1788	175
8.5.2	Federalism Under the New Constitution	176
8.5.3	The Failure of American Federalism and Democracy in the Civil War	178
8.5.4	Conclusions	179
8.6	Conclusions	179
	References	181

Abstract

The purpose of this chapter is to survey the literature on federalism and draw on the new institutional economics and political science to suggest aspects of an integration of the literature and hence a more complete theory of federal performance. The literatures on the performance of federal systems divide around three central questions: (1) the economic performance of federalism; (2) its political performance; and (3) the sources of stability for federal arrangements. Although these three literatures are complementary, they remain incompletely integrated, precluding a more complete theory of the performance of federal systems, one

B. R. Weingast (✉)
Stanford University, Stanford, CA, USA
e-mail: weingast@stanford.edu

encompassing both political and economic aspects. This chapter analyzes how federalism affects the incentives and objectives of subnational government officials. It explores the conditions under which federalism is self-enforcing and applies the framework to a discussion of the evolution of federalism in the United States since inception.

Keywords

Federalism · Stability of federalism · Revenue sharing · Self-enforcing federalism · Market preserving federalism · Decentralization

8.1 Introduction

The literatures on the performance of federal systems divide around three central questions. The first concerns the economic performance of federalism; the second, the political performance; and the third, the sources of stability for federal arrangements.

Economists typically focus on the first question, with an emphasis on the normative aspects, such as How ought federalism to be designed? In answering this question, the economic theory of federalism provides a theory about the optimal organization of the state, addressing issues, such as what powers should be assigned to what levels of government; why should not the central government do everything? This literature has strong positive implications, for it also explains the economic implications of alternative divisions of powers among the various levels of government.

Economists and political scientists have also focused on various political aspects of federalism, such as the degree to which the federal institutions affect the incentives of public officials. A range of results show how performance of a federation depends on these institutional arrangements underpinning the federal system. For example, a series of recent models show how different forms of tax system give subnational governments different incentives to provide public goods and foster economic growth.

Finally, the most recent questions involve how a federal system is sustained. The first two literatures take the structure of federalism as given and study its implications. This third literature treats federal structure as endogenous. To understand this question, we follow Riker (1964) and pose the twin dilemmas of federalism (de Figueiredo & Weingast, 2005): Although federations differ on many dimensions, all face the two fundamental dilemmas:

Dilemma 1: What prevents the national government from destroying federalism by overawing its constituent units?

Dilemma 2: What prevents the constituent units from undermining federalism by free-riding and other forms of failure to cooperate?

To illustrate the first dilemma, for much of the late twentieth-century modern Mexico was federal in name only, as the central government centralized political authority from 1940 through the 1980s (though this trend has been modestly

reversed in the last decade). To illustrate the second dilemma, consider the United States under the Articles of Confederation where the national government was too weak to provide national public goods, such as defense, policing the common market, or a stable monetary regime. In each policy area, some states had followed their common pool incentives, resulting in a series of welfare losses and disputes among states.

Although these three literatures are complementary, they remain incompletely integrated, precluding a more complete theory of the performance of federal systems. The purpose of this paper is twofold: first, to survey these three sets of works, and second, to draw on aspects of the new institutional economics and political science to suggest aspects of an integration and hence a more complete theory of federal performance.

The economics literature tells us the economic implications of alternative assignment of policies across levels of government, while the political science literature discusses the issue of what divisions are self-enforcing. As Riker (1964) demonstrates, not all types of divisions of powers within federal systems are self-enforcing. Unfortunately, the literatures on these two topics are largely separate.

The economic classics of federalism, identified with Hayek, Tiebout, and Musgrave, emphasized the felicitous aspects of federalism generated through interjurisdictional competition and through appropriate matching of authority over public goods with different levels of government. Modern economics tends to qualify the positive conclusions of the earlier literature, analyzing incentives problems with lower governments, often emphasizing various common pool problems (race to the bottom, internal trade barriers, soft budget constraints) and intergenerational spillovers.

The underlying conceptual distinction between these two literatures is important: the classical economic works assumed benevolent governments at all levels, so certain types of incentive problems simply did not arise. The modern economics literature takes the first step in relaxing this assumption by looking at the incentive effects of lower governments, assuming that they ignore all external effects of their decisions. This reveals several types of common pool problems where decentralized decision-making yields problems.

Although the modern literature takes the first step, two limitations remain. First, scholars in this tradition tend to assume that local governments are benevolent governments with respect to their own citizens, ignoring how this comes about. Second, students in this tradition often retain the assumption that the federal government is benevolent and thus argue that, in the face of common pool problems, the federal government should be responsible for policymaking. Both problems limit the ability of this literature to develop a complete positive model of a federal system, treating each level of government in parallel, with its own interests and incentive problems.

A more complete theory of federalism drops the assumption of benevolent government at all levels, attempting to explain what do governments really do. An alternative assumption, associated with James Buchanan, is the opposite from the benevolent government—that government is Leviathan. In one sense, this position is too extreme—like that of the benevolent government, it fails to explain how the government's maximand arises. Yet this perspective contains an important insight:

different institutional arrangements affect the incentives of various levels of government and hence affect their performance (Brennan & Buchanan, 1980).

The important lesson of Buchanan for present purposes is to ask how different types of institutional arrangements affect the incentives of different levels of government. In particular, the second substantive section of this paper asks under what conditions will governments find it in their interests to foster markets? And here, federalism organized in the appropriate way has felicitous effects.

Finally, surveying the three separate literatures points toward a more complete theory of federalism, one encompassing both political and economic aspects. This also allows us to consider the structure of federalism—indeed, whether a country can sustain federalism at all—as endogenous.

This essay is organized as follows. Section 8.2 discusses both the economic classics and the modern economics literature on federalism. Section 8.3 discusses the literature on how federalism affects the incentives and objectives of subnational government officials. Section 8.4 turns more generally to the question of the conditions under which federalism is self-enforcing. Section 8.5 applies the framework to a discussion of the evolution of federalism in the United States since inception.

8.2 Economic Theories of Federalism

The economic theory divides into two types of analyses, which I call here the classical and the modern. The classic economic contributions to federalism rest on the work of Hayek, Tiebout, and Musgrave and are strongly normative in character based on the assumption of a benevolent government. The modern work extends this tradition to include the study of various incentive problems among the many subnational governments created by federalism itself.

8.2.1 The Economic Classics of Federalism

The classic economics of federalism revolves around Hayek's (1939) emphasis on the role of differential information across levels of government, Tiebout's (1956) emphasis on interjurisdictional competition, and Musgrave's (1959) notions of the assignment of policy and tax authority across jurisdictions. Oates (1972) codified this work into a relatively coherent approach. These are straightforward but powerful ideas. As these works are the most well known and as many good surveys exist, I will summarize them briefly.¹

Hayek emphasized the importance of differential information. Subnational governments and their citizens typically have better information than the national government about local conditions and preferences. Thus, for local public goods,

¹For surveys of federalism, see Inman and Rubinfeld (1997), McKinnon and Nechyba (1997), Oates (1972, 2001), and Rubinfeld (1987).

local decision-makers are likely to make more informed and thus better decisions about matching local policy to local conditions than will national policymakers. Moreover, national governments have a tendency to promulgate “one-size-fits-all” policies that are insufficiently variable to adapt to differing local conditions.

National governments typically provide money for various subnational government expenditures. Yet the categories often fail to reflect local priorities. For example, a recent crime bill provided funds for hi-tech squad cars. These cars are quite expensive, and most smaller jurisdictions would not purchase them if they have to pay for the cars from their own funds. The bill offered to pay for 95% of the cost of the cars, so, at 5% of the cost, a large number of jurisdictions will elect to put up their (small) share and buy to obtain the cars. Yet for most jurisdictions, this is not the highest-priority use of the subsidy. Some need newer electronic communications; others need more policeman and so on. But the bill did not offer the items most needed by the local governments; so instead the local governments chose what was available. This example could be repeated almost endlessly given the huge range of federal expenditure programs in the United States that foster local public goods provision.²

Tiebout (1956) emphasized the critical importance of interjurisdictional competition, which has at least three separate components.³ In the presence of labor and capital mobility, this competition leads to matching policies to citizens and communities, to citizens and capital owners sorting across jurisdictions to reside in those with policies most favorable to their needs and circumstances, and to provide incentives for city managers, who must anticipate the effects of their decisions for citizens and firm location decisions. Interjurisdictional competition combines with the mobility of citizens and capital to imply that policies at variance with the population or firms means that citizens and capital leave for more hospitable locations, lowering the tax base. City managers, ever worried about their city’s tax base, are thus led to provide policies hospitable to those located in their jurisdiction. This last point implies that incentives of governmental officials are endogenous to the structure of federalism, a topic discussed at greater length below.

Economists provide lots of evidence for these propositions. The United States has a very mobile population, and the evidence suggests that the provision of local public services is important in these moves (see, for example, Oates, 1969; Gramlich & Rubinfeld, 1982; Nechyba & Strauss, 1997).

The third idea derives from Musgrave (1959), who studied the assignment problem, which asks for a federal system how should authority over public goods, policy, and taxes be assigned to the different levels of government to maximize citizen welfare? Musgrave observed that public goods differed along several dimensions, including the degree to which they exhibit economies of scale and congestion as the number of people consuming the public good grows. Some public goods, like

²The same holds for other federal governments, such as Mexico where, until recently, expenditures were highly centralized.

³In addition to the surveys cited in note 1, *supra*, Scotchmer (1994) surveys the theoretical literature following Tiebout.

national defense, a common market, and a stable currency, are truly national in scope in the sense that there are large economies of scale and little congestion. Other public goods, such as parks, schools, and sanitation, are more local in scope.

The assignment principle holds that authority over public goods should be assigned to that level of government that can most efficiently produce it. Thus, the national government should provide public goods which exhibit little congestion, such as national defense, while local governments should provide public goods that are more local in scope, such as parks, schools, and sanitation facilities, particularly as jurisdictions can adjust their provision of these goods to suit citizen tastes.

As Oates (1972) emphasized, these ideas combine to provide a powerful normative argument in favor of federalism: federalism enhances citizen's welfare through the appropriate assignment of functions across levels of government, through matching and sorting individuals and firms across jurisdictions that provide policies and public goods best suited for their firms and citizenry, and through the appropriate use of information.

8.2.2 Modern Economics of Federalism

The classic economic approach to federalism is relatively prodecentralization. Modern economic research on federalism has tempered the felicitous conclusions of the classic economic contributions, in part by studying various incentive problems created by virtue that subnational governments face various forms of free-riding and common pool problems.

Externalities provide the paradigmatic example in which a locality fails to bear all the consequences of its decisions. For example, consider environmental regulation in the United States. In many geographic areas, particularly those with prevailing winds, it is possible for one jurisdiction to export its pollution to other jurisdictions. Thus, part of the costs of pollution may be borne by others. This externality also implies that some of the benefits of this jurisdiction's addressing this issue accrue to residents outside the jurisdiction.

For example, consider the problem of sulfur dioxide pollution from electric power plants that burn high sulfur coalitional in Ohio. The prevailing winds blow east, carrying much of this pollution 100s and 1000s of miles away from the source. Suppose Ohio were to act alone and impose pollution controls within its jurisdiction. This would raise its firm's costs without affecting these firms' competitors located outside the jurisdiction; and further, much of the benefits would accrue outside the jurisdiction. Reflecting a classic common pool problem, Ohio has incentives to ignore the problem. When all states act this way, the problem remains unaddressed. Under the assumption of benevolent national government, this perspective implies that pollution control should be assigned to the national instead of the state or local governments.

A second and more serious common pool problem concerns internal trade barriers among the federal units, historically in the United States and common in many modern federations, such as Russia. Internal trade barriers were a major problem

under the Articles of Confederation, part of the rationale cited by the Federalists for a stronger national government. Under the Constitution, the United States was assiduous in policing state trade barriers. Indeed, examples abound in the nineteenth century of states trying to erect such barriers.

The importance of the common market for the Tiebout mechanism and, more generally, for the process of national specialization and exchange to gain the economies of scale associated with a large country requires that states be prevented from raising internal trade barriers. This is a task that must be assigned the national government.⁴

Another critical mechanism studied extensively in recent years concerns whether subnational governments face a hard or soft budget constraint (Dillinger & Webb, 1999, McKinnon, 1997, Qian & Roland, 1998, Rodden, 2000, Rodden et al., 2001, Sanguinetti, 1994, Wildasin, 1997). A hard budget constraint (HBC) requires that subnational governments bear the full financial consequences of their decisions: they cannot be bailed out or receive endless forgivable loans from the central banking system. A subnational government facing a hard budget constraint cannot spend beyond its means without risking bankruptcy. A soft budget constraint allows a subnational government to spend beyond its means. Although a soft budget constraint creates budget deficits, these are financed either through central government bailouts or by access to forgivable loans from a central bank.

Because subnational governments do not bear the full financial consequences of their fiscal decisions under a soft budget constraint, their spending is more profligate. In extreme cases, where subnational governments expect bailouts, spending is rampant on rents, special interest benefits, and corruption. This process is sadly illustrated by the explosive state deficits and consequent national macroeconomic imbalances in late 1990s Brazil and late 1980s Argentina.

The various forms of incentive problems facing subnational governments yield two conclusions. First, these problems qualify the normative argument favoring federalism, as common pool problem hinders a federation's performances. Second, it provides a role for the national government, one of policing these problems. Of course, studies of the national government in this role often find it lacking, as Revesz's (1997) analysis of environmental policy in the United States illustrates.

8.2.3 Market-Preserving Federalism

Market-preserving federalism provides a framework for assessing different forms of federal systems.⁵ The framework makes explicit a set of axioms implicit in the type of federalism studied by the classical era economists. Specifically, the classic

⁴Notice that, here too, the assumption of benevolent government is critical to this normative judgment. The reason is that something must prevent the federal government from abusing its power over the subnational governments to force these governments to adhere to particular policies, such as rent-seeking.

⁵Weingast (1995), Montinola et al. (1995), and Parikh and Weingast (1997).

economic approach requires the following assumptions about the political structure of federalism:

- (A1) A hierarchy of governments, each with their own sphere of policy autonomy
- (A2) That subnational governments have substantial regulatory controls over their economies
- (A3) That a common market exists, including the federal government's ability to prevent subnational governments from raising internal trade barriers
- (A4) That subnational governments face a HBC
- (A5) That a set of institutions protect the federal arrangement from encroachment by political officials at the various levels of government

Each of these axioms is implicit in the classical economic studies noted above. Moreover, making these assumptions explicit emphasizes that the economic approach also contains a comparative theory of federalism. Consider: Axiom A1 is a defining condition of federalism which all federations satisfy. Federal systems delegate very different types of powers to their subnational governments. Axiom A2 emphasizes that, for federalism to have the effects noted by Musgrave, Oates, and Tiebout, subnational governments must have meaningful policy authority over local public goods and other regulatory policy that affects their local economies. The common market (A3) is also essential for interjurisdictional competition. When subnational governments can raise internal trade barriers, they can protect their firms from outside competition, extract rents from constituents and asset holders, and indulge in a higher degree of corruption. These barriers can also hinder the factor mobility necessary for matching and sorting of factors and local public goods emphasized in the Tiebout literature. Problems with the failure of the HBC (A4) are also well known, as discussed above. When subnational governments do not bear the full financial consequences of their decisions, they are more likely to spend profligately, disperse rents, and engage in corruption. Finally, the failure of A5 compromises federalism. Thus, when the constitution allows the federal government to take over states (India) or fire state governors (Mexico), the federal government can use these tools to control subnational governments.

8.3 Political Incentives

The classic economic literature makes a heroic assumption—that governmental actors are nonstrategic, benevolent maximizers of citizen welfare. This assumption rules out corruption, rent-seeking, service to interest groups, and generally manipulation of public policy to serve private instead of public ends.

As the *Federalist* observed more than 200 years ago, “if men were angels,” we would not have to worry about structuring incentives through the appropriate design of political institutions; but as they observed, because men are not angels, we must

be concerned with creating political institutions that provide government officials with the appropriate incentives.⁶

No theory of federal design is complete without attention to the incentives of public officials. As the Federalists noted, to the extent that governments in some countries are concerned with citizen welfare, it is because their public officials face incentives to do so. The sad fact is that many governments in the twentieth century have been among the worst perpetrators of violence and destruction against their citizens.

In this section, we rely on positive political theory to discuss some insights into the types of mechanisms that help align the interests of governmental officials with those of their citizens. A look around the world demonstrates that most governments on the planet fail on this dimension, so the problem ignored by economists is a nontrivial one.

Once we relax the assumption of benevolent governments and understand that government officials generally face political pressure to use their policy authority to provide rents, private goods, and confiscate wealth to transfer to supporters, we see that federal governments have advantages over unitary states (Persson & Tabellini, 2001). When the central government holds a monopoly on violence, its officials often abuse its policymaking authority to serve private ends.

By decentralizing a range of policies to lower governments, federalism affects the ability and incentives of governments to serve private instead of public ends (as Hayek (1939) and Brennan and Buchanan (1980) emphasize). First, decentralization limits the powers of the federal government in a range of areas. Second, decentralization places lower governments in competition with one another (Tiebout, 1956), so that those that abuse their citizens' rights are likely to lose mobile factors. Third, appropriately structured federalism forces lower governments to bear the financial consequences of their decisions. In this section, we discuss the consequences of this structure, investigating the finer-grained details of institutions, incentives, and performance. The next investigates how this structure can be maintained given the political pressure to undo it for political ends.

8.3.1 The Fiscal Interest Model

Wallis, Sylla, and Legler (1994) provide an important perspective on these problems. They observe that whatever a government's goals, it must obtain revenue to further them. Governments therefore prefer *ceteris paribus*, institutions, and policies that relax their budget constraints. This observation implies that the microinstitutional

⁶In *Federalist* No 51: "If men were angels, no government would be necessary. If angels were to govern men, neither external nor internal controls on government would be necessary. In framing a government which is to be administered by men over men, the great difficulty lies in this: you must first enable the government to control the governed; and in the next place oblige it to control itself" [337].

details about the way in which governments tax their citizens affect the government's incentives and hence its performance.

This approach has profound implications for how governments behave. Thus, consider how governments tax banks. Governments that raise revenue through taxing bank charters—effectively taxing numbers of banks—have incentives to maintain a small numbers of banks. By creating monopoly rents through restricting the supply of banks, businesses are willing to pay more for bank charters and hence increase government tax revenue. In contrast, governments that tax bank capital have the opposite incentives—they want to maximize the number of banks and to foster a healthy climate for banks so that the tax base is the largest.

Interjurisdictional competition combines with the fiscal interest in interesting ways. Consider two illustrations. In the early United States, Philadelphia was a major banking center.⁷ As with most states, Pennsylvania taxed bank charters, having incentives to restrict the number of banks. Massachusetts, another home to banking, taxed bank capital and found the revenue from this sufficiently high that it could end the property tax, the principal source of revenue for most states. Sometime in the 1820s, however, New York switched to taxing bank capital and, at the same time, greatly relaxed the requirements necessary to become a bank. The result was a fast-growing industry in New York City that took over from Philadelphia as the nation's banking center. In the wake of New York's success, several other states switched to taxing bank capital.

Second, the contrast with the early United States and Mexico is also instructive.⁸ Wallis et al. (1994) also show that, to attract capital and labor, every new state had to create a banking industry, both to finance local enterprises and to finance (through bills of exchange) shipments of goods across state lines. Thus, the number of banks in the United States grew enormously. By the early twentieth century, the United States had over 10,000 banks.

Turn of the twentieth-century Mexico, in contrast, was a federal system in name only. As with so many dictators, Mexico's used national policymaking to create monopolies as a means of raising revenue and securing power by creating constituents dependent on maintaining the state (see Haber et al., 2004). The result was that each state had its monopoly bank, and the national government had two. In contrast to United States's 10,000 banks, Mexico had less than 50.

This illustration emphasizes the importance of federalism for government incentives. As Haber (2003) shows, governments in both the early United States and in late nineteenth-century and twentieth-century Mexico had incentives to restrict the number of banks. But the United States' federal structure induced competition among the states, providing some states with an advantage if they removed these restrictions, later forcing others to do so in response.

Careaga and Weingast (2002), Inman (1988), and Inman and Rubinfeld (1997) study variants of the fiscal interest model. Traditional fiscal federalism of revenue sharing emphasizes redistribution (Musgrave, 1959; Oates, 1972). Although the

⁷This discussion draws on the work of Wallis, Sylla, and Legler (1994).

⁸I draw on Haber (2003) for the contrast of the US and Mexican banks.

economics literature on federalism has traditionally ignored the incentives created by revenue sharing, the fiscal interest model has important implications for the behavior of subnational government officials.

These positive models study two different effects of revenue sharing: the incentive of the national government to allocate these funds and the incentives these funds create for subnational governments. We consider these in turn.

Inman (1988) and Inman and Rubinfeld (1997) study the incentives of national elected officials to allocate national revenue sharing funds to lower jurisdictions to finance public goods. Studying allocation by the legislature, they show that legislators face a form of common pool problem: because expenditures in their own districts yield each the full benefits but taxes are spread over the entire nation, each legislator has an incentive to demand a higher than optimal level of public goods expenditures.⁹ When each representative behaves in this way, the total level of such expenditures is too high. Inman (1988) and Poterba and von Hagen (1999) provide evidence for this type of common pool problem in a variety of fiscal contexts.

Revenue sharing also affects subnational government incentives. To see this, suppose that subnational governmental actors are concerned about maintaining their political careers and that may use public policies in two broad categories to further their goals. First, they can provide market enhancing public goods, P ; and second, they can provide private rents (corruption, service to interest groups), R , that have dead-weight social losses. What determines how they make this tradeoff? Careaga and Weingast show that the type of tax system affects how they make this tradeoff.

Providing rents has a direct benefit—providing individuals and groups with benefits increases their willingness to support governmental actors. Providing market-enhancing public goods has two benefits. First, in parallel with providing rents, providing public goods benefits some individuals and groups and thus increases their willingness to support governmental actors. Second, it also relaxes the government's budget constraint: by increasing the market, these public goods expand the government's tax base and thus allows it to provide more of both types of services.

Next, consider the proportion of revenue of locally generated revenue, α , retained by the subnational government, with the federal government capturing the remainder, $1-\alpha$. Careaga and Weingast's main comparative static result is that, as the proportion of locally generated revenue, α , that the subnational government captures increases, the amount of market-enhancing public goods they provide also increases. The reason is that higher α raises the increase in revenue captured by the subnational government from increasing the tax base through an increase in P .

This approach has considerable implications for the type of revenue sharing system used by a federal country. Under complete revenue sharing, where the central government collects all revenue and then distributes this to states based on its own political criterion, the amount of revenue received by a given subnational

⁹Weingast, Shepsle, and Johnsen (1981) provide a model of this process.

government is only minimally responsive to P . In contrast, in a system where the subnational government raises the lion's share of its own revenue, increases in P enhance the local market, raise the tax base, and increase the tax revenue that the local subnational government captures.

Put simply, the type of revenue sharing system affects the incentives of subnational governments to foster markets. Governments that raise their own revenue are more likely to focus on market-enhancing public goods, while governments that obtain most of their revenue from the central government are more likely to be corrupt.

Consistent with this result is the following pattern among federal systems. During the United States's rise from a small economy on the periphery of the developed world in the late eighteenth century to become the richest nation by early twentieth century, states depended almost exclusively on their own sources of revenue. So too do provinces in modern China (see Jin, Qian, and Weingast (2005), although this has changed somewhat in the late 1990s). Iaryczower, Saiegh, and Tommasi (2000) suggest that this characteristic also held in Argentina during its high growth phase in the latter part of the nineteenth century and early twentieth centuries. In contrast, this condition fails for modern Argentina, India, Mexico, and Russia.

Rodden (2001) provides systematic evidence for a strong effect of own taxes on size of government. Previously Stein (1998), among others, had shown that greater degrees of fiscal decentralization is associated with larger governments. Rodden studying a larger sample added a variable measuring the degree to which subnational governments fund their spending through their own taxes. He shows that federal systems in which subnational government raise their own taxes tend to have smaller governments.

Finally, Epple and Zelenitz (1981) and Epple and Romer (1991) study how the Tiebout process limits the ability of local jurisdiction to extract tax revenue from constituents. Epple and Zelenitz find that more jurisdictions limit this extraction but do not eliminate it. Epple and Romer show that local jurisdiction can extract rents from landowners because land is a fixed rather than a mobile factor.

In short, economists since Hayek and Buchanan have argued that federalism affects the incentives of state actors. Brennan and Buchanan (1980) studied this in the context of the assumption that governments sought to maximize tax revenue. More recent work applies Wallis et al.'s (1994) fiscal interest model, showing how federalism affects the policy and fiscal behavior of subnational governments with more diverse goals.

8.4 Endogenous Federalism

As the survey thus far indicates, most analyses of federalism take the structure of federalism as given, analyzing its effects. This yields an impressive and well-developed theory about the effects of federalism. Nonetheless, this approach neglects the question of what sustains federalism over the long run. This question is of interest for its own sake but also because it interacts with the above concerns about performance. To the extent that only certain types of federal structures can be sustained,

some normative prescriptions will focus on worlds that cannot exist. The literature on federal stability is sufficiently new that no such conclusions have yet to be derived; but it holds the promise of adding significantly to the theory of federal performance.

The question of federal stability relates to a more general problem. In the United States, we take constitutional stability—and with it, the stability of both American federalism and democracy—for granted. Yet looking around the world, it is clear that most countries have difficulty adhering to constitutional and political rules supporting critical national institutions, such as federalism, democracy, and the market. Moreover, we tend to ignore the significant problems with the stability of American democracy, federalism, and the constitution during the first 100 years of the American Republic, as the failure of these institutions in the Civil War amply demonstrates (Weingast, 1998). In short, a theory of the performance of federalism requires that we endogenize it; that is, we must understand the forces that produce federal stability and hence explain why some countries can sustain federalism while others not.

In this section, we study some of the principles emerging that help explain how federalism is sustained. In the language of modern rational choice theory, this requires that federalism be self-enforcing in the sense that relevant national and subnational political actors have incentives to adhere to the rules of federalism.

The two fundamental dilemmas of federalism noted in the introduction demonstrate the fragility of federalism: As Riker (1964) observed—and in keeping with the modern economics literature—subnational governments have a tendency to destroy federalism through common pool problems. Similarly, the national government has a tendency to agglomerate power and overawe the subnational governments. The two fundamental dilemmas are especially difficult to solve because institutional mechanisms designed to mitigate one tend to exacerbate the other. To police the subnational governments, the federal government must have the power to punish them. But this power also allows a nonbenevolent federal government to abuse the states. In other words, sustaining federalism requires a careful balance between the two dilemmas (de Figueiredo & Weingast, 2005). The center must have sufficient power to prevent the states from free-riding yet not be so powerful as to be able to overawe the states.

How is this balance accomplished? The literature provides only partial answers. To address this question, I draw on the emerging literature on self-enforcing constitutions. I first present three principles of constitutional stability, for the issue of federalism's stability is part of the larger question of how some constitutions are sustained.¹⁰ I then turn to some additional insights provided in the literature.

The first principle concerns the ability of both levels of government to deter each other if the other oversteps the bounds of federalism.¹¹ Thus, the center must have the ability and incentive to police common pool problems by the subnational

¹⁰Works in this new literature include Bednar (1996), de Figueiredo and Weingast (2005), Fearon (2000), Przeworski (1991), and Weingast (1997).

¹¹This discussion draws on recent game theoretic approaches to self-enforcing federalism, especially de Figueiredo and Weingast (2005), Bednar (1996), and Weingast (1997).

governments, while the latter must have the ability to police the center if it abuses its authority and threatens the subnational government's independent authority. The difficulty in this setup is a coordination problem: the subnational governments must somehow have the ability to act in concert against potential abuses by the federal government. In game theoretic terms, the latter requires that they have a common focal point on which to react (de Figueiredo & Weingast, 2005). Such focal solutions to the subnational government coordination problem typically arise through pacts and constitutions that make explicit the rules of the federal game that specify readily identifiable bounds on the center. When subnational governments have the ability to act in concert, they can police the center because the center knows it will be punished if it oversteps the agreed upon bounds. As Madison observed in *Federalist* 46, "But ambitious encroachments by the federal government on the authority of the state governments would not excite the oppositive of a single state, or a few states only. They would be general signals of alarm. Every government would espouse the common cause."

Consider two examples. As discussed at greater length in Sect. 8.5, the American Constitution helped create limits on the new national government through several mechanisms. First, it articulated specific limits on the national government, such as a series of rights accorded to citizens. Second, it specified a series of institutions that defined how national policy was produced. In particular, the separation of powers system limited the types of policies that the national government could pursue. Third, specifying both rights and restrictions on the national government and procedures by which policy could be made helped create a series of focal solutions so that, per *Federalist* 46, citizens could react in concert against potential abuses. This trigger strategy, in turn, provided the incentives for the national government to adhere to these limits.

As a second example, consider China's fiscal decentralization of the 1980s, granting provinces considerable policy freedom of their economies (notably, to create a transition from socialism). Provinces were also granted considerable fiscal authority, with a remarkably high degree of tax retention at the margin (Jin et al., 2005). After the violence in Tiananmen Square in 1989, reactionary forces in the central government sought to undermine many of the reforms by recentralizing. At a critical meeting where this was proposed, the governor of Guangdong argued forcefully against this, with most other governors supporting him. As a consequence, the center had to back down (Shirk, 1993; Montinola et al., 1995).

The second principle of sustaining federalism draws on a general point about constitutional stability. All successful constitutions limit the stakes of politics (Przeworski, 1991, Chap. 2).¹² The reason is that, in the absence of credible limits on the government, circumstances inevitably arise where individuals and groups feel their livelihood and their property threatened. In such circumstances, they are likely to favor extraconstitutional means of protecting themselves, such as supporting a coup. Chile in 1973 and the Second Republic in Spain during the 1930s both

¹² Weingast (1997) provides a model of this principle.

illustrate this principle. In both cases, property owners felt their property threatened by the democratically elected regime; in both cases, they supported a military coup, resulting in an immediate transfer of power in Chile and in a bloody Civil War in Spain (1936–1939) culminating in the military's success against the regime.

Per the second principle, a critical aspect of federalism is that by decentralizing authority over a range of important policy issues over which people disagree, federalism lowers the stakes of national politics. When a critical issue over which people disagree is decentralized, citizens avoid both the conflict at the national level of whose view will win and also the risk that the winners impose their will on the losers.

As an illustration, consider the issue of slavery in the early United States. Northerners and Southerners strongly disagreed about the appropriate form of rights in slaves. The huge value of slaves in the Southern economy was so large as to give Southerners little choice but to support strong rights in slaves. By decentralizing this issue, federalism helped the country survive for three generations. Had the national government sought to deal with this issue for the entire nation, disagreements over this issue imply that the country would likely have been dissolved much earlier than 1860.

The third principle, first proposed by Riker (1964), concerns the role of elections and the party system. The need to win elections is one of the principal incentives that sustains federalism. When local and national politicians do better by coordinating their activities, officials at each level have incentives to respect the power and prerogatives of the other levels. Consider a two-party system (though the same logic holds with more parties). Suppose that party officials at one level of government attempt to usurp the power of another level. Then officials at the other level are likely to resist, thus hindering the cooperating necessary to win elections. This lack of coordination, in turn, grants their opponents a competitive advantage in the elections. To avoid losing, therefore, officials at each level have an incentive to respect the federal allocation of authority.

In the United States, parties have long been federal organizations, with the bulk of supporters at the state level. To win national elections, members in many localities had to cooperate, allowing them the ability to protect their prerogatives. In contrast, Ordeshook and Svetsova (1997) show how Russia's electoral institutions imply lack of coordination among subnational and national officials and how this in part underpins Russia's noncooperative federalism. Garman, Haggard, and Willis (2001) show that this principle underlies much of the differences in federalism across Latin America.

8.4.1 Further Insights

Two recent approaches focus on the national government's policing of subnational governments. Bednar, Eskridge, and Ferejohn (2001) argue that national courts often play this role. Yet such courts by their nature cannot as easily police the national government. They show how this approach helps illuminate the history of federalism in the United States and Canada. Bednar et al. (2001) develops a formal model

arguing that the national government must develop a reputation for punishing subnational governments that pursue common pool incentives. In a game theoretic context, she provides conditions for an equilibrium in which the national government's policing efforts succeed. Both these approaches, however, study only the problem of subnational governments, ignoring the problem of policing the national government.

In addition, two approaches endogenize aspects of federal structure. Cremer and Palfrey (2002) are the first to endogenize the assignment problem by considering a polity divided into subnational jurisdictions. Voters in these jurisdictions are to choose a level of public goods. But voters may also decide to nationalize the public good, that is, to provide one level of the public good at the national level. They show that under some configurations of preferences, citizens will nationalize the local public good. Suppose there are three jurisdictions, each with three citizens with the following configuration: two jurisdictions have a majority of two who want a low level of public good and a minority of one who wants a high level. In the third jurisdiction, all citizens want a high level of public goods. Nationally, then, there is a majority of five of nine citizens who (weakly) favor providing a high level of the public good: the three citizens from the jurisdiction who want high public goods are weakly in favor, and the minority in each of the other jurisdictions strictly favor nationalization.

Wibbels (2003) provides an important insight into the problem of how federations maintain a hard budget constraint for subnational governments. He observes that most studies of this problem are exhortations—they advise the federal government to impose hard budgets and not to give in to state demands for deficit relief. But exhortations are not self-implementing. Wibbels observes that if most states run up large debts simultaneously, the pressure on the national government to take over this debt is likely to be impossible to avoid.¹³ So where does the pressure not to bail out states arise? The answer is in part the states that have maintained sound finances: they must bear their costs of a bailout without any of the benefits and will thus oppose bailouts. This has two implications. First, the distribution of profligate states matter. Second, constitutional rules governing how national financial decisions affect the likelihood of bailouts. For example, if an upper chamber represents the states, the bias in representation (e.g., in favor of poor, small states) can affect the national bailout. Wibbels (2003) shows that this process was important for preventing the proposed bailout of bankrupt and near-bankrupt American states in 1841 and thus for maintaining hard budget constraints on the states.

8.4.2 Conclusions

To summarize, this section surveys a new, exciting, but incomplete approach to studying the stability of federalism. To survive, federalism must be self-enforcing.

¹³This suggests that maintaining a hard budget constraint is in part a coordination game, so that mechanisms that prevent subnational governments from coordinating are important.

Yet the two fundamental dilemmas of federalism make this problematic. The center needs sufficient power to police common pool problems among the states; and yet more power makes it more likely that the center will abuse its power. Three principles of self-enforcing federalism were noted. Although the works discussed here do not yet constitute a coherent theory, they do suggest the difficulties in maintaining a federal system and that the set of federal rules that can be sustained in practice may be far smaller than the set of potential rules discussed in the normative literature.

8.5 Federalism in the United States

The history of American federalism illustrates the various principles studied above. Indeed, most of the major turning points in American history reflect changes in federalism, including the formation of the Constitution, the conflict over slavery and the Civil War, and the changing relationship of the national government to the economy during the New Deal and during the 1960s.

The theories summarized above provide considerable insight into the structure and performance of the American economy and political institutions in its first century.

8.5.1 The United States Under the Articles of Confederation, 1781–1788

Per the modern economics literature, the United States under the Articles of Confederation exhibited a range of common pool problems. Three such problems stand out as threatening the economic and political future of the new nation: the inability to defend itself; the lack of a stable, national currency, and hence the ability of states to export part of the costs of monetary inflation (and hence SBC); and the rise of internal trade barriers. The Articles represented a loose federation, with the center unable to police common pool problems. Indeed, these problems were central to the Federalists' arguments in favor of their proposed Constitution in 1787. The inability to defend itself was one of the Federalist's major arguments (and the substantive subject raised in the *Federalist Papers*—numbers 2–5 entitled, "Concerning Dangers from Foreign Force and Influence"). The Congress could not raise its own taxes to pay for national defense. Instead, the national government could pass a law authorizing defense expenditures. To finance this authorization, it had to request that each state provide its share of the funds, which required that each state pass laws raising new taxes. This created a classic common pool problem, as each state had an incentive to free-ride. Many of the states failed to make their contributions, greatly hindering the ability of the national government to defend the country. Similarly, some states raised internal trade barriers hindering the common market. Other states inflated their own currencies rather than raising taxes, creating monetary problems. Moreover, per the assignment problem, it is clear that the Articles of Confederation

failed to provide adequate powers and authority to the national government to address these public goods problems.

Per the market-preserving federalism framework, the United States under the Articles of Confederation failed the common market condition and the hard budget constraint condition. Internal trade barriers allowed states to create local monopolies, keep out products from neighboring states, and thus hindered the benefits noted by the classic economic treatment of Tiebout. This political framework had high opportunity costs, as it left considerable gains from cooperation uncaptured.

8.5.2 Federalism Under the New Constitution

The debates prior to the Constitution corresponded to the problems identified by the two fundamental dilemmas of federalism. The Federalists emphasized the second dilemma's common pool problems and the dangers for the future of the new republic. To address the national government's inability to provide essential public goods, the Federalists argued to increase the national government's powers. On three occasions prior to the proposed new Constitution, they failed. The reason is that the Antifederalists had the upper hand. In combating the Federalists, the Antifederalists emphasized the first federal dilemma, namely, the dangers that a national government with too much power would overawe the states and destroy the liberty so recently won from Great Britain. To resolve this problem in 1787, the Federalists proposed to grant the national government more powers while, at the same time, embedding that government in a framework that limited its powers and lowered its potential to abuse those powers.

The new Constitution dramatically altered the federal arrangement. Consistent with the classic economic analysis, the Constitution created a federal structure that helped promote economic development and put the United States onto the path of becoming the richest nation in the world. First, it created institutions and powers allowing the national government to provide the three national public goods lacking under the Articles: defense, the common market, and a stable currency. Per modern economic's focus on common pool problems, the Constitution also granted the national government the power to police the common market and allowed the federal government to police the propensity of states to create internal trade barriers and to export their bad monetary behavior. Although states proved quite imaginative in the ways that they invented new means of internal trade barriers, the Supreme Court was equally imaginative at creating new Constitutional doctrine to rule these unconstitutional.¹⁴ The result was one of the largest unregulated common markets in the world, combined with a strong system of private property rights and the enforcement of contracts. Federalism thus helped to

¹⁴ For example, Supreme Court *Bank of Augusta v. Earle* (1839) used the privileges and immunities clause to prevent the state of Alabama from taxing out of state banks at a higher rate than in state banks. In *Gibbons v. Ogden* (1824), the Court used the commerce clause to prevent the state of New York from granting a monopoly to operate steamships in that state.

underpin the thriving national market, the engine of American economic growth that brought the United States from the periphery to the richest country in the world.

Second, the new Constitution's division of policy authority between national and state governments corresponded to that following Musgrave's assignment logic: the national government retained policy authority over national public goods, such as those noted above, and foreign policy, while the states retained power over nearly all local issues, including the types of property rights system, social regulation, and religious freedom. Per Tiebout, states faced strong competition with one another. This was especially true for new states and territories on the frontier which needed to compete for scarce capital and labor. After 1825, most citizens who moved to the frontier did so to gain land and become participants in a national and international economy; indeed, many grew grain shipped east, often bound for Europe. The evidence suggests that these territory and state governments focused attention on providing market-enhancing public goods.¹⁵

As North (1961) first argued, this system result in a remarkable degree of regional specialization and exchange that emerged after 1815.¹⁶ By the Civil War, the South produced primarily export crops, such as cotton, but also tobacco, indigo, sugar, and rice. The Northeast produced transportation, financial, and marketing services for the South's export economy. By the end of this period, the Northeast was also home of the emerging manufacturing economy, exporting its products West and South. The Northwest, largely self-sufficient at the beginning of the nineteenth century, slowly became specialized producers of food for the rest of the nation, particularly the Northeast.

The theory of self-enforcing federalism also helps explain the stability of the new Constitution. Whereas the Antifederalists had defeated the Federalists' earlier attempts to increase the national government's powers, the Federalists succeeded with the Constitution. The Federalists' failed attempts focused exclusively on addressing the second dilemma. In contrast, the Federalists attended to both dilemmas of federalism, not just the common pool problems, when they devised the new Constitution.

The genius of the new Constitution was to create a new set of institutions that at once provided critically needed national public goods while limiting the national government's powers to these domains. Per the theory of self-enforcing federalism, subnational governments must have a means of coordinating against the national government when it attempts to overstep its bounds. To address this, the Federalists attempted not only to articulate new powers for the national government but also a series of limits on its powers. First, the national government was to be one of enumerated powers only. Federalism was explicitly recognized, with many of the most important policy domains explicitly reserved for the states. At the same time, the Bill of Rights charged the national government to respect a range of citizen rights.

¹⁵ See Heckleman and Wallis (1997) on railroads; Haber (2003) and Wallis, Sylla, and Legler (1994) on banks; and Wallis (2001) on canal building in Indiana.

¹⁶ See also Callender (1901).

Second, the national government could act only under a series of well-specified procedures governed by the separation of power system. This not only ensured supramajority support before the national government could act, but, as the Federalist articulated, it would pit “ambition against ambition,” making rash action less likely.

Third, these provisions were to be enforced in part through trigger strategies that made adhering to these bounds self-enforcing for national political officials. As quoted above in Sect. 8.4, Federalist 46, states and citizens were to react in concert in the event of an attempt by the national government to overstep the Constitution’s bounds. Expectations of this reaction, in turn, provided the critical incentives for national political officials to honor the Constitution’s prescriptions. Of course, the fact that the states were to maintain militias while the federal government was likely to have a very limited domestic standing army meant that as a group the states had far more military resources than the national government.

8.5.3 The Failure of American Federalism and Democracy in the Civil War

American democracy and federalism failed in 1860 with the election of Republican Abraham Lincoln. Many Southerners feared that the national government would not adhere to the Constitution’s historic bargain. For three generations, rights in slaves had been the domain of the states, although the federal government had from the beginning legislated about slavery in the territories and hence the future states.¹⁷ Many Southerners felt the Republicans antislavery sentiments would lead them, once in power, to break the bargain. Of course, many other factors also underlay the Civil War, but the threat to slavery was one of the most important. The value of slaves in 1860 was phenomenally large—on the order of the entire American GNP (see Atack & Passell, 1994, Chap. 13). This implies that even modest threats against slavery would have large expected effects on Southerners.

A variety of institutions helped protect slavery within the American system, including the Constitution’s separation of powers system making national legislation difficult, and federalism that left issues of property to the states. The issue of whether these institutions protecting slavery were self-enforcing was of central concern to Southerners (Weingast, 1998). One of the principle institutions protecting slavery was the balance rule, the notion that the United States would be composed of an equal number of slave and free states, maintained (though with occasional difficulties) until 1850. The difficulty in the 1850s was that the North was growing far faster than the South and that slavery had no legal territories within which to expand. The long-run political stability of slavery was therefore in doubt.

¹⁷Part of what made this system self-enforcing was the balance rule, the notion that free states and slave states would be admitted in pairs so that the two sections would have equal representation in the Senate. Equal representation, in turn, gave each section a veto over national policy. With its veto, Southerners could protect the federal arrangement to leave slavery alone. See Weingast (1998) for details.

Although the reasons for southern secession were complex and hardly pre-ordained, the issue of maintaining self-enforcing federalism that protected rights in slaves within the American system was one of the principal reasons for the Union's breakup. The main point is that even a well-designed federal system like the United States had problems that caused massive failure.

After the Civil War, federalism reemerged as Southerners' regained power in the nation. In this period, too, federalism helped underpin long-term economic growth. The growing integration of the economy generated a range of new problems that could not be solved by the states, leading to greater national regulation, notably railroad regulation and antitrust in the late nineteenth century. This regulation foreshadowed the question of the relationship of the national government to the economy that would dominate much of the politics and economic policymaking of the twentieth century, but that is another story.

8.5.4 Conclusions

This brief survey of American history illustrates the importance of the various approaches to federalism surveyed above. All aspects were shown to be relevant. Moreover, the fact that the American federal system fell apart in 1860 combines with the fragility of federalism elsewhere (e.g., the large federations in Latin America) reinforces the importance of the question of what conditions provide for the stability of federalism. Nonetheless, American federalism survived the Civil War and helped foster American economic growth, creating the richest nation in the world.

8.6 Conclusions

This paper surveys several literatures on federalism and provides a partial integration. Three categories of works were discussed. First, it considers the economic literature on federalism as an optimal theory of the organization of the state. This includes the classic economic works on federalism of Hayek, Musgrave, Oates, and Tiebout, emphasizing the benefits of federalism through the efficient use of information, the appropriate assignment of policies to levels of government, and the important effects of interjurisdictional competition. More modern economic works have tended to qualify the classic economic literature's conclusions, typically emphasizing various types of common pool problems. The literature on market-preserving federalism was also discussed.

The second literature steps beyond the first by emphasizing the economic literature's assumption of benevolent government. Although this assumption is useful for some questions, it hinders the development of a positive theory of federalism. Thus, I discussed several works studying how federalism affects the incentives of sub-national government officials. The traditional economic framework, taking government goals as given, misses an important aspect of federalism emphasized by Buchanan, Hayek, and Tiebout, among others, namely, that federalism helps shape preferences of political officials in ways consistent with furthering citizens' welfare.

In particular, interjurisdictional competition in the presence of several conditions—an HBC, policy authority, and a common market—helps align subnational government officials' incentives with those of their citizens. A range of recent works also demonstrates the importance of fiscal system for subnational government officials' incentives.

The third literature discussed is the least developed but opens the possibility of greatly extending the breadth of the literature on federalism. It begins with the observation that both the first two literatures take federalism's structure as given. The third literature makes the structure of federalism endogenous and asks what makes federalism self-enforcing in the sense that political officials have incentives to abide by federalism's rules.

The common theme among all these works is the role of institutions in economic and political performance. By analyzing the three literatures and suggesting how they fit together, this essay points toward a more complete theory of federalism.

Let me conclude this essay by suggesting some implications for developing countries, where it is now widely recognized that government incentives toward rent-seeking and corruption rather than citizen welfare represents a major hindrance of development (e.g., Easterly, 2001). The literature on federalism has long emphasized the importance of federal institutions for aligning the incentives of political actors with citizen welfare. Competition among jurisdictions in combination with factor mobility places limits on the degree to which governments can extract taxes and rents from citizens. This process is not perfect, as the work of Epplé and Zelenitz (1981) and Epplé and Romer (1991) suggest.

Because security of property rights and protection against government confiscation or rapaciousness is a major problem in the developing world, the institutions of federalism hold significant promise of mitigating these problems. Of course, not just any type of federalism performs this role. As the market-preserving federalism framework suggests, federal systems differ significantly in how they devolve power and authority, and only some types of devolution are likely to improve political and economic performance in the developing context. In particular, appropriately structured federalism requires that subnational governments have real policy authority, face a common market and factor mobility, and face a hard budget constraint. Institutions supporting each of these seem necessary to obtain the beneficial effects of federalism.

Unfortunately, the political reality in most developing countries implies unstable regimes that typically use various forms of rents, subsidies, and market intervention to make political deals necessary to survive. These political forces work against implementing the type of federal system just outlined, in part because it forces political officials to break the deals used to maintain power.

Acknowledgments The author thanks Rui de Figueiredo, Stephen Haber, Robert Inman, Wallace Oates, Yingyi Qian, Antonio Rangel, John Wallis, and David Wildasin for helpful conversation.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Atack, J., & Passell P. (1994). *A new economic view of american history: From colonial times to 1940*. 2nd ed. W.W. Norton & Company.
- Bednar, J. L. (1996). *Federalism: Unstable by design* [MS, University of Michigan].
- Bednar, J. L., Eskridge, W., & Ferejohn, J. A. (2001). A political theory of federalism. In J. Ferejohn, J. Rakove, & J. Riley (Eds.), *Constitutions and constitutionalism*. Cambridge University Press.
- Brennan, G., & Buchanan, J. M. (1980). *The power to tax*. Cambridge University Press.
- Callender, G. S. (1901). *Selections from the economic history of the united states, 1765–1860*. Boston: Ginn & Company.
- Careaga, M., & Weingast, B. R. (2002). Fiscal federalism, good governance, and economic growth in Mexico. In D. Rodrik (Ed.), *In search of prosperity: Analytic narratives on economic growth*. Princeton University Press. Forthcoming.
- Cremer, J., & Palfrey, T. R. (2002). Federal mandates by popular demand. *Journal of Political Economy*, 108(5), 905–927.
- de Figueiredo, R. J. P., & Weingast, B. R. (2005). Self-enforcing federalism. *Journal of Law, Economics, and Organization*, 21, 103.
- Dillinger, W., & Webb, S. B. (1999). *Fiscal management in federal democracies: Argentina and Brazil* (Working paper). World Bank.
- Easterly, W. (2001). *The elusive quest for growth: Economists' adventures and misadventures in the tropics*. MIT Press.
- Epplé, D., & Romer, T. (1991). Mobility and redistribution. *Journal of Political Economy*, 99, 828–876.
- Epplé, D., & Zelenitz, A. (1981). The implications of competition among jurisdictions: Does Tiebout need politics? *Journal of Political Economy*, 89, 1197–1217.
- Fearon, J. D. (2000). *Why use elections to allocate power?* W.P., Department of Political Science, Stanford University.
- Garman, C., Haggard, S., & Willis, E. (2001). Fiscal decentralization: A political theory with Latin American cases. *World Politics*, 53, 205–236.
- Gramlich, E., & Rubinfeld, D. (1982). Micro estimates of public spending demand functions and tests of the Tiebout and median voter hypotheses. *Journal of Political Economy*, 90, 536–560.
- Haber, S. (2003). Banks, financial markets, and industrial development: Lessons from the economic histories of Brazil and Mexico. In J. A. Gonzalez, V. Corbo, A. O. Krueger, & A. Tornell (Eds.), *Macroeconomic reform in Latin America: The second stage*. University of Chicago Press. Forthcoming.
- Haber, S., Razo, A., & Mauer, N. (2004). *The politics of property rights: Political instability, credible commitments, and economic growth in Mexico, 1876–1929*. Cambridge University Press.
- Hayek, F. A. (1939). The economic conditions of interstate federalism. *New Commonwealth Quarterly*, V, 131–149. Reprinted in his *Individualism and economic order*. University of Chicago Press, 1948.
- Heckleman, J. C., & Wallis, J. J. (1997). Railroads and property taxes. *Explorations in Economic History*, 34, 77–99.
- Iaryczower, M., Saiegh, S., & Tommasi, M. (2000). *Coming together: The industrial organization of federalism*. Working Paper, Universidad de San Andres, Argentina.
- Inman, R. P. (1988). Federal assistance and local services in the United States: The evolution of a new federalist fiscal order. In H. S. Rosen (Ed.), *Fiscal federalism: Quantitative studies*. University of Chicago Press.
- Inman, R. P., & Rubinfeld, D. L. (1997). The political economy of federalism. In D. C. Mueller (Ed.), *Perspectives on public choice theory*. Cambridge University Press.
- Jin, H., Qian, Y., & Weingast, B. R. (2005). Regional decentralization and fiscal incentives: Federalism, Chinese style. *Journal of Public Economics*, 89, 1719.

- McKinnon, R. I. (1997). Market-preserving fiscal federalism in the American monetary union. In M. I. Blejer & T. Ter-Minassian (Eds.), *Macroeconomic dimensions of public finance* (pp. 73–93). Routledge.
- McKinnon, R., & Nechyba, T. (1997). Competition in federal systems: The role of political and financial constraints. In J. Ferejohn & B. R. Weingast (Eds.), *The new federalism: Can the states be trusted?* Hoover Institution Press.
- Montinola, G., Qian, Y., & Weingast, B. R. (1995). Federalism, Chinese style: The political basis for economic success in China. *World Politics*, 48, 50–81.
- Musgrave, R. (1959). *Public finance*. McGraw Hill.
- Nechyba, T., & Strauss, R. (1997). Community choice and local public services: A discrete choice approach. *Regional Science and Urban Economics*, 28(1), 51–73.
- North, D.C. (1961). *The economic growth of the united states 1790–1860*. Prentice-Hall.
- Oates, W. (1969). The effect of property taxes and local public spending on property values: An empirical study of tax capitalization and the Tiebout hypothesis. *Journal of Political Economy*, 77, 757–771.
- Oates, W. (1972). *Fiscal federalism*. Harcourt Brace Jovanovich.
- Oates, W. (2001). Fiscal federalism. *Journal of Economic Literature*, 37(3), 1120–1149.
- Ordeshook, P. C., & Shvetsova, O. (1997). Federalism and constitutional design. *Journal of Democracy*, 8(1), 27–42.
- Parikh, S. & Weingast, B. R. (1997). A comparative theory of federalism: India. *Virginia Law Review*, 83(7), 1593–1615.
- Persson, T., & Tabellini, G. (2001). *Political economics*. MIT Press.
- Poterba, J., & von Hagen, J. (Eds.). (1999). *Fiscal institutions and fiscal performance*. University of Chicago Press.
- Przeworski, A. (1991). *Democracy and the market*. Cambridge University Press.
- Qian, Y. & Ronald G. (1998). Federalism and the soft budget constraint. *American Economic Review*, 88(5), 1143–1162.
- Revesz, R. (1997). Federalism and environmental policy: A normative critique. In J. Ferejohn & B. R. Weingast (Eds.), *The new federalism: Can the states be trusted?* Hoover Institution Press.
- Riker, W. H. (1964). *Federalism: Origins, operations, and significance*. Little, Brown.
- Rodden, J. (2000). *The dilemma of fiscal federalism: Hard and soft budget constraints around the world* (Working paper). Department of Political Science, MIT.
- Rodden, J., Eskeland, G., & Litvack, J. (Eds.). (2001). *Decentralization and the challenge of hard budget constraints*. MIT Press. Forthcoming.
- Rubinfeld, D. (1987). Economics of the local public sector. In A. U. J. Auerbach & M. Feldstein (Eds.), *Handbook of public economics* (Vol. II, pp. 571–646). Elsevier.
- Sanguinetti, P. J. (1994). Intergovernmental transfers and public sector expenditures: A game-theoretic approach. *Estudios de Economía*, 21(2), 179–212.
- Scotchmer, S. (1994). Public goods and the invisible hand. In J. Quigley & E. Smolenski (Eds.), *Modern public finance*. Harvard University Press.
- Shirk, S. (1993). *The political logic of economic reform in china*. Berkeley: University of California Press.
- Stein, E. (1998). Fiscal decentralization and government size in Latin America. pp. 95–119. In Fukasako, K. & Hausmann, R. (Eds.), *Democracy, Decentralization and Deficits in Latin America*. InterAmerican Development Bank.
- Tiebout, C. (1956). A pure theory of local expenditures. *Journal of Political Economy*, 64, 416–424.
- Wallis, J. J. (2001). *The property tax as a coordinating device: Financing Indiana's mammoth internal improvement system, 1835–1842* (Working paper). University of Maryland.
- Wallis, J. J., Sylla, R. E., & Legler, J. B. (1994). The interaction of taxation and regulation in nineteenth century U.S. banking. In C. Goldin & G. D. Libecap (Eds.), *The regulated economy: A historical approach to political economy*. University of Chicago Press.
- Weingast, B. R., Shepsle, K., & Johnsen, C. (1981). The political economy of benefits and costs: a neoclassical approach to distributive politics. *Journal of Political Economy*, 89, 642–664.

- Weingast, B. R. (1995). The economic role of political institutions: Market preserving federalism and economic development. *Journal of Law, Economics, and Organization*, 11, 1–31.
- Weingast, B. R. (1997). The political foundations of democracy and the rule of law. *American Political Science Review*, 91, 245–263.
- Weingast, B. R. (1998). Political institutions and civil war: Institutions, commitment, and American democracy. In R. Bates, A. Greif, M. Levi, J.-L. Rosenthal, & B. R. Weingast (Eds.), *Analytic narratives*. Princeton University Press.
- Wibbels, E. (2003). Bailouts, Budget Constraints, and Leviathans: Comparative Federalism and Lessons from the Early United States. *Comparative Political Studies*, 36(5), 475–508.
- Wildasin, D. (1997). *Externalities and bailouts: Hard and soft budget constraints in intergovernmental fiscal relations*. Typescript. Vanderbilt University.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Local Politics in Nations and Empires

9

Roger B. Myerson

Contents

9.1	Introduction	186
9.2	Local Order and Property in Village Communities	187
9.3	Credibility of Leadership and Moral-Hazard Rents in the State	192
9.4	Mandarins and Gentry in Traditional Autocratic States	194
9.5	Representative Assemblies and Constitutional Democracy	199
9.6	Conclusions	203
9.7	Cross-References	205
	References	206

Abstract

Successful states maintain a functional relationship between local politics and national politics. Property rights have depended on recognition by local communities since before any states existed. To maintain unity in an extensive domain, a state needs a cadre of agents (mandarins) who expect national leaders to reward them for serving the state above any local connections. But investments require locally rooted investors with confidence in the state's protection, so a stable prosperous state must earn the trust of local elites (gentry). Local and national leadership can be aligned in two fundamentally different ways: by top-down influence of national leaders in local politics and by bottom-up influence of local leaders in national politics. Top-down influence is dominant in successful autocratic states. In successful democracies, national leaders need bottom-up approval from local groups throughout the nation, and local leaders who perform well can become competitive candidates for national leadership. International assistance for democratic development can fail when local politics is neglected.

R. B. Myerson (✉)
University of Chicago, Chicago, IL, USA
e-mail: rmyerson@uchicago.edu

Keywords

Village communities · Local leadership · Property rights · Mandarins · Gentry · Representative assemblies · Political decentralization · State-building

9.1 Introduction

In federal republics, state powers are constitutionally divided between the national government and lower levels of subnational government where officials owe their positions to local politics (see ► [Chap. 8](#) in this Handbook). In the study of comparative politics, such political decentralization is often viewed as fundamentally applicable only in nations that have adopted a federal system of government, which is seen as merely one among many alternative ways of organizing a political system. However, both from a broader historical perspective and from a deeper theoretical perspective, we should recognize that local politics can have fundamental importance in any state that maintains order over an extensive territorial domain.

Discussions about foundations of the state often begin from an assumption that the alternative to a well-organized state would be a Hobbesian anarchy where each individual stands alone against all others, but this is never the case. Nowhere do humans live without basic social structures to help them get along with their neighbors, and the importance of local groups becomes even greater when people cannot rely on a wider state. The establishment of a national state means imposing its national leaders' authority over communities which may already have their own forms of local leadership with local political accountability (Murtazashvili, 2018; Myerson, 2022).

To maintain unity, any successful state needs an executive team that can act forcefully and coherently in any part of an extensive territorial domain. This unifying mission requires a cadre of agents who are dedicated to the central leadership of the state above any local identification. But the resources for a successful state will depend on long-term economic investments that need effective protection from neighboring communities, as well as from the state's central agents. Thus, the stability and prosperity of a national state depend on its ability to maintain a constructive relationship between the centralized agents of the state's national leadership and local groups that are rooted in their communities.

When we approach comparative politics from this perspective, we can find important differences between states, because there are two fundamentally different ways to develop a basic alignment between local politics and national politics in a state. One way is to make local politics dependent on national politics, as happens regularly in traditional autocratic or authoritarian regimes, where national leaders may bestow a privileged status on some favored individuals, who are then recognized as leaders of their community on the basis of their special connections to the wider state. Such privileged status may be granted as a reward for past service to the national leader. Alternatively, alignment can be achieved by making national politics dependent on local politics, by institutionally empowering local leaders in national

politics, as happens in modern democratic states where national leaders are accountable to locally elected representatives in a national assembly. The dependence of national politics on local politics may be even stronger in federal democracies where successful local leaders can become competitive candidates for national political leadership.

These issues are examined in more detail below. After noting the fundamental primacy of local politics from the earliest evolution of human society, Sect. 9.2 offers some general observations about autonomous village communities, where people must maintain a broad consensus about the distribution of ownership and use rights for the land and other resources in their neighborhood. Section 9.3 examines the political foundations of wider states, showing how the problem of maintaining a credible long-term allocation of moral-hazard rents can make the central leadership of a state look like a community that has specialized in maintaining group trust across great distances. Section 9.4 focuses on the general principles for organizing provincial and local government in a traditional autocratic or authoritarian state that grants privileged status to those who have served it and then governs through these privileged elites. To maintain the coherent unity of the state across an extensive domain while promoting local investments, the state's privileged elites may be divided into two broad classes: *mandarins* who have widely itinerant careers in provincial administration and *gentry* who remain rooted in their local districts. The vital importance of trust between the localized gentry and centralized mandarins is discussed with examples from the history of Spanish America. Section 9.5 reviews how the role of local politics in successful states can be fundamentally transformed by the institutional development of representative assemblies and democracy, under which all citizens share the political privileges that autocracies extend only to a limited class of gentry. Section 9.6 summarizes our main conclusions and suggests their potential applicability in international assistance for democratic development.

9.2 Local Order and Property in Village Communities

We humans are social mammals, adapted to living in communities.¹ The folk theorem of repeated games can be interpreted as a general model of how people who live together in a small community can discipline themselves to maintain virtually any pattern of behavior that may be adaptive for their survival. The folk theorem is proven by strategies in which anyone who deviated from his or her prescribed proper behavior would then suffer an adverse change of status in the community and so be treated worse by others.² We may conjecture that some aspects

¹ A comparative analysis by Dunbar (1998) of brain sizes and group sizes among different species of social mammals suggests that human brains may be particularly adapted for life in a community of about 150 individuals, who all know each other and can develop personal relationships with each other.

² See, for example, Sect. 7.5 in Myerson (1991).

of such strategies might be hard-wired in our human brains, such as an inclination to judge the propriety of the behavior of others in our community and a reciprocal fear of losing status in their eyes. An innate expectation that one's good standing in the community may depend on respecting the generally recognized rights and property of others could be particularly adaptive for humans, as social animals with hands for making many useful things (Myerson, 2023).

Coordination in some communal endeavors may require leadership. Trusted public leadership depends on a reputational equilibrium where an individual can expect to be recognized by the community as a leader and so to be accorded special powers and benefits of this high status, as long as these powers are used to provide some expected public services. Improper use of this power can be deterred if leadership entails expected rewards (or moral-hazard rents) which are not less than the benefits that the leader could get by abusing the powers of this position. Successful societies must be able to get people focused on such equilibria with some generally accepted leaders, who can take responsibility for essential public goods that require coordination or management by one person.

Such problems had to be solved among hunter-gatherer bands long ago when our species first spread out of Africa to transform the world. We may conjecture that some of the first uses of human language might have been for a band's leader to assign roles in a hunt or battle and then to distribute shares of any rewards from success but also for others in the band to gossip about whether their leader had exercised this coordinating power appropriately. In this sense, local community politics may have deep roots in the evolutionary history of our species. People seem to have an innate propensity for identifying leadership in their communities and an innate inclination to accept the norms that are affirmed and adjudicated by their community's leaders.

The development of a capability for language also enabled humans to recognize social and kinship links with other communities that share a common language and culture, thus defining a broader social or tribal identity among the communities that are so connected. A propensity for people to identify themselves as part of such a multicomunity society or tribe may also go far back in the evolutionary history of humanity (see ► Chap. 33 in this Handbook; Richerson & Boyd, 1999; Chapais, 2008; Moffett, 2018; and Newson and Richerson, 2021). Groups that were unable to recognize long-term cooperative relationships with anyone outside their local community would have had difficulty defending their place in the world against groups that could use wider cooperation to organize raids to take resources and territory (Mathew, 2022). Before the development of national states with specialized ruling elites, however, the broader coordination of a tribe would depend on negotiations among leading members of its various communities.

Then, from about 10,000 years ago, the great transformative development of agriculture depended not just on some basic understanding of plants but also on the ability of people in farming communities to defend their rights to benefit from the crops that they had worked to plant and cultivate. Before the rise of states that could provide law and order over extensive regions, each village must have had the

necessary leadership to fight for the defense of its territory against its neighbors and to negotiate agreements and alliances with them.

As North (1981, Chap. 7) has observed, the development of agriculture is fundamentally dependent on a system of reliably enforced property rights that can encourage farmers to invest in the cultivation of crops, which they are willing to do only if they have some confidence in their ability to benefit from the resulting harvest. So the fact that farming communities existed in the Middle East for thousands of years before the establishment of the first national state (about 5000 years ago) shows that reliable enforcement of property rights does not necessarily depend on the existence of an effective national government to enforce law and order.

Since the dawn of human history, systems of property rights have been maintained and enforced by people in their local communities. Questions about whether some specific individual has taken benefits from improperly using some specific physical asset (like land or a manufactured item) can only be answered by observers who are locally present and have some familiarity with the individual and the asset. So the enforcement of rights to use valuable assets requires some cooperation among people in the neighborhood of the asset.

Indeed, we might take the definition of a local community to be a group of people who not only reside near each other but also share a broad consensus about the rights of use or ownership that each of them has over the resources in their neighborhood. The terms of this consensus about social order and property rights may be derived from local custom or from administrative decrees of a wider state; but in any case, the community's consensus can generate a strong social imperative for individuals to accept the limits of their locally recognized rights.

A communal consensus about property rights can be maintained only with some general agreement about social procedures for adjudicating disputes, which may depend on some generally accepted form of local leadership. But such power of local leaders to preside over the confirmation of socially enforced property rights ultimately depends on people's general acceptance of their authority in the community. One could get into trouble by denying the authority of leaders who are accepted by everyone else, and so local politics may be fundamentally understood as a coordination game with multiple equilibria. Thus, the identification of local leaders can be guided by local traditions or by salient external directives, according to Schelling's (1957) focal-point effect.

A useful perspective on local order in autonomous village communities may be provided by the observations of Henry Sumner Maine (1861, 1871).³ Maine (1871)

³Henry Sumner Maine [1822–1888] was a British jurist who in 1861 published a successful text on *Ancient Law*. Then he went to India to serve as a senior legal advisor to the British imperial government, where he studied legal problems from traditional Indian communities, applying a sensitivity that was rooted in his long study of the laws of ancient Rome and medieval Europe. After returning to England, he wrote Maine (1871), *Village-Communities in the East and West*, to describe a conceptual framework that could fit both ancient European history and contemporary Indian developments. See further discussion in Redfield (1950) and Myerson (2021a).

observed that nineteenth-century villages in India shared many basic structural features with communities in ancient and medieval Europe, which were familiar to him from his study of ancient law. The territory of the village-community would be composed of a cultivated area, areas of common fields or waste, and the residential area of the village. The cultivated land would be divided among the village households, but they had to cultivate their plots in a coordinated manner according to village rules, while the fields and waste areas could be used in common by the whole village. In the village, each household was ruled by the recognized head of the family. Disputes among the households would be resolved according to customary rules as defined by a village council or assembly. When its peaceful order was disturbed, the village would rely on a chief or head of one preeminent family for leadership of the village's defense.

Maine (1861) had observed that, before the growth of the Roman state, the laws of ancient Rome were applied, not to individuals but to families; and each family was ruled by a patriarch who held despotic power over everyone else in the household. Relationships were regulated by people's generally recognized status, not by bilateral contracts. Maine (1871) found that these communitarian principles were also general characteristics of traditional villages in medieval Europe and in his contemporary India. Maine (1861) noted that individual property rights were introduced into Roman law as Rome became a great state and that property that an individual acquired in military service to the state was the first kind of property that a Roman was legally allowed to own as an individual, not subject to the head of his family. That is, an impetus to separate individuals' legal rights from their families may have been initially driven by a state's need to recruit individuals into its service.

Economists have often suggested that advantages of efficiency can be derived from legal principles that allow any individual to own property and allow any group of individuals to voluntarily form a contractual relationship. But when broader state institutions do not provide enforcement of laws and adjudication of disputes, property rights must be adjudicated and enforced by people in the community. So we cannot evaluate the social efficiency of legal rights without considering the community's cost of maintaining them. This basic point can help us to understand Maine's observation that people's relationships in traditional villages are generally regulated more by social status than by voluntary contractual relationships. A poor village, which could hardly afford to offer public adjudication for arbitrary contracts, might naturally limit its civil protection to rights of social status that are recognized by the community. Then individuals can be motivated to provide various public services for the community if they anticipate that doing so can help them achieve a higher recognized social status.

In a community where people cannot rely on protection from a publicly funded police force, an individual could find it difficult to enforce a claim to valuable property alone without the support of other family members who share an interest in the claim. So collective ownership by a family could become the norm as a way of credibly covering the costs of enforcing these ownership rights. A rule of recognizing rights only for the senior head of each family could also help to reduce the

community's responsibilities for dispute resolution, as the adjudication of intra-family disputes would be left to the head of the family.

Although traditional autonomous villages are often described as democratic, Maine (1871, pp. 164–168) argued that they should actually be considered little oligarchies, and the element of oligarchic inequality in such villages actually tends to worsen when a wider state provides even a minimal quality of regional government. To understand this effect, we must remember that local hardships have always driven some people to leave their homes and move to other communities, whether as migrants or refugees or indentured servants. Such immigrants arrive in a village with no claim to its land and resources, but they may earn a livelihood by the labor that they can provide. If nothing compels the old village families to share the privileges of their status with these newcomers, then their descendants may eventually form a permanent landless underclass in the village. Such inequalities can become part of the accepted order in a community if they are rooted in its history.

However, in the absence of a wider state to maintain regional order, an autonomous village-community could regularly face existential threats of conflict against invaders or other communities in the region. The urgency of mounting a vigorous communal defense in such crises could motivate villagers to offer full citizenship to all residents who fight for the village against its enemies. However, this force for inclusion is eliminated when a wider state protects the village in a regional peace. Thus, if a state provides even tolerably good government while permitting traditional villages to define their own citizenship autonomously, then local village democracy can develop into an oligarchy that may face problems of collective ownership without avoiding problems of class inequality. Alternatively, a state which claims the power to define citizenship could try to compel communities to recognize equal rights for all residents, including new immigrants; but we should understand that such assertions of central state power over citizenship may be resented and resisted by the established residents of many communities.⁴

Even so, the creation of a regional peace should be counted as a benefit that a new state provides to communities in its domain, if a community's costs of taxes or tribute to the state are less than its costs of conflict and defense without the state. An independent community's ability to protect its members and their property is fundamentally limited by the threat of raiding by outsiders. When a community is annexed into a wider state, the state offers protection against intercommunal raids but imposes its own right to tax local property (Scott, 2017). Then local investors need protection both from theft by neighbors and from expropriation by state agents, and the state's top leaders hold responsibility for the discipline of state agents. But while forces of the state may be capable of seizing local residents' property and even taking their life or liberty, many common property disputes will

⁴So customary land tenure rules in which rights to use communal land are allocated to members of the community can exacerbate ethnic tensions by raising questions about who should have such membership rights (Boone, 2009).

still involve details of local conditions that distant state officials cannot be expected to know. Thus, secure property rights may depend on cooperation between local community leaders and national leaders of the wider state. To consider how such cooperation can be sustained, however, we need a basic theory of states.

9.3 Credibility of Leadership and Moral-Hazard Rents in the State

States are established by political leaders, who organize groups of people and mobilize them for action toward shared goals. In many periods of history, the founder of a new state could begin a military career by taking responsibility for defense of his local community and then, after developing a reputation for effective leadership, could recruit broader support to build forces capable of establishing order over a wider region.

Ancient social philosophers perceived that the basic foundations of a state may depend on its leaders' reputations for rewarding those who serve them. According to Xenophon's *Education of Cyrus*, Cyrus the Great founded the Persian Empire with one essential quality of leadership: a reputation for reliably rewarding good service. The captains who flocked to Cyrus' banner were confident that they would be well rewarded for serving him in battle, and their recognition of Cyrus as the reliable paymaster of a great army made him a leader who could build a state which was then the greatest in the world (Xenophon & Ambler, 2001).

To compete for power in any society, a political leader needs the active voluntary support of many people in a group or faction, and these supporters must be motivated by some expectation of future reward in the event of their success. But when rivals have been defeated, a leader may be able to enjoy the fruits of power without such broad support, and so an established leader may be tempted to ignore the claims of past supporters. Thus, a successful leader must somehow be credibly committed to reward those who have supported him in the past. This is the central credibility problem of political leadership.

This central credibility problem can be solved within a leader's own faction of supporters, without relying on any external court. A leader becomes politically accountable to a group of supporters when he could not hold power without their confidence in his promises of rewards for good service. This accountability can be effective when these key supporters meet regularly in a central council or court where they can monitor their leader's distribution of rewards and share any evidence of his denying appropriate rewards to any of them. Myerson (2008) showed that, in negotiation-proof equilibria of a simple model of sequential contests for political power, a contender for power would be unable to credibly recruit any supporters if he did not organize them in such a way that he would fall from power if he failed their trust. Then the standards of behavior which a leader must maintain to keep his supporters' confidence can effectively become a fundamental political law or

personal constitution for the leader. Such personal constitutions for a state's political leaders may underlie the enforcement of all other constitutional provisions in a state.

When the domain of a state is too large for its ruler to supervise personally, the state cannot exert control over any part of its domain unless it has some agent who is close enough to monitor compliance with the state's laws and ensure that any disobedience would be punished. So states regularly have appointed governors who exercise delegated powers of the state within a province of manageable size, to supervise enforcement of the state's laws and collection of the state's revenues in the province.

However, as governors and other agents exercise power in the name of the state, they may find many ways to profit personally from abusing their delegated powers. Any state must solve the moral-hazard problems of getting its agents in every province to faithfully implement the state's policies instead of using their delegated powers for personal gain. To solve these moral-hazard problems, any agent of the state must expect that the rewards for good service can be better than what the agent could get by abuse of power. The rewards from good behavior that must be expected by agents of the state, to deter them from abusing the powers of their office, are called the *moral-hazard rents* for their office.

Becker and Stigler (1974) analyzed the problem of motivating government officials to serve appropriately during a long career in which opportunities for abuse of power can arise every day. Their analysis showed that how the cost of moral-hazard rents for such officials can be minimized by an incentive system in which large late-career rewards are promised for officials who maintain a record of good service. Any evidence of bad service would be punished by early dismissal without such rewards. By back-loading the promised payment of moral-hazard rents to a later point in time (without changing their expected present discounted value), these rewards can be used to motivate more years of good service in offices of increasing responsibility.

During an official's career, however, promises of back-loaded moral-hazard rents become a debt of the state, which its ruler could be tempted to repudiate by falsely finding fault in the official's performance. So the credibility of these promised rewards may depend on the official having personal connections in the ruler's central council or court, so that a wrongful dismissal of the official would raise political risks for the ruler as it became widely known among other high officials whose trust the ruler needs. That is, responsible officials of the state must form a tightly connected social group. Literacy of state agents has been essential for strong states throughout history, not only so that the ruler's commands can be reliably conveyed to agents throughout the domain but also so that responsible officials who are dispersed throughout the domain can maintain personal connections with each other by exchanging letters.

The advantages of delaying the payment of moral-hazard rents can actually be pushed beyond the end of an official's career if officials also want good things for their children. That is, the cost of paying moral-hazard rents to one agent of the state can be reduced by promising that the agent's children will be also given priority for

valuable responsible positions in the state. Thus, responsible agents of the state may tend to become a hereditary elite (Myerson, 2015a).

From this perspective, we should understand that the elite who exercise power in a great state may form a community of individuals with a strong network of social and family connections. That is, many familiar structures of community life that have deep roots in human evolution can be found also in privileged elite around the center of power of a great state, as if the rulers and their courtiers formed one more village in the realm. The one basic difference from traditional local communities, however, would be that the members of a state's governing elite must learn (generally with writing) to maintain their connections across much greater distances than members of traditional local communities.

9.4 Mandarins and Gentry in Traditional Autocratic States

In an insightful analysis of the origins of nationalism, Anderson (1991, pp. 53–56) compared the career paths of administrative officials in an absolutist monarchy to the routes of religious pilgrimage that have created a sense of meaningful unity for adherents of great religions. But Anderson also saw a fundamental difference between these absolutist functionaries, whose jobs required them to be ready for reassignment to any province in the realm, and feudal nobles, whose careers were largely tied to their family's estate, with occasional visits to the monarch's court. We may suggest, however, that the fundamental distinction here is not between absolutist and feudal monarchies as two types of political systems but between two classes of agents on whom rulers throughout history have relied for exercising power over an extended domain. We may refer to these two elite classes as *mandarins* and *gentry*, and the difference between feudal and absolutist monarchies may then be defined by the greater relative importance of mandarin functionaries under absolutist monarchs. Recognizing the essential complementary roles of the resident land-owning gentry and the perambulatory mandarin governors in traditional monarchies can help us to understand the fundamental relationship between local politics and national politics in any political system.

As we have seen, the ruler of a great realm can solve the problem of asserting power in distant parts of his domain by appointing local governors to exercise delegated powers within a province of more manageable size. Within the province, each resident would fear to disobey the governor's commands when everyone else in the province is expected to obey; so it could be a rational equilibrium for everyone in a province to obey their appointed governor, even without any immediate threat of external force. Thus, local acceptance of the governor's authority would enable the governor to maintain stable order in the province with little or no action by the ruler's armed forces, which could be devoted instead to countering foreign challengers elsewhere on the frontiers of the realm.

However, if people in a province got so accustomed to obeying their governor that they forgot that his power was supposed to be contingent on the ruler's approval, then the governor could effectively become an independent local ruler. Thus, to

maintain unity of the realm, it has been a common practice for rulers to transfer their governors from one province to another at regular intervals. A transfer to a larger province could be a welcome promotion for the governor, but the real point of the transfers is to keep reminding everyone that their province owes obedience, not to their governor personally but to only their governor as a designated appointee of the ruler. These cross-country transfers form the circuitous routes of mandarin careers that Anderson saw as defining the unity of the realm.

The regular movements of mandarin administrators imply that they cannot take a long-term perspective on the development of any one province.⁵ But the wealth of a province depends on people making economic investments there that may yield profits only over a long time horizon. Thus, the prosperity of the realm requires another class of agents, the gentry, who have enough power to protect investments in a province where they maintain a long-term presence, while the mandarins who supervise their province may come and go. As the ruler's appointed representative, a mandarin governor must have substantial power to command the gentry of a province. But the provincial gentry must have also some right of appeal to the ruler's central council or court when a governor's actions threaten their basic property rights.

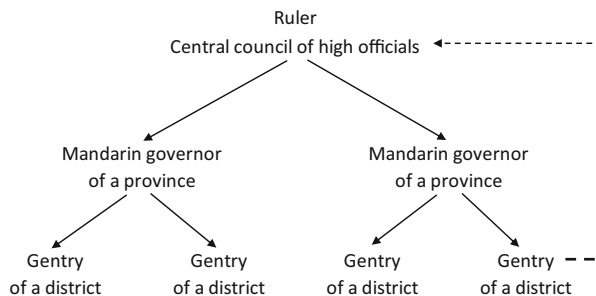
The ability of the local gentry to get high-level attention for their concerns requires that they have some local organization and that at least some leading members of the local gentry should maintain personal contacts with the state's central court. To ensure that any organized group of local gentry is subject to supervision by a mandarin governor, the gentry would typically be organized in local districts that are contained within a governor's province. Thus, empires have commonly been divided into provinces where a centrally appointed mandarin governor exercises broad powers, and then each province may be further subdivided into local districts where leaders of the resident gentry help to maintain local order (see Fig. 9.1). Within this framework, an imperial elite of peripatetic mandarins can maintain cohesive government of the empire, while local elites of resident land-owning gentry maintain a system of local property rights in which they can profitably invest.⁶

Historical monarchies have differed greatly in the relative ability of their centrally directed mandarin administrators to regulate the locally based gentry, with minimal central control in feudal states and maximal central control in absolutist monarchies (Finer, 1997). Even in the first known kingdom, scribes of ancient Egypt already

⁵If governors' rewards depended on the long-term economic performance of their former provinces, the result would be to induce problems of moral hazard in teams, as the long-term performance would also depend on the actions of all subsequent governors. In team problems, commitments to rewarding many agents can generate an adverse incentive for the central paymaster to reduce performance; see Holmström (1982).

⁶Myerson (2021b) shows a simple model in which local investment must be inefficiently undersupplied if the central state takes responsibility for the local justice system but gives local residents no political voice.

Fig. 9.1 Structure of power in a simple autocratic empire



exemplified the category of mandarin administrators, supervising the delivery of taxes and tribute from the provinces.

The creation of a privileged land-owning gentry can arise naturally from two problems that confront a new state. First, the state must find a way to establish some basic control over communities throughout its domain, to extract taxes and suppress insurgency. Second, the new state must fulfill its promises to those who supported its claims to sovereignty over this domain, including those who served in its conquering armies as well as indigenous local leaders who helped induce their communities to accept the new state. A natural solution to both problems is for the new state to reward its supporters by granting them property rights and authority in a district where they will then be expected to maintain order. Then privileges of the gentry may be generally understood as rewards for past service to the state, and the ruler's power to certify new members of this privileged class could also be used to motivate service to the state in subsequent generations.

Maine (1871) offered a detailed analysis of how weak feudal states in medieval Europe transformed autonomous village-communities into feudal manors under local landowners. The territory of an autonomous village-community would become the manor of a feudal lord, who would fit our definition of landed gentry when the manor was his principal residence. The community's common lands would become the lord's private domain, and village households became tenants who might retain customary rights in the community's cultivated areas. Then the village council would become the lord's court, with the lord as judge and with villagers as jurors. Maine noted that rights of lordship over a manor were sometimes derived from a grant by the ruler to reward past service, but local lordships could also be established by the state's recognizing and elevating the position of an indigenous chief from a leading local family.

Even when feudal lordship was bestowed on the traditional war-leader or chief of a formerly autonomous community, one may find substantial historical injustice in the conversion of communal property into private property of the lord. Most importantly, the state's recognition of the indigenous chief would make him a local lord whose position no longer depends on the community's approval of his

public service. In effect, the wider state would endorse general obedience to the local lord as a focal equilibrium for people in the community, provided that the lord kept the community as an orderly tax-paying component of the state's domain.

It may be instructive also to consider the institutional structures for local and regional government that were applied in Spain's vast American empire after 1492. Haring (1947, Chap. 9) notes that on entering a new region to subdue and occupy it, the Spanish conquistadors' first step would be to establish a town and set up its basic municipal organization. The first citizens of the town came with the expedition of conquest and formed a town council (*cabildo*), which thereafter became an essentially self-perpetuating institution, as new members were regularly selected by old members subject to approval by a provincial governor. The town then functioned as an organization of the local gentry to maintain local order and property rights for its citizens, including those who received large estates in the district around the town. The town's militia would be the primary force for suppressing any uprising of the indigenous population.

The towns in Spanish colonial America were supervised by an appointed provincial governor, who fit the mandarin pattern of being an itinerant functionary from Spain or elsewhere in the empire. This system of local government in the Spanish empire was ultimately derived from the Roman Empire, where imperial governors supervised provinces that contained formerly autonomous city-states, which had primary responsibility for order and taxes within their local districts. In the Roman Empire, a city council could complain to the emperor about a governor's actions. The towns of the Spanish Empire in America could appeal disputes to the *audiencia* or administrative court of a wide region, as well as to the king's council in Spain. These regional audiencias were key institutions for maintaining imperial control across the great distances of Spain's American empire. The members of an *audiencia* were professional judges who exercised their broad powers as a group, and so their ability to monitor each other closely could help to mitigate many problems of moral hazard in the exercise of distantly delegated powers.

Our general distinction between gentry and mandarins in imperial elites can be made particularly sharply in the Spanish colonial empire, where these groups were generally recruited from different populations. After the first generation, the gentry in colonial towns were generally creoles, born in America to families of Spanish ancestry, while the mandarins who staffed the higher levels of imperial administration (including the governors and *audiencia* judges) were mostly *peninsulares*, born in Spain. Higher administrators might be less inclined to engage in local colonial corruption when their families and personal connections were rooted in Spain (Salgado, 2022; Guardado, 2022), while the gentry's local ties gave them more interest in maintaining the empire's essential local infrastructure. Together they constituted an imperial system of power that could overwhelm the autonomous communities and states of the indigenous native Americans. None of the indigenous native-American states could match the Spanish Empire's ability to maintain local militia forces under the creole gentry, which could contain a native revolt, while the higher network of mandarin administrators could call in greater imperial forces from far away.

Machiavelli (1513, Chap. 4) observed in history that some kingdoms were easy to invade but hard to conquer, while other kingdoms were much harder to invade but then more easily conquered after a successful invasion. This distinction can be derived from the relative strengths of the local gentry and the central mandarin administration of a kingdom. A stronger gentry could retain a larger share of local revenues, leaving less resources for the central government's defense of the kingdom's frontiers; but in that case, a successful invader would meet continuing resistance from local leaders who were determined to defend their privileges of power. Thus, administrative reforms in the Spanish Empire after 1782 that increased the central government's share of revenues also decreased the local gentry's motivation to preserve the Empire's authority in their provinces after the Napoleonic invasion of Spain in 1807 (Chiovelli et al., 2023).

In any case, successful states depend on local political elites' ability to trust the basic parameters of their relationship with the central administration of the state, and such trust requires broadly shared expectations that may take time to develop. This point is illustrated by the remarkable fact that, when the Spanish Empire in America finally dissolved, the territories of the new states were generally defined by the jurisdictions of the Empire's regional audiencias. As the seats of audiencias became national capitals, regional power relationships from the colonial period could provide the foundation for shared expectations and trust between the leading citizens of provincial towns and the central administration of the new nation.

We have defined the gentry here as a class of individuals who are rooted in their local communities but who have an elite status with sufficient voice in the state to be confident that their local property rights will be respected by officials of the state. The existence of such nationally respected local elites would depend on their having some reliable way of conveying complaints against state officials to a central forum where an evident failure of justice could damage the ruler's vital political reputation for being a trustworthy patron. The overall size of the gentry could be constrained by the inability of a traditional state with limited administrative capacity to assure such reliable access for more than a small fraction of the population.

However, in dysfunctional regimes that lack any reliable institutions for protecting residents far from the seat of power, there may be wide regions where a class of gentry so defined does not exist, and then we must expect local economic investment to be small and vulnerable. People who cannot trust their national leaders often prefer to rely on informal structures of local leadership, which are harder for outsiders to monitor and manipulate (Scott, 2009). In such cases, strengthening the central state administration could actually weaken security of local property rights, when agents of the strengthened state get greater ability to expropriate local property (as has been observed in the history of Afghanistan by J. Murtazashvili and I. Murtazashvili (2021)). We should understand that the state's national leaders have a general interest in registering ownership of property for the purpose of taxing it, but they would have less interest in actually protecting local property rights when local owners have no national political voice or influence.

If a central state is too weak to keep disciplined security forces in local communities throughout its domain, then other providers of local protection may emerge under independent leadership, but the weak state could still try to maintain its general sovereignty by co-opting successful local leaders and offering them positions of authority in other provinces. In that case, the central state would be perversely contributing to a spread of roving bandits in the sense of Olson (1993); see also Verweijen (2015, Sect. 10.1.2) and Sanchez de la Sierra (2020). Thus, some of the worst forms of economic underdevelopment can result from a lack of any basic trust between local residents and the state that claims sovereign power over them.

9.5 Representative Assemblies and Constitutional Democracy

The domination of local politics by an elite who owe their privileged status to recognition by the state's central authorities is one way of assuring some alignment between local leadership and the wider state. An alternative way of promoting such alignment would be to make the state's supreme political leadership more dependent on local politics, but incumbent rulers might naturally consider such dependence to be an undesirable inconvenience. So it is not surprising that institutions that compel national leaders to seek approval from representatives of local elites are hard to find in the history of states before 1200. However, when such representative assemblies were introduced in England and other states of medieval Europe, they substantially increased local elites' willingness to trust and support their national leaders, so that the net result could be a significant increase in the strength of the state (Congleton, 2001).

The English Parliament and other early representative assemblies developed in the thirteenth century as institutions where the ruler could seek support from local elites for new tax revenue after hearing their complaints about abuses of power by the ruler's agents. In England at this time, one notable cause of complaints was that the king's sheriffs (or governors) increased their opportunities for extracting legal fees from local gentry by summoning them more frequently to local courts, where local gentry could develop a shared sense of grievance against royal officials (Maddicott, 1984). The early English Parliaments brought together representatives from town councils and rural gentry in every part of the realm, as these were the people who had basic responsibility for maintaining local order and raising local revenues for the state. These local elites were more inclined to trust and support a monarch who regularly gave them a national forum for communicating their local complaints against abuses of power by royal officials (Maddicott, 1984, 2010; Coss, 2003; O'Callaghan, 1969). As the English Parliament began meeting with some regularity under Edward I, it quickly became evident that a demand for taxes would yield more revenue if the king got it approved in Parliament (Prestwich, 1997, Chap. 17). Thus, the institution of Parliament increased the king's ability to credibly promise privileges for local elites in exchange for their raising new revenues for the state. Within a couple of generations after the first English Parliament, Parliament

showed it could play a key role in approving a new king's accession to the throne, in reforming the system of local government, and in raising revenues for military operations that shifted the balance of power against France in what would become the Hundred Years War (Ormrod, 2012).

The fact that towns were represented in Parliament could make them more effective as engines of economic growth. Unlike the tradition in Spain, towns in medieval England did not have jurisdiction over the rural areas around them, and so the town councils and rural gentry each had their own representatives in the English Parliament. To see the importance of such representation, it may be useful to compare the political institutions of England, Spain, and Poland. Around 1600, all of these countries had representative national assemblies (called the Cortes in Spain, the Sejm in Poland), and in some ways Poland's national assembly might have seemed the most advanced by modern standards. But only the rural gentry had representation in the Polish Sejm; Polish towns had no voting representatives (Jędruch, 1998). The Spanish Cortes included representatives from the major towns, but these towns also controlled much of the rural land around them. Only in England (and Netherlands) do we find towns that were represented in the national assembly but did not control the agricultural land around them. For the local leaders of a town that did not control a larger region, the only taxable source of wealth was the town's reputation as a good place for people to do business. When such towns exercised power in the national political system, they would naturally favor policies that enabled them to assure good protection of property rights for their residents, and they could support rural peasants' rights of mobility to come work in the towns.

In the 1700s, rural representation in Parliament also had a key role in the development of turnpike trusts, which were local companies that built toll roads throughout England, giving England the best transportation system of the eighteenth century, and thus setting the stage for the industrial revolution (Albert, 2007). These toll roads were managed by local county leaders, who retained profits from tolls on well-maintained roads, but the tolls had to be nationally regulated as part of a national transportation network. The national government's dependence on local elites through their representatives in Parliament made such local investments secure against central expropriation. Thus, the development of the world's first modern land transportation network depended essentially on England's system of parliamentary government, which made the national government responsible to locally elected representatives. (The development of standardized macadam road-building technology in this period facilitated the subsequent development of modern roads elsewhere in the world.)

The early development of representative assemblies in Europe generally included representation only for recognized local elites, which could include town councils and rural gentry. Under this system, only a small subset of the population would have sufficient political voice to be confident of their local property rights being respected by officials of the state, and so local economic investments by others outside the privileged elite would be severely limited. But the base of local investments can be enlarged when political rights are extended democratically to broader groups in the population. For an early observation of the significance of such democratic extension of political voice for modern economic growth, we may recall this remarkable passage from Adam Smith (1776):

In England, a lease for life of forty shillings a year value is a freehold, and entitles the lessee to vote for a member of parliament; and as a great part of the yeomanry have freeholds of this kind, the whole order becomes respectable to their landlords on account of the political consideration which this gives them. There is, I believe, nowhere in Europe, except in England, any instance of the tenant building upon the land of which he had no lease, and trusting that the honour of his landlord would take no advantage of so important an improvement. Those laws and customs so favourable to the yeomanry have perhaps contributed more to the present grandeur of England than all their boasted regulations of commerce taken together. (*Wealth of Nations*, Book III, Chap. 2, p. 415)

The central focus of the economics profession since Adam Smith has been on regulations of commerce that can encourage greater economic growth. But in this passage, Smith tells us that basic political and legal rights for small farmers, which enabled them to invest in improving their land without fear that the benefits of such improvements would be expropriated by a landlord increasing the rent, may have been the most important factor in the prosperity of England in his time. He indicates that this empowerment of poor tenant farmers was a result of particular legal and political institutions of England, including the participation of small farmers in local parliamentary elections. Political developments in England that gave political voice to a great mass of small farmers in turn enabled them to invest securely in economic improvements which marked the start of modern economic development.

Mass local political participation was also cultivated after 1620 in English colonies of North America, where institutions of local self-government were granted to encourage British settlers to come to America and to offer loyal service in local militias. As argued by Tiebout (1956), mobility of people and resources can motivate a ruler to offer good government services, to attract tax-payers and investors into his domain; but long-term promises of good treatment by the government become more credible for people when they get democratic voting rights (see also Myerson (2010)).

Since the nineteenth century, democratic states have extended voting rights to more of their resident population, and when all adult citizens can vote, then there should be no formal distinction between citizens and gentry, which could essentially make the latter term obsolete. Extending democratic privileges to broader groups has enlarged the set of people who can invest in improving their economic status, with confidence that their investments will have protection under the state. Thus, miracles of economic growth in modern democracies have been based on people throughout a nation confidently making significant new investments to improve themselves and their communities.

It is sometimes assumed, however, that successful democratic development depends only on citizens' rights to vote in elections of national leaders, as if national democracy needed no connection with local politics. Such an assumption would ignore the historical fact that representative institutions of national democracy originally developed from institutions for maintaining a balanced working relationship between local community leaders and the state's national leadership. A comparison between the American Revolution and the French Revolution provides further evidence of the potential importance of local political roots for successful national democracy.

The United States was established as an independent nation by a Congress of delegates from 13 provincial assemblies, each of which consisted of representatives elected by their communities. Since America's first national elections in 1788, American politics has always been characterized by a broad concern for maintaining an appropriately balanced distribution of powers between elected local governments and the elected national government. In contrast, the revolutionary establishment of democratic government in France from 1789 was accompanied by a centralization of power. As Alexis de Tocqueville noted after visiting America in the 1830s, the French Revolution was against both royalty and provincial institutions (Tocqueville, 1945, p. 100). Regional parliaments, which had provided significant political decentralization in France's Old Regime, were viewed as vestiges of hereditary privilege and were suppressed by the Revolution. Within a dozen years after the French Revolution, the underfunded and chaotic system of revolutionary local governments had been brought under central direction by appointed agents of a national dictator (Cobban, 1943; Myerson, 2015b).

We should understand that successful democracy requires more than just elections. For democratic competition to effectively benefit the public, voters must have a choice among candidates with proven records of public service who have developed good reputations for exercising power responsibly in elected office. When such trusted leadership is lacking, democracy is inevitably fragile. This essential supply of trusted democratic leadership can develop best in responsible institutions of local government, where successful local leaders can prove their qualifications to become strong competitive candidates for higher office. In effect, responsible democratic local governments can strengthen national democratic competition by lowering barriers against entry into national politics (for a formal model, see Myerson (2006)). From this perspective, we could argue that even a small nation may benefit from having some decentralization of power to autonomous subnational governments, where future candidates for national leadership can demonstrate their ability to serve the people.

An incumbent national leader, however, might naturally prefer not to face competition from candidates who have demonstrated effective leadership in local government. Thus, we may expect national leaders to advocate a more centralized state, where autonomous local political institutions are weak or nonexistent. Given that elected national leaders are generally the most influential people in their country, it can become very difficult to introduce responsible democratic local government in countries where democracy has been introduced first only at the national level. However, a constitutional system with democratic local government can become politically stable once it is established. When governors and mayors have been locally elected, they become local power-brokers from whom national politicians must regularly seek support in competition for national power, and then it would be very costly for any national leader to threaten the constitutional powers of these elected local officials.

Locally accountable local government can also have fundamental advantages in the provision of local public goods that are essential for modern economic growth. With democracy in local politics, local leaders may win their positions by providing

public goods and services to earn people's approval, instead of depending on central connections. Even a poor community may be able to mobilize resources for essential public investments when members of the community are coordinated by local leaders whom they can trust to appropriately reward contributors and discipline free-riders (Fortmann, 1983). Such trust can be expected only from leaders whose authority is based in local politics. Administrative officials who are appointed by the national government will inevitably be more concerned about national political priorities and less committed to earning people's trust in a community from which the officials may expect to be transferred at their next promotion (see also Myerson (2021b)).

Local property-rights enforcement should be included among the local public goods that can benefit from local accountability. This is the basic reason why greater local private investments may depend on investors having some political voice in an appropriately responsible local government.

Local defense against invaders may also be counted among the local public goods that can benefit from democratic decentralization. In a nation where locally elected officials exercise substantial authority, every part of the nation will have local leaders who have a proven ability to mobilize their neighbors for political action and who have a stake of power in the national state that can motivate them to lead their neighbors in its defense. This point extends Machiavelli's classic observation that some decentralization of political power can make a nation harder to conquer.

Other advantages and disadvantages of democratic decentralization have been analyzed in the literature on fiscal federalism (see an insightful summary by Oates (2005) and a thoughtful critique by Treisman (2007)). Decentralized governments can have advantages for flexibly providing local public goods that vary across regions, in response to different local conditions. If a centralized government had no constraints or norms against offering different public services in different regions, then a geographically based governing coalition could use such flexibility to concentrate public spending in the regions where it gets its political support, neglecting people elsewhere. On the other hand, a centralized government may be better able to take account of externalities that local public goods in one region may have for people in other regions. With decentralization, a possibility of subsidies across regions could generate soft local budget constraints that encourage people to misrepresent their demand for local public goods. We should note that all of the advantages of democratic decentralization depend on hardening of local public budgets. A soft local budget constraint would mean that local public funding depends on negotiations with national authorities, and this dependence would generate a basic political imperative for electing local leaders whom the national authorities would favor.

9.6 Conclusions

We have developed a theory of the state that is compatible with the historical fact that local politics has been an essential part of human life since long before the first establishment of territorially extensive states. Humans evolved to live in

communities that maintain systems of ownership or usage rights that define how members of the community can exploit the resources around them. Individuals will not invest in improving their local resources without some confidence that their community will respect and protect their right to profit from the improvement, and such confidence must have existed at least since first origins of agriculture.

But a community's ability to protect its members and their property is necessarily limited by the possibility of raiding by outsiders. Thus, wider states could establish themselves by offering regional protection against intercommunal raids and by taxing the surplus that is generated by this regional peace. Any state needs a centralized team that can project sovereign force throughout an extensive realm, to keep the internal peace and implement state policies coherently. To maintain the coherent unity of the state, its responsible agents should identify their long-term interests with the state's central leadership, effectively separating themselves from local interests of any one community.

So the political economy of property rights essentially involves both local politics and national politics. A broad consensus within a community about who has rights to use and exploit various resources in its vicinity can be maintained only with some form of generally accepted local leadership that can provide for adjudication of disputes. But the wider national state also has an interest in registering property rights, for the purpose of taxing them, and so local investors need also wider protection against confiscatory tax increases or expropriation by centralized agents of the national state.

Thus, the stability and prosperity of a national state depend fundamentally on its ability to earn the trust of local elites. Such trust must be sustained by a social equilibrium that involves a general recognition of some basic rights, channels for communicating complaints about violations of these rights, and some broad class identification to ensure that a local rights violation could provoke a wider loss of essential trust from groups across the nation. That is, a successful state needs some fundamental connection between its national politics and local politics in communities throughout its domain. We have found two general ways establishing such connections between local and national politics. In traditional autocratic states, national political leaders can bestow a privileged status on favored individuals who may then use their privileged connections in the wider state to establish themselves as local leaders in their communities. Conversely, in successful modern democracies, the powers and positions of national leaders depend on approval from local groups throughout the nation, and autonomous local leaders who perform well can become competitive candidates for national leadership. That is, local politics may become dependent on national leadership in autocratic states, but successful democratic states make national leadership more dependent on local politics.

One area where this theoretical perspective offers significant policy implications is in international assistance to support democratic national development. A national government's ability to regulate foreign aid can effectively make the aid a valuable external resource that increases the national leader's ability to govern with less local support. Thus, international assistance has commonly been a force for the centralization of power in poor countries, which can undermine the balanced relationship

between local and national politics that is essential for successful democratic development. So if international assistance is not to be counterproductive, aid coordinators need a deeper understanding of local politics and its role in national development (Khemani et al., 2016). With such understanding, donors might insist that some portion of their aid should be directed through local institutions with locally elected leadership.

The need for deeper understanding of the foundations of successful democratic states was tragically demonstrated by the 2021 collapse of the US-supported Republic of Afghanistan, where the last President, Ashraf Ghani, had been widely regarded as an expert on fixing failed states. Ghani and Lockhart (2008) cogently argued that the ultimate goal of state-building assistance should be to help a country establish an effective government that is accountable to citizens. But from our perspective, serious questions should have been raised about whether effective democratic accountability could be achieved in a state that centralized all responsibility for government under one elected official, the President of Afghanistan, leaving no public responsibilities for local leaders who could be directly accountable to their communities.

Theories of constitutional democracy have regularly emphasized different ways of constitutionally dividing power, particularly among various offices of the executive, legislative, and judicial branches of the national government; but our perspective here suggests a vital primacy in the division of power between national and local political leaders. This point should be particularly relevant for missions to support democratic state-building. Broadly accepted forms of local leadership can be found even in failed states, where a lack of basic consensus about national leadership could make proposals to divide it into separate branches seem premature or quixotic. But even for the authors of the US Constitution in 1787, the balance of power between the new Federal government and the existing local authorities was a primary concern, because local authorities controlled the ratification process; and a separation of powers among branches of the Federal government was introduced as a mechanism for credibly constraining Federal power.

In the language of the American Revolution, the *people* who formed the fundamental basis for the new nation were understood to be the enfranchised inhabitants acting together in their local communities (Waldman, 2014). If this understanding had been applied in Afghanistan before 2021, interveners for democratic state-building might have had more appreciation for the autonomous authority of traditional village institutions (Murtazashvili, 2016), instead of trying to establish a centralized state without any role for locally accountable leadership.

9.7 Cross-References

- [Institutions Before States: Aggressive Egalitarianism and Private Property](#)
- [The Performance and Stability of Federalism: An Institutional Perspective](#)

Acknowledgment of Funding Open access is made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Albert, W. (2007). *The turnpike road system in England: 1663–1840*. Cambridge University Press.
- Anderson, B. (1991). *Imagined communities: Reflections on the origin and spread of nationalism*. Verso.
- Becker, G., & Stigler, G. (1974). Law enforcement, malfeasance, and compensation of enforcers. *The Journal of Legal Studies*, 3, 1–18. <https://www.jstor.org/stable/724119>
- Boone, C. (2009). Electoral populism where property rights are weak: Land politics in contemporary sub-Saharan Africa. *Comparative Politics*, 41(2), 183–201. <https://doi.org/10.5129/001041509X12911362971990>
- Chapais, B. (2008). *Primeval kinship: How pair-bonding gave birth to human society*. Harvard University Press.
- Chiovelli, G., Fergusson, L., Martinez, L., Torres, J. D., & Caicedo, F. V. (2023). *Bourbon reforms and state capacity in the Spanish empire* (SSRN working paper). <https://doi.org/10.2139/ssrn.4545407>
- Cobban, A. (1943). Local government during the French revolution. *English Historical Review*, 58 (229), 13–31. <https://doi.org/10.1093/ehr/LVIII.CCXXIX.13>
- Congleton, R. D. (2001). On the durability of king and council: The continuum between dictatorship and democracy. *Constitutional Political Economy*, 12, 193–215. <https://doi.org/10.1023/A:1011605310916>
- Coss, P. (2003). *Origins of the English gentry*. Cambridge University Press.
- de Tocqueville, A. (1945). *Democracy in America* (Vol. 1). Knopf.
- Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, 6(5), 178–190. [https://doi.org/10.1002/\(SICI\)1520-6505\(1998\)6:5<178::AID-EVAN5>3.0.CO;2-8](https://doi.org/10.1002/(SICI)1520-6505(1998)6:5<178::AID-EVAN5>3.0.CO;2-8)
- Finer, S. E. (1997). *The history of government from the earliest times*. Oxford University Press.
- Fortmann, L. (1983). *The role of local institutions in communal area development*. Report to Botswana Ministry of Local Government and Lands. http://pdf.usaid.gov/pdf_docs/PNAAT392.pdf
- Ghani, A., & Lockhart, C. (2008). *Fixing failed states: A framework for rebuilding a fractured world*. Oxford University Press.
- Guardado, J. (2022). Hierarchical oversight and the value of public office: Evidence from colonial Peru. *Journal of Politics*, 84(3), 1353–1369. <https://doi.org/10.1086/717757>
- Haring, C. H. (1947). *The Spanish empire in America*. Oxford University Press.
- Holmström, B. (1982). Moral hazard in teams. *The Bell Journal of Economics*, 13(2), 324–340. <https://www.jstor.org/stable/3003457>
- Jędruch, J. (1998). *Constitutions, elections, and legislatures of Poland, 1493–1993*. Hippocrene Books.
- Khemani, S., Dal Bó, E., Ferraz, C., Finan, F., Johnson, C. S., Odugbemi, A., Thapa, D., & Abrahams, S. (2016). *Making politics work for development* (World Bank policy research report 106337). World Bank Group. <http://documents.worldbank.org/curated/en/268021467831470443/Making-politics-work-for-development-harnessing-transparency-and-citizen-engagement>
- Machiavelli, N. (1513). *The prince*. English translation by Luigi Ricci. Entreacacias SL, 2020.
- Maddicott, J. R. (1984). Magna Carta and the local community 1215–1259. *Past & Present*, 102, 25–65. <https://www.jstor.org/stable/650759>
- Maddicott, J. R. (2010). *The origins of the English parliament 924–1327*. Oxford University Press.
- Maine, H. S. (1861). *Ancient law*. John Murray. <https://archive.org/stream/ancientlawitsco18maingoog#page/n6/mode/2up>
- Maine, H. S. (1871). *Village-communities in the east and west*. John Murray. <https://archive.org/stream/villagecommunit00mainuoft#page/n9/mode/2up>
- Mathew, S. (2022). Turkana warriors' call to arms: How an egalitarian society mobilizes for cattle raids. *Philosophical Transactions of the Royal Society B*, 377, 20210144. <https://doi.org/10.1098/rstb.2021.0144>

- Moffett, M. (2018). *The human swarm: How our societies arise, thrive, and fall*. Basic Books.
- Murtazashvili, J. (2016). *Informal order and the state in Afghanistan*. Cambridge U Press.
- Murtazashvili, J. (2018). A tired cliché: Why we should stop worrying about ungoverned spaces and embrace self-governance. *Journal of International Affairs*, 71(2), 11–29. <https://www.jstor.org/stable/26552327>
- Murtazashvili, J. B., & Murtazashvili, I. (2021). *Land, the state, and war: Property institutions and political order in Afghanistan*. Cambridge University Press.
- Myerson, R. (1991). *Game theory: Analysis of conflict*. Harvard University Press.
- Myerson, R. (2006). Federalism and incentives for success of democracy. *Quarterly Journal of Political Science*, 1, 3–23. <https://doi.org/10.1561/100.00000002>
- Myerson, R. (2008). The autocrat's credibility problem and foundations of the constitutional state. *American Political Science Review*, 102, 125–139. <https://www.jstor.org/stable/27644502>
- Myerson, R. (2010). Capitalist investment and political liberalization. *Theoretical Economics*, 5(1), 73–91. <https://econtheory.org/ojs/index.php/te/article/view/20100073/0>
- Myerson, R. (2015a). Moral hazard in high office and the dynamics of aristocracy. *Econometrica*, 83, 2083–2126. <https://www.jstor.org/stable/43866410>
- Myerson, R. (2015b). The strength of U.S. federal democracy. *Horizons: Journal of International Relations and Sustainable Development*, 5, 178–189. <https://www.jstor.org/stable/48573598>
- Myerson, R. (2021a). Village communities and global development. *Horizons: Journal of International Relations and Sustainable Development*, 18, 228–241. <https://www.jstor.org/stable/48614124>
- Myerson, R. (2021b). Local agency costs of political centralization. *Theoretical Economics*, 16(2), 425–448. <https://doi.org/10.3982/TE3763>
- Myerson, R. (2022). Local politics and democratic state-building. *Journal of Democracy*, 33(4), 62–73. <https://muse.jhu.edu/article/866642>
- Myerson, R. (2023). Focal coordination and language in human evolution. *To appear in Human Nature*. <https://doi.org/10.1007/s12110-024-09476-4>
- Newson, L., & Richerson, P. J. (2021). *A story of us: A new look at human evolution*. Oxford University Press.
- North, D. (1981). *Structure and Change in Economic History*. W. W. Norton & Co.
- O'Callaghan, J. F. (1969). The beginnings of the Cortes of León-Castile. *American Historical Review*, 74(5), 1503–1537. <https://www.jstor.org/stable/1841323>
- Oates, W. E. (2005). Towards a second-generation theory of fiscal federalism. *International Tax and Public Finance*, 12(4), 349–375. <https://doi.org/10.1007/s10797-005-1619-9>
- Olson, M. (1993). Dictatorship, democracy, and development. *American Political Science Review*, 87(3), 567–576. <https://doi.org/10.2307/2938736>
- Ormrod, W. M. (2012). *Edward III*. Yale University Press.
- Prestwich, M. (1997). *Edward I*. Yale University Press.
- Redfield, R. (1950). Maine's Ancient Law in the light of primitive societies. *Western Political Quarterly*, 3(4), 574–589. <https://doi.org/10.2307/442515>
- Richerson, P. J., & Boyd, R. (1999). Complex societies: Evolutionary origins of a crude superorganism. *Human Nature*, 10, 253–289. <https://doi.org/10.1007/s12110-999-1004-y>
- Salgado, M. (2022). Building loyalty through personal connections: Evidence from the Spanish Empire. https://drive.google.com/file/d/191o52pgX4Jeyy4hwGuu62lkMoHjDlh_y
- Sanchez de la Sierra, R. (2020). On the origins of the state: Stationary bandits and taxation in eastern Congo. *Journal of Political Economy*, 128(11), 32–74. <https://www.journals.uchicago.edu/doi/10.1086/703989>
- Schelling, T. C. (1957). Bargaining, communication, and limited war. *Journal of Conflict Resolution*, 1, 19–36. <https://www.jstor.org/stable/172548>
- Scott, J. C. (2009). *The art of not being governed: An anarchist history of upland Southeast Asia*. Yale University Press.
- Scott, J. C. (2017). *Against the grain: A deep history of the earliest states*. Yale University Press.

- Smith, A. (1776). *An inquiry into the nature and causes of the wealth of nations*. Reprinted by University of Chicago Press, 1976.
- Tiebout, C. (1956). A pure theory of local expenditures. *Journal of Political Economy*, 64, 416–424. <https://www.jstor.org/stable/1826343>
- Treisman, D. (2007). *The architecture of government: Rethinking political decentralization*. Cambridge University Press.
- Verweijen, J. (2015). *The ambiguity of militarization: The complex interaction between the Congolese armed forces and civilians in the Kivu provinces, eastern DR Congo* [Utrecht University dissertation]. <https://rug.academia.edu/JudithVerweijen/Dissertation>
- Waldman, M. (2014). *The second amendment: A biography*. Simon & Schuster.
- Xenophon, & Ambler, W. (trans.). (2001). *The education of Cyrus*. Cornell University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Authoritarian Institutions

10

Georgy Egorov and Konstantin Sonin

Contents

10.1	Introduction	210
10.2	Why Nations Fail	212
10.2.1	The Problem of Commitment	213
10.2.2	Multiple Equilibria and Path Dependence	216
10.3	Institutions of Disenfranchisement	218
10.4	Censorship, Propaganda, and Information Control	220
10.4.1	Censorship and Propaganda	221
10.4.2	The Costs of Information Control	223
10.5	Institutions of Control over Collective Action	225
10.5.1	Authoritarian Coalition Formation	226
10.5.2	Institutionalized Ruling Party	227
10.5.3	Balancing Power	228
10.6	Conclusion	229
	References	231

Abstract

Political institutions of authoritarianism are designed to ensure the elite's lasting hold of power. They might be growth-enhancing at certain stages of a country's development and become a major obstacle to economic progress in other stages, but their primary role is to keep the elite unaccountable. Thirty years ago, the economic collapse of the USSR-led communist system gave a hope for the "end of history," the convergence of all political systems to a democracy. Since then, the share of nondemocracies has stabilized after a dramatic fall; their share of the world's GDP has been increasing, largely due to China's enormous contribution.

G. Egorov (✉)

Kellogg School of Management, Northwestern University, Evanston, IL, USA
e-mail: g-egorov@kellogg.northwestern.edu

K. Sonin (✉)

Harris School of Public Policy, University of Chicago, Chicago, IL, USA
e-mail: ksonin@uchicago.edu

In this essay, we focus on institutional connections between the monopoly on political power and the monopolistic, centralized organization of economic life under authoritarianism.

Keywords

Authoritarian regimes · Dictatorship · Extractive institutions · Multiple equilibria · Information manipulation

10.1 Introduction

Authoritarianism has been the dominant form of government throughout human history. With the gradual increase of franchise over the nineteenth and twentieth centuries, democratic institutions have replaced authoritarian ones in dozens of countries. By now, in polities as different and diverse as India, the United States, Sweden, Chile, the Philippines, Switzerland, and many others, citizens have been replacing their leaders at the voting booth for decades. The end of the twentieth century saw a remarkable one-time switch from authoritarian form of government to democracy with the demise of the Soviet Union and the satellite communist dictatorships. Fukuyama (1992) famously suggested that this marks “the end of history,” the ultimate win of democracy as a form of government.

Yet authoritarian political institutions are still a widely used form of government: nearly half of the world population lives under some form of nondemocratic rule today. After 30 years of transition, many formerly communist countries are hybrid regimes or full-blown dictatorships. The phenomenon is hardly limited to new democracies: for example, Venezuela’s periods of competitive, democratic elections spanned decades, yet this did not prevent the country from sliding into autocratic rule in the twenty-first century. Largely due to the China’s enormous contribution, the authoritarian countries’ share of the world’s GDP has been constantly increasing. (See Shirley and Xu ► Chap. 35), for an extended discussion of institutional foundations of China’s economic development.) In China itself, the hopes that the pressure of the expanding economy and trade and the improving quality of life will lead to democratization have largely dissipated as the government gradually tightened its grip over the society and eliminated restrictions on the paramount leader’s power.

In this chapter, we reflect on the growing literature on authoritarian institutions. In doing so, we do not focus on those institutions that play similar roles in dictatorships and democracies. We discuss them only if there is a major difference between the function they serve under a democracy and under an autocracy. Electoral institutions are a prime example here. In a democracy, elections are a key mechanism of political selection and an instrument, with which people hold the rulers accountable. In authoritarian regimes, these institutions serve different functions: dictators use elections to gather information, discipline bureaucrats, re-align their coalitions, or project strength but *not* to give the population an opportunity to replace them. Other

institutions are specific to and characteristic of authoritarian regimes and have virtually no democratic counterparts; we discuss such institutions as well.

While we pay as much attention to empirical work on authoritarianism, there is a point to be made about theoretical models. Sometimes, formal models of politics seem to overemphasize the role of significant individuals as they invariably feature a small number of actors set to play a game, a now-standard way to model strategic interactions in economics and political science. When modeling nondemocratic politics, there is a special temptation to focus on few individual decision-makers: many dictatorships are organized, either formally or informally, in a way that the individual at the top or a small junta has a lot of power. The analytical convenience of using game-theoretic apparatus should not be confused with a description of an actual dictatorship. The institutional structure of a dictatorship is as much represented by a game of two players, a dictator and a subordinate, as the structure of a large firm is represented by a game of a principal and an agent. In other words, game-theoretic models are merely instruments to study what matters in non-democratic politics: institutions, organization, networks, social relations, etc.

To make the discussion more systematic, we make a distinction between two types of authoritarian institutions. Some institutions are designed and built by dictators in order to survive challenges, maintain control, or enrich themselves. Other institutions are part of the autocratic environment: they describe the rules of the game where dictators—actual and potential and past, present, and future—are players, not designers. Institutions that limit citizens' participation in the government—repression, disenfranchisement, censorship, etc.—is an example of the former. They reduce the government accountability and allow the regime to pursue its preferred, unpopular (or, at least, lacking a majority support) policy. If the dictator cannot repress the opposition and make elections noncompetitive, he would have to either distribute resources or make policy concessions. Instead, authoritarian regimes build up institutions to carry out repressions and to effectively disenfranchise the majority of the population (see Sect. 10.3).

The other group of authoritarian institutions, designed by autocrats, is primarily concerned with information manipulation (Sect. 10.4). Dictators censor news and feed citizens with propaganda. At the same time, a major prerequisite for efficient governance and ultimately for the dictator's political survival is his ability to gather and process information. To keep power, even the most sultanistic of dictators need to understand the ever-changing needs of the people they govern. This makes it even more surprising that, almost as a rule, dictators end up in an informational vacuum, surrounded by loyal yet incompetent subordinates. Organizations that are responsible for gathering, screening, aggregating, and processing information in democratic systems—various ministries, committees, advisory bodies, intelligence services, and military units—fail to do their job in authoritarian regimes. In fact, they are often structured to do the opposite.

For an individual autocrat, the key element of survival in power is building a support coalition. One way to sustain the power of leaders is the establishment of an

institutionalized ruling party such as the Communist Party of the Soviet Union, the Chinese Communist Party, or the Institutional Revolutionary Party of Mexico, each of which was the primary governing mechanism in their country for the large part of the twentieth century. Absent a well-institutionalized government structure, the authoritarian equilibrium might feature a balance between different factions that are unable to defeat each other (Newson & Trebbi, 2018).

Applying Douglass North's definition of institutions as "the rules of the game in a society, or more formally, the humanly devised constraints that shape human interaction" (North, 1990), we also discuss those authoritarian institutions that are not a result of elite's optimizing designs but rather institutions that are a result of their interaction with their competitors in the power struggle. To take an extreme example, if the current way of dealing with ousted leaders is to execute them, and new leaders follow this pattern, this is certainly not an institution that the incumbent has designed. However, it is "humanly devised constraints that shape human interaction."

This chapter does not aim to be an all-encompassing survey of nondemocratic politics. Gehlbach et al. (2016) is a survey of formal models in nondemocratic politics with emphasis on the critical role that institutions play. For a survey on long-term institutional change, which necessarily deals with long-term institutions of nondemocracies, we refer to Acemoglu et al. (2020). In Egorov and Sonin (2024), we paid special attention to dictators' survival strategies such as purges, election result falsification, censorship, propaganda, etc. and citizens' responses such as protests and revolutions.

The rest of this essay is organized as follows. In Sect. 10.2, we discuss the current consensus model that relates economic and political institutions and development. In Sect. 10.3, we discuss political *institutions of disenfranchisement* that autocrats construct to prevent accountability by citizens. Section 10.4 deals with institutions of information control, while Sect. 10.5 focuses on institutions that prevent collective action such as inside coups, mass protests, and revolutions.

10.2 Why Nations Fail

From an institutionalist perspective, the field of political economics is currently close to a consensus view that secure property rights and democracy are critical for development. The general approach, which builds upon foundation work in the New Institutional Economics, is best expressed in Acemoglu and Robinson (2005), a graduate-level textbook by Daron Acemoglu and James Robinson, and widely popularized in their *Why Nations Fail* magnum opus (Acemoglu & Robinson, 2012). It makes good institutions of property rights protections the critical element of development on the economic side; the effective access to vote is the critical element on the political side.

The main concepts that Acemoglu and Robinson use to explain economic development are that of *inclusive institutions* that allow for broad participation in value-enhancing economic activity and provide incentives for creative destruction

and development and *extractive institutions* which hinder economic growth. Inclusive institutions allow citizens to participate in the shared prosperity. Inclusive political institutions allow citizens to play a role in creating rules of the economic game. Democracy often, though not always historically, provides the political foundation for inclusive economic institutions. In contrast, extractive institutions benefit a small elite-holding monopoly over political power. The elite is interested in protecting its political power as political power allows them to extract disproportionate rents from economic activity via monopolies protected by law, tradition, or force. By its nature, extractive political institutions are authoritarian.

The extractive vs. inclusive institutions model, a synthesis of decades of research by numerous economists and political scientists, is deeply rooted in the New Institutional Economics. The ideas of multiple equilibria and path dependence were formulated by Douglass North in North (1981). The theory of social orders in North et al. (2009) is, in many respects, another version of the inclusive vs. extractive institutions theory. North, Wallis, and Weingast define a set of institutions to be a *limited access order* if they serve the purpose of keeping the elite's control over political and economic rents, which closely corresponds to the Acemoglu and Robinson's extractive institutions. Both big theories are results of decades of research on the role of commitment and institutions of property rights protection initiated in North and Weingast (1989).

Below, we start with the credible commitment paradigm in analysis of institutions. The current consensus, building on work of North, Weingast, Acemoglu, Robinson, and others, considers institutions a product of political process in the society. Even if the process is very long term, it is ultimately a result of choices made by rational agents at some point in time (Acemoglu et al., 2011). Identifying an inability to commit to a certain course of actions and analyzing institutions that compensated for this inability has become a standard approach for institutional economists over the last decades. Then, we turn to the concepts of multiple equilibria and path dependence, which are critical to explain why extractive (and so authoritarian) institutions are persistent.

10.2.1 The Problem of Commitment

Over the last 40 years, the *credible commitment paradigm* proved immensely fruitful in many areas of economics. Institutionalist Yoram Barzel highlighted the lack of commitment as a central problem in his analysis of slavery (Barzel, 1977). Macroeconomists Finn Kydland and Edward Prescott used it in their groundbreaking discussion of rules vs. discretion in monetary policy (Kydland & Prescott, 1977). Scholar of socialism János Kornai focused on the inability of a socialist government to commit, which he dubbed *the soft budget constraint*, to explain the ills of planned economies (Kornai, 1986). Working on a theory of the firm, Oliver Williamson invoked the Thomas Shelling's metaphor of credible commitment to differentiate between types of investment (Williamson, 1983). Douglass North and Barry Weingast analyzed the consequences of the Glorious Revolution, opening up a

whole new world of applications (North & Weingast, 1989). Political scientist Kenneth Shepsle applied the commitment approach to analysis of government structures (Shepsle, 1991). In his studies of conflict, James Fearon identified the countries' inability to commit to a certain course of actions as one of the major causes of war (Fearon, 1995).

Yet nowhere the commitment paradigm was as fruitful as in explaining why some institutions have formed in the ways they have formed. In North and Weingast (1989), the transition of the authority over the standing army from the crown to the parliament was the way for the new sovereign to resolve the commitment problem and gain the necessary support. This in turn led to secure property rights for would-be entrepreneurs and, eventually, laid institutional foundations for the Industrial Revolution. In Acemoglu and Robinson (2000), the inability of the elite to commit to future payments to the commons is resolved by the extension of the franchise. The rich elite agrees to establishing democratic institutions, because otherwise it cannot commit not to exploit the opportunities to expropriate the poor. By similar logic, in Acemoglu and Robinson (2001), the inability of the rich to commit to redistribution and the inability of the poor to commit not to revolt lead to full democratization in some circumstances. In other circumstances, it is the inability of a democratic government to commit not to increase taxes results in the rich elite launching a coup.

For authoritarian regimes, the problem of commitment is exacerbated because authoritarian leaders, by definition, face less constraints on their power. With less constraints, they are bound to have more severe commitment problems than their democratic counterparts. For example, a US president has not, with a few exceptions, had problems committing to leave their post after their two terms. With less mature institutions, a leader has incentives to stay in power inefficiently long. Of course, no ruler has absolute power—and dictators sometimes operate under stringent constraints—yet consolidated authoritarian regimes typically have less checks and balances than democratic ones. Thus, the commitment problem that they have manifests itself in what institutions they construct. We will see this below discussing the succession problem, nearly unsolvable in a personalistic dictatorship. However, perhaps the most significant consequence of the inability of dictators to commit is their inability to make partial concessions or pursue gradual democratization.

There are many reasons why an authoritarian regime might turn to democratization. (See ► Chap. 11 on regime change.) Democratizations themselves do produce growth benefits in both short (Rodrik & Wacziarg, 2005) and long runs (Papaioannou & Siourounis, 2008). Even if he does not care about broad prosperity, a dictator might want to democratize to avoid facing mass protests or internal coups, which are costly even when unsuccessful (Balima, 2020). Partial democratization might be desirable if it provides a commitment device to protect property rights and thus improves agents' incentives to involve in productive activities (Acemoglu & Robinson, 2005; Myerson, 2010).

Naturally, a dictator might not relinquish power, even partially, unless he is pressured to do so. Yet why do so many dictators fail to democratize even when the pressure is very strong? The commitment perspective offers an answer in a

general model of strategic enfranchisement and disenfranchisement with forward-looking agents (Acemoglu et al., 2012, 2015). In the simplest possible version, the current decision-maker, for example, a unitary dictator or the median voter of the enfranchised coalition, makes two decisions. First, she chooses this-period policy. Second, she determines who is going to be the decision-maker in the next period. Suppose that someone else's rule in the future is preferred by the current decision-maker.¹ This is a major general rationale for enfranchisement: for example, with a dictator in power, the middle class fears expropriation, thus providing low efforts and generating small surplus. When enfranchised, the middle class has its property rights protected better and exerts high efforts. Thus, the dictator would prefer to transfer power (to democratize) and benefit from citizens having secure property rights in the future. However, the dictator might dislike the consequences of the political choices by those whom she transferred power to. As a result, the country is stuck with an inefficient autocratic rule despite the fact that the autocrat herself would prefer partial democratization. In other words, a rational dictator would not want to partially relinquish her power, fearing that this will lead to her ouster through the "slippery slope" (Schwarz & Sonin, 2008). The prospects of slippery slope might be highly uncertain, but the fear is certainly real. With forward-looking decision-makers, the country is stuck in a bad equilibrium.

The dictator's unwillingness or inability to have committed to a certain course of actions might result in a different kind of an inefficient outcome. Greif (1998) analyzes *podesteria*, the institute of delegation of power by a group of competing oligarchs in medieval Venice, as an instrument to resolve their common pool problem. Without an arbitrator, the oligarchs would spend an inefficiently large amount of resources on competing against each other. Focusing on the experience of transition from socialism to market, Guriev and Sonin (2009) argue that even if oligarchs are interested in appointing a leader who can contain destructive rent-seeking, they would opt for a weak dictator fearing that a strong one will end up expropriating their property. Thus, an efficient outcome is not achieved. In a democracy, institutions of reputation, political parties, etc. make the problem far less severe.

As our discussion of the commitment problem above emphasized, dictators avoid creating institutions that constraint their own freedom of choice. In a democracy, the new leader has limited or no power over the fate of the predecessor. In an autocracy, a successor might have a lot of power over the former dictator's fate. Therefore, the loyalty of a possible successor is critically important for autocrat. At the same time, not having a successor has always been considered a destabilizing factor, threatening the incumbent regime. (See Kurrild-Klitgaard (2000) and Kokkonen and Sundell (2014), for empirical evidence on advantages of European monarchies in solving the succession problem.) Not surprisingly, few dictators have truly solved the succession problem (Bueno de Mesquita & Smith, 2017; Herz, 1952; Konrad & Mui, 2017).

¹Lizzeri and Persico (2004) apply this logic in a model of franchise extension to discuss the evolution of public spending.

In the twenty first, the aging leaders of Egypt, Libya, Malaysia, Tunisia, Turkmenistan, Uzbekistan, Yemen, Zimbabwe, and other countries for years failed to delegate any power to designated successors; ultimately, their succession plans failed.

One governing mechanism that does solve the succession problem is an *institutionalized ruling party*, which we discuss below in Sect. 10.5. With a party calling the shots, both the commitment and the loyalty problems become less acute for the incumbent. Authoritarian parties that were able to ensure regular rotation at the top such as Mexico's PRI in 1930–1994 or the Chinese Communist Party in 1976–2019 were able to deal with the succession problem.

10.2.2 Multiple Equilibria and Path Dependence

Douglass North made the concept of multiple equilibria central in the institutional theory of economic development. The idea is straightforward: it is possible to have very different sets of institutions, created by the people, in otherwise identical countries. Over time periods that span decades of even centuries, there is no visible convergence of institutions. Countries could have the same climate, the same reserves of natural resources, the same neighbors and international ties, the same history prior to the bifurcation point, and the same anything which was already baked in before the paths have diverged and yet end up living with different institutions and, as a result, be on different paths of development.² The development of this theory was warranted as economists have repeatedly hit dead-ends in their quest for determinants of growth failures.

If two equilibria with the same fundamentals are possible, and one equilibrium is vastly preferable to the other—as it is, for example, in Murphy et al. (1993)—the natural question is why would not the country move from the bad equilibrium to a good one? The theory of inclusive vs. extractive institutions provides an answer: the extractive institutions, often very inefficient, are a mechanism of protecting economic rents of the narrow elite. Authoritarian political institutions are, in short, the political institutions that allow the elites to keep political power, which in turn allows them to extract rents. In other words, the main purpose of authoritarian institutions is to perpetuate the political status quo, the rule by a narrow elite.

Consider the following highly stylized example. Suppose that in some country economic agents have a chance to invest in productive activities but also need to spend resources on private protection of their property rights. The resources that are spent on protection—building private armies or corrupting public officials—could also be used to prey on others' productive capital. In Sonin (2003), there might be multiple equilibria: in one equilibrium, with low productive investment and a fast growth rate, agents spend a lot of resources on unproductive protection and

²North's argument was developed in numerous works; Murphy et al. (1993) is perhaps the simplest possible formal model. Acemoglu et al. (2020) discuss the necessity of multiple-equilibria approach in studies of institutional development.

redistribution from each other. In another, “good,” equilibrium, there is little unproductive investment in redistribution, and the growth rate is high. The wealthiest members of these societies are the primary beneficiaries of private protection of property rights as maintaining a private army or political connections have, naturally, high returns to scale. So the rich are better off in the bad equilibrium than in the good one. Thus, in contrast to the naive intuition and in alignment with the logic of the extractive institutions, it is the rich who are the source of political demand for bad institutions—they are interested in poor *public* protection of property rights.³

The concept of *path dependence* is intimately linked to the concept of multiple equilibria. If a nation stays in a bad equilibrium, while a good equilibrium potentially exists as in the model described above, why would not it move to a good equilibrium? (In a simplistic setting of Murphy et al. (1993), *every* citizen would benefit from moving to a good equilibrium.) Path dependence refers broadly to situations, in which a move to a good equilibrium is constrained by the past choices on the country’s trajectory. In many models, the stability of a bad equilibrium hinges upon the assumption that switching to a better equilibrium requires a coordinated action of a mass of citizens.

Path dependence does not need to rely on coordination failure of a large number of agents. Let us consider a game in which a succession of potential challengers fight for power against the incumbent dictator. After the fight, the winner of the fight decides whether or not to repress the loser. We allow players’ strategies—what do to with losers of the power fight—to depend on the “reputation,” the number of killings that have been ordered by the loser in question during her tenure in power.⁴

The welfare-maximizing equilibrium in this game is such that each winner spares the loser. There is no repression on the equilibrium path: leaders replace each other and exit the game only because of a natural death. However, there is another, “killing” equilibrium in this game, in which the winner always represses the loser. The mechanism at work is as follows. If the current dictator executed her predecessor, then the challenger has stronger incentives to execute the overthrown incumbent fearing that, if not executed, he will be a challenge in the future, and, if he returns back to power, would repress those whom he will have defeated. This results in path-dependence of repressive regimes: the winner of a power fight values his options differently depending on the type of his fallen enemy. If somebody takes over from a bloody dictator, he is more likely to become a bloody dictator himself than if he

³The first model with the rich interested in poor protection of property rights is Polishchuk and Savvateev (2004). The reason is that poor public protection of property rights make it easier to benefit from rent-seeking, which exhibits relatively high return to scale; thus, the rich prefer poor protection of property rights (see also Hoff & Stiglitz, 2004). Sonin (2003) extends this logic to a dynamic setting and a range of voting rules over institutions of property rights protection and demonstrates the possibility of multiple long-term equilibria.

⁴The “killing game” in Egorov and Sonin (2015) is a complete information game. The “reputation” is understood as a characteristic of an equilibrium strategy; there is no learning along the way. It is straightforward to extend the logic to a model of commitment-type-based reputation in an imperfect information game (Kreps et al., 1982).

comes to power after a natural death of the previous ruler. Thus, the new ruler cannot switch to another equilibrium path, even though he knows that he would be better off in a “peaceful” equilibrium path. Killing predecessors becomes, effectively, an institution.

The bad path-dependent equilibrium described above relies, ultimately, on the commitment problem. In many circumstances, the loser of an authoritarian power fight might be willing to commit not to be a contender in the future as such commitment would spare his life. Such commitment might be impossible for a dynastic ruler, whose rights to contend the throne are “divine” and thus virtually indispensable. Not surprisingly, countries with a limited and clear-cut set of contenders such as dynastic monarchies are more likely to have repressions of predecessors than, say, military dictatorships. Egorov and Sonin (2023) use this framework to explain why many personalistic dictatorships, even if highly institutionalized, end up in a vicious cycle, in which the leader ends up surrounded by incompetent loyalists as he fears betrayal and accelerates repressions to compensate for poor policy choices.

10.3 Institutions of Disenfranchisement

By definition, a nondemocratic government assumes that many citizens are excluded from having a say in political decisions. Dictators of the twentieth century, such as Hitler, Stalin, and Mao, among others, as well as many lesser-known leaders, purged their political supporters and repressed millions who did not support their policies. In extreme cases like the totalitarian dictatorships of Stalin’s Soviet Union or Mao Zedong’s China, the entire population was effectively disenfranchised, having no say in policy choices or the selection of leaders. Many other dictatorships disenfranchised or purged entire social groups based on factors like wealth status, religious affiliation, or ethnicity (Egorov & Sonin, 2015).

The key argument for why disenfranchisement benefits an autocratic leader comes from the standard spatial model of politics. In this model, candidates must commit to the median voter position as a policy platform to have any chance of winning. While it might not be the median voter playing the pivotal role in an authoritarian power contest, the general logic applies. The larger the share of the population that supports the dictator’s chosen policy over the opponent’s alternative, the easier it is to stay in power. To make his hold on power stronger, the dictator does not need to shift their policy platform toward the median the way a democratic politician would. Instead, the dictator might “trim the electorate,” making the median of the remaining population closer to the dictator’s own preferred policy. Repressing those who are most opposed to the dictator’s preferred policy—whether through imprisonment, violence, or disenfranchisement—makes it easier to implement the policy (Gregory et al., 2011).

Early economic theories of nondemocratic government focused on a simple trade-off: how to optimally allocate resources between repression and benefits (Wintrobe, 1990, 1998). Modern theories of repression assume strategic targeting and selection.

For example, Myerson (2015) suggests that the best incentives for the autocrat's supporters are provided via randomized purges, combined with rewards for service. Tyson (2018) and Dragu and Przeworski (2019) combine an agency model of dictatorship with targeted repressions. Montagnes and Wolton (2019) and Rozenas (2020) use communist purges in Stalin's Russia and Mao's China to illustrate the effects of violence on performance and the selection of subordinates.

Regarding propaganda as a substitute for repression, Guriev and Treisman (2019) argue that the dictator uses either repression or persuasion for different social groups. Gitmez and Sonin (2023) suggest that repression and informational control could complement each other, with repression of skeptics allowing for increased manipulation of information for others. Initially, propaganda slant is limited by the incentive constraints of skeptics, but when the skeptics are repressed, the incentive constraint is relaxed, and the rest of the population receives more proregime information. In George Orwell's *1984*, the people of Oceania are forced to use the *newspeak*, a special language designed to limit their ability to articulate antigovernment concepts, cannot switch off radio that translates propaganda, and are forced to participate in indoctrination meetings, yet what actually makes them love Big Brother is the physical torture.

Another form of effective disenfranchisement is electoral fraud. Stripping citizens from the opportunity to replace the incumbent at the ballot box is in the dictator's self-interest. In fact, it is a question why authoritarian regimes have elections at all. In a democracy, electoral fraud might tip the outcome of a close election. If citizens are unaware of the fraud or consider its extent negligible, they may perceive the outcome as legitimate. Therefore, if the fraud remains undetected, it can influence the outcome. In nondemocracies, electoral fraud is often extensive, and violations of due procedure are so numerous and transparent that it begs a question: why is it rational to hold fraudulent elections that citizens easily identify as such? Why organize such elections at all?

There is a substantial literature in political science that strives to explain elections held by autocrats. Przeworski (2009) describes "plebiscitary elections," which the regime uses to demonstrate that it can "force everyone to appear in a particular place on a particular day and perform the act of throwing a piece of paper into a designated box." Egorov and Sonin (2021) model dictators "projecting strength" by organizing authoritarian elections to prevent an insurrection. Though citizens know that elections are manipulated by the incumbent and realize that the reported outcome exaggerates the incumbent's true popularity, they still interpret his overwhelming "victory" as a sign of genuine support. (See Simpson (2013), for empirical evidence on fraudulent elections from all over the world.)

Another explanation why autocrats organize elections is that they use them to gather information. Even if the dictator controls the voting process and has means to manipulate the announced final tally, she might be able to gauge the extent of citizens' dissatisfaction or to learn about their specific concerns. Martinez-Bravo et al. (2022) argue that local (village-level) elections in China inform the central authorities about the peasants' concerns. Using a data base on world-leaders' exits from 1975 to 2004, Cox (2009) argues that elections help autocratic regimes to

gather information that is needed to optimize succession. Miller (2015) finds that a negative shock to the election results prompts autocracies to spend more on education and social welfare. In Little et al. (2015), the results of an election convey the same information to the dictator and the citizens, yet the dictator uses this information to decide whether or not to step down voluntarily.

Authoritarian elections might play yet another role—they may define and enforce power-sharing or rent-sharing agreements among the elites (Boix & Svolik, 2013; Gandhi & Przeworski, 2007; Magaloni, 2010). In some circumstances, such agreements might help the authoritarian elite to survive a transition to competitive elections (Martinez-Bravo et al., 2017). In Gehlbach and Simpson (2015), electoral fraud, modeled as Bayesian persuasion, is the instrument the dictator uses to manipulate bureaucrats' incentives. This makes efforts of different bureaucrats complementary—they participate in fraud if they expect others to participate—and results in a high variance of the winning (or losing) margins. Finally, Wig and Rod (2016) document the risks associated with authoritarian elections: if the election outcome points to a possible revolution, it might trigger a preemptive coup.

Empirical literature on electoral fraud is growing fast, though most of the studies are focused on imperfect democracies rather than authoritarian regimes. Enikolopov et al. (2013) used a field experiment to estimate the extent of electoral fraud in the Russian parliamentary elections in 2011. Rundlett and Svolik (2016) used data from the same election to illustrate the mechanism of beliefs formation when citizens observe electoral fraud. Cantu (2019) described mechanisms of electoral fraud in a one-party dictatorship employing data on the Mexican presidential elections in 1988.

10.4 Censorship, Propaganda, and Information Control

Understanding the critical role of information in authoritarian politics was made possible by the development of modern contract theory and principal-agent models. Information control is critical to authoritarian government as any collective action of multiple agents requires them to coordinate. Individually, participating in a revolution is costly. However, when a mass of citizens participate, the costs are substantially lower and the chances of success are higher. Thus, *informational control* is a major part of any dictator's survival strategy. Although informational control has arguably played as great a role for first-century emperors or twentieth-century caudillos as for modern dictators, it was not conceptualized as such. Throughout centuries, authoritarian leaders used censorship, propaganda, and other forms of information manipulation to limit the reach of their competitors and mobilize the population. Still, histories and biographies that would analyze information management as a strategic tool that the political economists now consider one of the central mechanisms of maintaining control over the society are yet to be written.

Any authoritarian leader faces at least two major problems related to information flows. First, she has to design the optimal information structure that would determine how much and what kind of information should be available to citizens. Second, she needs to organize the information flow in the other direction so that she can gather

information, be it by allowing a certain degree of media freedom, relying on secret surveys and opinion polls, or inferring the public mood via organizing elections that she fully controls.

10.4.1 Censorship and Propaganda

We start with the simplest possible observation about information manipulation. An authoritarian leader chooses an information structure, under which information is released to citizens whose interests are not necessarily the same as that of the autocrat. Manipulating information in an optimal, from the dictator's standpoint, way requires providing truthful information with some positive probability. In other words, optimal propaganda is necessarily a mix of some truthful information with some information favorable to the dictator.

The first model of government censorship of media as Bayesian persuasion appeared in Gehlbach and Sonin (2008), subsequently published as Gehlbach and Sonin (2014). (See Bergemann and Morris (2019), for the general framework of Bayesian persuasion.) Little (2023) discusses theoretical models of information manipulation based on other communication protocols such as no-commitment persuasion in Crawford and Sobel (1982), verifiable messaging in Milgrom (1981), and signaling in Spence (1973). Across all communication protocols, Bayesian persuasion results in the maximum impact of information manipulation (Kamenica & Gentzkow, 2011). At the same time, using Bayesian persuasion to model information manipulation by the regime comes at a cost as it requires to assume a significant commitment power on the sender's part. In some circumstances, this assumption is too strong to be realistic: a dictator's ability to commit to a certain signal structure might be limited compared to that of a democratic leader.

The theoretical observation that any targeted information manipulation should contain some truthful information has tangible empirical implications. Even the most authoritarian regime needs to structure institutions that govern media freedom, censor information, punish violators, etc. in a way that give the information that citizens' receive some credibility. The propaganda machine in the George Orwell seminal novel *1984* did not satisfy this criteria: it has always insisted on the party line. A more credible authoritarian propaganda in USSR and other communist dictatorships last century or in Russia and China today has indeed contained a mix of truthful and manipulated information.

Naturally, the optimal slant in the propaganda messages decreases with the opportunity cost of getting information: if citizens do not pay attention to what the government says, the latter survives for a narrower range of parameters. Knight and Tribin (2018) demonstrate, in the context of Venezuelan dictatorship, that availability of propaganda-free channels reduced the impact of state propaganda. Glaessel and Paula (2020) confirm that access to alternative sources of information limits propaganda with 1989 data on German Democratic Republic television: recipients disapproved of censorship if they were able to use Western television to detect misinformation. Still, access to information is not sufficient unless citizens have

incentives to consume information. Chen and Yang (2019) conducted a field experiment in China that provided citizens' access to an uncensored internet. In a nutshell, the finding is that propaganda works: once citizens acquire new information, their knowledge, beliefs, attitudes, and intended behaviors change. At the same time, the demand for uncensored information is low; on their own, citizens do not look for additional information even if they have access to it.

The informational theory of modern dictatorships is growing fast. In Shadmehr and Bernhardt (2015), the state does not censor moderately bad news to prevent citizens from making inferences from the absence of news that the news could have been far worse. In different circumstances, the autocrat might prefer transparency to reduce the risk of an inside challenge (Hollyer et al., 2018; Kosterina, 2017) or to strengthen a power-sharing agreement (Sheen et al., 2022). In Boleslavsky et al. (2021), the autocrat prefers transparency as it helps to mobilize citizens to protect him in the case of an inside coup.

Guriev and Treisman (2022) offer a sweeping theory of *spin dictators*, a modern form of authoritarianism, in which information manipulation replaces cruel repression as the main instrument of authoritarian control. Instead of glorifying the leader as the Stalin's or Hitler's cults did, modern dictators present themselves as steady and competent stewards of national prosperity. Subtle censorship and targeted propaganda make repression largely unnecessary. The book provides systematic evidence that violence is much less common in modern dictatorships than in dictatorships of the past.

Relying on extensive archival data and modern methods of econometric inference, Adena et al. (2015) delved into the impact of Nazi propaganda. They confirmed a significant effect of radio propaganda postconsolidation of Hitler's regime: it incited anti-Semitic acts and led ordinary citizens to denounce Jews to authorities. Interestingly, the propaganda sometimes had unexpected outcomes: in areas with positive attitudes toward Jews, the anti-Semitic propaganda paradoxically strengthened these favorable views. Such propaganda could serve multiple purposes, targeting potential opponents or fueling totalitarian mobilization (Arendt, 1951). (Zhuravskaya et al. (2020) provide a comprehensive survey of the empirical literature on the political effects of the Internet and social media.)

Empirical research on the role of propaganda in authoritarian regimes is rapidly expanding. (See Rosenfeld and Wallace (2024) for a recent survey of empirical work.) Qin et al. (2018), utilizing data on government-owned newspapers in China from 1981 to 2011, demonstrated that market competition reduced political propaganda, confirming theoretical predictions in Gehlbach and Sonin (2014). Pioneering empirical analyses by King et al. (2013) and King et al. (2014) unpacked the strategic use of censorship by the Chinese government. They revealed that the government avoided censoring criticism but blocked information facilitating citizen coordination on grievances. Using daily news reports from Russia's largest state-owned television network, Rozenas and Stukal (2019) discovered that instead of censoring economic facts, the media attributed positive news to government officials' competence while attributing negative news to external factors.

10.4.2 The Costs of Information Control

Many decisions by authoritarian leaders seem to be appallingly misguided in retrospect. Decisions by Argentina's military junta led by General Leopoldo Galtieri to seize Falkland Islands, communist Romania's chairman Nicolae Ceausescu to call a mass meeting in December 1989 and, before that, to launch unrealistically ambitious economic reforms, Iraq's dictator Saddam Hussein to annex Kuwait in 1991 each caused a chain of events that led to their regimes' demises.⁵ In each of these cases, the dictators failed to be well-informed about political developments of the first-order importance to them. Ceausescu, Mubarak, or Galtieri's failure to be properly informed is a consequence of the dictator's *informational dilemma*: to govern, any leader relies on institutions that collect, aggregate, and process information about the state of affairs and the attitudes of his subjects. If information is collected and aggregated by free media or some other sources independent from the dictator, then citizens might learn about leader's corruption or incompetence. Furthermore, they might learn about other citizens' attitudes and willingness to challenge the leader. On the other hand, if the dictator controls the sources of information, then the information these sources produce might not be reliable.

There are two very distinct broad types of institutions that autocrats use to collect information. First, the government can use public sources of information, relying on strong institutions that protect media freedom, freedom of speech and assembly, etc. Alternatively, the government can rely primarily on information gathered by various secret services or new-era methods such as digital surveillance (Xu, 2021). With the former, the downside for the autocrat is that free media provides the same information to citizens, which reduces the impact of propaganda and might help to facilitate antigovernment protest. With the latter, the problem is that relying on information from secret services is, effectively, sharing power and rents with them.

As we discussed above, censoring information flows increases the chances of an autocrat's survival. However, there might be efficiency costs associated with restrictions on media freedom. In Egorov et al. (2009), a resource-poor dictator allows media freedom as he is concerned with providing his bureaucrats with proper incentives. The paper confirms empirically the relationship between oil wealth and media freedom: in dictatorships, more oil means less media freedom, whereas in democracies the effect disappears. An oil-rich dictator can afford to stay out of touch with reality, censor media and yet stay in power; an oil-poor dictator does not have this luxury. Similar efficiency vs. propaganda trade-offs appear in the model of strategic protest restrictions (Lorentzen, 2013) and censorship (Lorentzen, 2014). As a result, media freedom varies a lot across nondemocratic regimes, from levels comparable to mature democracies to that of totalitarian regimes.

The most straightforward way to present the dictator's informational dilemma in a simple theoretical model is to focus on the loyalty vs. competence trade-off that each

⁵Egorov and Sonin (2023) focus on Russia's dictator Vladimir Putin's decision to invade Ukraine in February 2022 as another example.

dictator faces (Egorov & Sonin, 2011; Jia et al., 2015; Zakharov, 2016). A dictator needs a competent agent that controls the information flow. A more competent subordinate is more likely to differentiate a real threat to the dictator's power such as a change in public mood, a need for an economic reform, an emergence of a new opposition leader, or even a looming foreign invasion. At the same time, a more competent subordinate is more likely to side with the dictator's enemies when the dictator is vulnerable, that is, exactly when the subordinate's loyalty is critical. An insecure or cautious dictator will therefore choose incompetent loyalists as ministers because he fears that a competent minister will betray him more easily than an incapable one. In other words, a rational dictator ends up in an informational bubble of her own choice: though the dictator is free to choose an agent of any competence in this model, the choice is incompetent.

The informational model of loyalty and competence provides a partial answer to the following puzzle. It is only natural that countries that have frequent coups and recurrent revolutions exhibit poor economic performance. Yet why do countries that have the same leaders for decades provide a disproportional number of growth failures (Besley & Kudamatsu, 2009; Bueno de Mesquita et al., 2003)? Jones and Olken (2005) demonstrated, using unexpected deaths of leaders in office as a source of exogenous variation in leadership, that leaders matter for growth, and negative effects of individual leaders are strongest for unconstrained autocrats.⁶ Mançur Olson has offered a powerful metaphor of "roving vs. stationary bandit" (Olson, 1991), further developed in the concept of "encompassing interest" (McGuire & Olson, 1996). The power of this metaphor is in its consistency with a basic premise of economics: incentives matter. The higher the leader's stake in the country's welfare and the longer his horizon, the more interested she should be in the country's prosperity. A major problem with the Olson "stationary bandit" metaphor is that it contradicts the accumulated empirical evidence on modern dictatorships. The dictators who have had the most power and, in Olson's model, the highest stake in their countries—Germany's Hitler, Russia's Stalin, China's Mao—have led their countries to massive humanitarian disasters and destruction of social welfare on a historic scale. In general, the longer a dictator's tenure is, which, presumably, means a higher stake in the country's fate, the lower were economic and societal benefits of his rule (Bueno de Mesquita et al., 2003). The loyalty vs. competence model demonstrates, theoretically, that stagnation might be a likely flip-side of the prolonged political stability under dictatorship. Loyalty at the expense of competence might be in part responsible for massive disasters such as the Great Famine in China (Meng et al., 2015) or Holodomor in the Soviet Union (Naumenko, 2021). In both cases, it was gross mismanagement, coupled with dictator's disregard to human life and welfare, that caused deaths of millions. Inefficient government is the flip side of the strong authoritarian control.

⁶Easterly and Pennings (2017) replicated, using an expanded data set, the Jones and Olken's results with respect to very high or very low growth episodes; they also confirmed that autocracies produce higher growth volatility than democracies as suggested by Rodrik (2000) and Besley and Kudamatsu (2009).

Bai and Zhou (2019) confirmed the existence of the loyalty vs. competence trade-off using data on the Cultural Revolution in China (1966–1976), Mao’s successful attempt to renew his power coalition after the disaster of the Great Leap Forward. As it turned out, it was the most competent elite members who were purged and replaced by mediocre substitutes. Shih (2022) coined the term “coalition of the weak” for the appointment strategy Mao Zedong pursued in his last years. Not coincidentally, the Mao’s years in power saw economic and societal stagnation, accompanied by renewed elite purges and mass repression. Focusing on the post-Mao period, Jia et al. (2015) demonstrated that the Chinese Communist Party has avoided the “loyalty vs. competence” trap through a system of job rotation and promotion within the party. In contrast, Shih et al. (2012) provide evidence that nepotism plays a major role in promotions within the highest tier of the CCP hierarchy. Analyzing the database of 12,000 appointments to the People’s Liberation Army of China, Mattingly (2024) concludes that when an autocratic regimes faces a shift from concerns over the foreign threats to concerns about domestic threats, the balance of military appointments shifts toward appointment of politically loyal, rather than combat-experienced officers.

10.5 Institutions of Control over Collective Action

Since the classic work of Mancur Olson (Olson, 1965), preventing collective action by potential opposition has been well understood as an important objective of authoritarian leaders. To control elite opposition, dictators employ both sticks, such as purging elites and punishing dissenters, and carrots, by making potential challengers beneficiaries of the status quo. For nonelites, repression and information manipulation are common tactics. While dictators may also “buy” support from nonelites, it is important to note that they maintain their subjects in a bad equilibrium. Thus, any benefits provided are arguably less than what citizens would receive in a good equilibrium. When people attribute their well-being to the dictator’s actions, it is often a result of misinformation or information manipulation.

The basic models of political dynamics assumed away the collective action problem. In Acemoglu and Robinson (2001, 2005), potential dissidents under the elite-controlled regime (“the poor”) are able to overcome, from time to time, the collective action problem and coordinate on protests. In other circumstances, “the rich” would overcome the collective action problem and launch a coup.⁷ As a next step, Ellis and Fender (2010) added a model of mass protests as information cascades (Lohmann, 1994) to the Acemoglu and Robinson’s framework. Shadmehr and Bernhardt (2011), in a two-person coordination game, demonstrated that limiting public information available to citizens might increase the likelihood of protests as each individual citizen is forced to rely on others’ information to a larger extent. In

⁷Tullock (1971) argued that the coups d’état are more common in nondemocracies than revolutions because the collective action problem is much less acute in the case of a coup.

an important theoretical contribution, Edmond (2013), the dictator, has a costly technology to jam the signal available to citizens who might otherwise coordinate to protest.

Even if the majority of citizens desire the removal of the dictator, the absence of free elections can make it challenging for them to organize and coordinate protests and revolutions. One crucial aspect of collective action is the aggregation of information. For a citizen who seeks the leader's removal, it is vital to understand how many others share the same preferences, what information they possess, and what actions they plan to undertake.

The problem of collective action remains difficult to resolve even if the dictator does not directly manipulate people's beliefs. In protests and revolutions, citizens must be aware of what others intend to do: an individual's payoff depends critically on the actions of others. In such scenarios, controlling information regarding collective action involves manipulating the process of information exchange rather than providing manipulated information or censoring media. Furthermore, even in the absence of uncertainty, the collective action problem may persist: when others choose to stay at home, remaining at home might be the optimal strategy even for a dedicated opponent of the regime.

10.5.1 Authoritarian Coalition Formation

Assuming that there always exists a single incumbent leader who makes all critical decisions himself is analytically convenient, yet in reality, no dictator rules alone. Even the most personalistic of dictators need to resolve the problem of power-sharing to deal with elites and the problem of control over masses (Bueno de Mesquita et al., 2003; Meng, 2020; Svobik, 2012).

Political coalition-formation in a nondemocratic context is radically different from the democratic one. One important difference is that a dictator can share the country's resources with his supporters in ways in which democratic leaders cannot. For an authoritarian leader, the most obvious way to build a coalition of support is to buy allegiance by sharing rents or making policy concessions. Gandhi and Przeworski (2006) theorize that when a dictator faces a threat of rebellion, she makes larger policy concessions but also shares more rents and confirm this prediction using data on all dictatorships that existed between 1946 and 1996.

Bueno de Mesquita et al. (2003) organize their analysis of nondemocratic regimes using the "selectorate theory." The *selectorate* are those who participate in choosing the *winning coalition*, a subset of the selectorate that guarantees power control. It might consist of a single person in a personalized dictatorship or of all citizens eligible to vote in a perfect democracy. Members of the winning coalition stick to the current leader as they are not sure they will be included in the winning coalition of a challenger. Thus, any challenger who wants to upend the *status quo* has to offer a premium over what the members of the winning coalition receive from the incumbent.

Myerson (2008) offers a game-theoretic model, in which supporting the leader is a focal point for his (would-be) lieutenants. Unless the leader agrees, *ex ante*, to limit his authority, the lieutenants cannot be sure that she would remunerate them for the past support. As a result, in any renegotiation-proof equilibrium, the leader offers some constitutional restrictions on his power and the lieutenants' support is conditional on the leader observing these restrictions.

Empirical research supports, broadly, the idea that autocrats buy their support. Desai et al. (2009) use a panel of 80 nondemocracies from 1975 to 1999 to demonstrate the existence of the "authoritarian bargain": autocrats pay off their citizens to surrender their political freedoms. Caselli and Tesei (2016) find that in moderately entrenched autocracies, windfalls significantly exacerbate the autocratic nature of the political system: when a dictator has more rents to share, he has more control. Leon (2014) demonstrates empirically that military coups are more likely in countries that spend a relatively low share of GDP on the military. That is, sharing the rents with the military does reduce the probability of a military revolt. Using unique archival data on the allocation of cars for mid-level bureaucrats, Lazarev and Gregory (2003) analyze the microlevel of the dictator's distribution of rents. Their evidence strongly suggests that the allocation of cars, a prized commodity in the Soviet Union, was primarily a purchase of the bureaucrat's political loyalty.

Another avenue of empirical research has tried to measure the value of political connections to autocratic leaders. In a pioneer contribution, Fisman (2001) demonstrated that the share price of firms connected to Indonesian dictator Suharto suffered more than that of less-connected firms in response to news about serious health issues of the dictator. González and Prem (2020) use firm-level data from Chile to document resource misallocation in favor of politically connected firms during the transition from dictatorship to democracy: firms linked to the Pinochet regime (1973–1990) were relatively unproductive and benefited from resource misallocation under dictatorship, but then, after learning that the dictatorship was going to end, firms in the dictator's network increased their productive capacity, experienced higher profits, and obtained more loans from the main state-owned bank.

10.5.2 Institutionalized Ruling Party

One important example of a government structure in a nondemocracy is an institutionalized ruling party, a quasi-state body that is structured like a political party in a democracy, yet does not actually compete in elections. Instead, its main function is to maintain control over the rest of the society.

There have been relatively few attempts to build a model of an institutionalized ruling party and mechanisms that it uses to maintain discipline and cohesiveness inside and maintain control over the populace. Francois et al. (2016) model warring factions within the Communist Party of China. Gehlbach and Keefer (2011) suggest an informational approach toward understanding the size and scope of operation of a ruling party. Citizens are divided into two groups, the insiders and the outsiders. The difference between the former and the latter is that when the party leadership

expropriates property from an insider, other insiders are informed. Therefore, they have incentives to protect each other against expropriation. In contrast, outsiders have their property expropriated with other agents unaware, thus creating a premium for belonging to the elite. As Vladimir Lenin, the founder of the Communist Party of the Soviet Union, put it in 1921, the fourth year of the Russian Revolution: “We need full and truthful information. And the truth should not depend upon whom it has to serve. We can accept only the division between the unofficial information (for the Comintern Executive Committee only) and official information (for everybody).”

The model of Gehlbach and Keefer (2011) and Example 4.1 in Egorov and Sonin (2024) are two examples of the Coasean approach to modeling a political party. Ronald Coase’s celebrated theory of the firm puts emphasis on the distinction between in-house and out-sourced production, the difference being attributed to “transaction costs” (the agency problems). Similarly, an optimal organization of a political party would allocate some tasks in-house (what is done by party members) and some to outside producers (what is done by party supporters). The agency problems within the party would define the hierarchy of authority within the party the same way they define it in a firm. The promise of the Coasean approach is that it potentially encompasses political parties in both democratic and nondemocratic environments.

10.5.3 Balancing Power

A classic method of nondemocratic control is “divide and rule,” whereby the leader maintains control by playing different factions of the society against each other. Structurally, this is similar to the pork-barrel model of democratic politics, in which the agenda-setter might build a coalition that comprises a majority in the parliament, imposing a tax on the rest. Roemer (1985) is an early model of coalition-building aimed at maintaining nondemocratic power: the challenger proposes income redistribution; the incumbent offers a list of penalties for joining the challenger’s coalition and then failing.

Still, the “divide-and-rule” framework misses another, perhaps more salient, form of authoritarian government. Newson and Trebbi (2018), analyzing authoritarian elites in such diverse polities as sub-Saharan Africa and China, conclude that “the prevailing view of winner-take-all contests can be clearly rejected.” In an authoritative study of the last decade of Stalin’s rule, historians conclude that the most appropriate model would be that of balance of power, in which no individual politician, even Stalin himself, was able to move without building a temporary coalition of support (Gorlizki & Khlevniuk, 2005).

As we discussed at the beginning of this essay, the essential difference between democracies and dictatorships is that models of the latter cannot rely on commitment ensured by democratic institutions, established procedures, independent courts, etc. In the most extreme case, no commitment is possible at all. Next, we will use theoretical arguments from Acemoglu et al. (2008, 2009, 2012) to illustrate what happens in an environment with total absence of commitment in nondemocratic politics. In this highly stylized example, a group that has sufficient number of votes within the ruling coalition can eliminate the rest. However, actors need to be strategic

about participation in such a group, at there might be more than one round of eliminations. Being included in the initial winning coalition does not guarantee belonging to the ultimate winning coalition.

The fact that the losers in a power struggle are eliminated forever, without a possibility of return, is a drastic departure from the theory of democratic coalition formation, where coalitions can be perpetually formed and dissolved. And assuming that loser of a power struggle in an authoritarian regime lose their lives as well is not far-fetched. The Soviet political history in 1919–1952 is the case in point. A small group of the Communist party officials, the Politburo, was effectively the ruling coalition. All top government positions were held by its members. Though formally its members were elected at the Communist Party meetings, for all practical purposes, the Politburo determined the fates of its members, as well as those of ordinary citizens. Of 40 Politburo members (28 full, 12 nonvoting) appointed between 1919 and 1952, only 12 survived through 1952. Of these 12, 11 continued to hold top positions after Stalin's death in March 1953. There was a single Politburo member (Grigory Petrovsky) in 33 years who left the body and survived. Of the 28 deaths, there were 17 executions decided by the Politburo, 2 suicides, 1 death in prison immediately after arrest, and 1 assassination.

Acemoglu et al. (2008) use the following example to tell the story of Joseph Stalin's succession fight. Power configuration $\frac{3}{42}, \frac{4}{42}, \frac{5}{42}, \frac{10}{42}, \frac{20}{42}$ is stable as elimination of any single member results in further eliminations. As a result, in any coalition that has a majority, there is always a member who would not want to take part of it, expecting to be eliminated in the next round. Now, consider the demise of the strongest member with power $\frac{20}{42}$ (Stalin). The new power configuration is unstable. Then, the three weakest members, $\frac{3}{22}, \frac{4}{22}, \frac{5}{22}$ (Khrushchev, Malenkov, and Bulganin), eliminate the strongest of the remaining members, $\frac{10}{22}$ (Beria), and form a new stable coalition $\frac{3}{12}, \frac{4}{12}, \frac{5}{12}$. (The denominators change as the power of the eliminated member is redistributed proportionally among those who voted for elimination.)

The above example demonstrates that nondemocratic stability might be an equilibrium outcome even if there is no single dominating force, for example, a leader or an agenda-setter who punishes a deviator. In this equilibrium, the ruling coalition is not necessarily minimal, and the agent with the highest amount of individual power is not necessarily included in the ultimate winning coalition. Finally, this model squares well with evidence on some modern autocracies. Francois et al. (2015) show that African ruling coalitions are relatively large and ethnic groups are proportionally represented. Newson and Trebbi (2018) found similar results analyzing authoritarian elites in sub-Saharan Africa.

10.6 Conclusion

There is little doubt that authoritarian institutions impede economic development. With very few and very special exceptions, nearly all of today's developed countries are democracies. Authoritarian institutions are designed to limit citizens' access to

political and, ultimately, economic rents. They tend to preserve the status quo, as creative destruction is antithetical to them.

Drawing on an extensive global dataset, with some time series going back to the early nineteenth century, Knutsen (2015) reports robust evidence that democracy increases not only technology-induced growth but also net economic growth rates. Notably, the results hold when accounting for the endogeneity of democracy, country-fixed effects, and sample-selection bias. The mechanism of authoritarian failures is broadly understood as well: authoritarian institutions that keep dictators in power provide poor protection of property rights, which is critical for growth. Duong et al. (2022) confirm this general observation, finding a positive relation between democracy and shareholder protection, proxied by the difference between the initial price and the eventual market price of firm shares. Using data for political regimes, income, and human capital for a sample of 141 countries over 1500–2000, Madsen et al. (2015) find democracy, instrumented by linguistic distance-weighted foreign democracy, to be a significant determinant of income and growth, controlling for human capital, among other key variables. Papaioannou and Van Zanden (2015) present evidence that a dictator's long years in office reduce economic growth, increase inflation, and harm the quality of institutions. The negative effect is particularly strong in young states and in Africa and the Near East.

Still, there is some evidence that an authoritarian regime might be conducive to economic activity—perhaps at some stages of economic development or in some specific areas. (Luo and Przeworski (2019) attribute autocratic “growth miracles” to the fact that to grow very fast, a country needs to have a low starting point, and poor countries are typically autocratic.) For example, Giavazzi and Tabellini (2005) demonstrate, using difference-in-difference estimates, that countries that first liberalized their economies and then became democracies perform much better than countries that pursued the opposite sequence. (See Riedl et al. (2020) for a recent review of the political science literature on authoritarian-led democratizations.)

Even if concentrated political authority aids economic development at some stages, authoritarian institutions discourage innovation along various dimensions and thus impede modern technology-based growth. Acemoglu and Robinson (2005) argued that concerns about losing control slowed down the embrace of industrialization and railroads by the absolutist Russian and Hapsburg Empires in the early nineteenth century. The lesson gains new relevance now that the Chinese economy, after 40 years of rapid growth under an authoritarian regime, is approaching the technological frontier. For decades, Chinese growth exploited the advantage of technological backwardness and access to enormous reserves of cheap labor. By the third decade of the twenty-first century, both resources have been depleted. Can China count on productivity-led growth without a dramatic expansion of democratic institutions? The country seems to be shifting from a regime structured around an institutionalized ruling party to a more personalistic, archaic regime, that is, in the direction opposing democratization. If there is any accumulated wisdom in the literature on nondemocratic politics, it is that the emerging regime is inconsistent with fast growth and development.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., & Robinson, J. A. (2000). Why did the west extend the franchise? Democracy, inequality, and growth in historical perspective. *The Quarterly Journal of Economics*, 115(4), 1167–1199.
- Acemoglu, D., & Robinson, J. A. (2001). A theory of political transitions. *American Economic Review*, 91(4), 938–963.
- Acemoglu, D., & Robinson, J. A. (2005). *Economic origins of dictatorship and democracy*. Cambridge University Press.
- Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Crown Publishers.
- Acemoglu, D., Egorov, G., & Sonin, K. (2008). Coalition formation in non-democracies. *Review of Economic Studies*, 75(4), 987–1009.
- Acemoglu, D., Egorov, G., & Sonin, K. (2009). Do juntas Lead to personal rule? *American Economic Review Papers and Proceedings*, 99(2), 298–303.
- Acemoglu, D., Egorov, G., & Sonin, K. (2011). Political model of social evolution. *Proceedings of the National Academy of Sciences*, 108(supplement 4), 21292–21296.
- Acemoglu, D., Egorov, G., & Sonin, K. (2012). Dynamics and stability of constitutions, coalitions, and clubs. *American Economic Review*, 102(4), 1446–1476.
- Acemoglu, D., Egorov, G., & Sonin, K. (2015). Political economy in a changing world. *Journal of Political Economy*, 123(5), 1038–1086.
- Acemoglu, Daron, Georgy Egorov and Konstantin Sonin. 2020. Institutional change and institutional persistence. In *Handbook of Historical Economics* ed. Alberto Bisin and Giovanni Federico. Elsevier.
- Adena, M., Enikolopov, R., Petrova, M., Santarosa, V., & Zhu-ravskaya, E. (2015). Radio and the rise of the Nazis in prewar Germany. *Quarterly Journal of Economics*, 130(4), 1885–1939.
- Arendt, H. (1951). *The origins of totalitarianism*. Harcourt Brace.
- Bai, Y., & Zhou, T. (2019). “Mao’s last revolution”: A Dictator’s loyalty — Competence tradeoff. *Public Choice*, 180(3), 469–500.
- Balima, H. W. (2020). Coups d’Etat and the cost of debt. *Journal of Comparative Economics*, 48(3), 509–528.
- Barzel, Y. (1977). An economic analysis of slavery. *The Journal of Law and Economics*, 20(1), 87–110.
- Bergemann, D., & Morris, S. (2019). Information design: A unified perspective. *Journal of Economic Literature*, 57(1), 44–95.
- Besley, T., & Kudamatsu, M. (2009). *Making autocracy work* (pp. 452–510). Harvard University Press.
- Boix, C., & Svolik, M. W. (2013). The foundations of limited authoritarian government: Institutions, commitment, and power-sharing in dictatorships. *Journal of Politics*, 75(2), 300–316.
- Boleslavsky, R., Shadmehr, M., & Sonin, K. (2021). Media freedom in the shadow of a coup. *Journal of European Economic Association*, 19(3), 1782–1815.
- Bueno de Mesquita, B., & Smith, A. (2017). Political succession: A model of coups, revolution, purges, and everyday politics. *Journal of Conflict Resolution*, 61(4), 707–743.
- Bueno de Mesquita, B., Smith, A., Siverson, R., & Morrow, J. D. (2003). *The logic of political survival*. MIT Press.
- Cantu, F. (2019). The fingerprints of fraud: Evidence from Mexico’s 1988 presidential election. *American Political Science Review*, 113(3), 710–726.

- Caselli, F., & Tesei, A. (2016). Resource windfalls, political regimes, and political stability. *Review of Economics and Statistics*, 98(3), 573–590.
- Chen, Y., & Yang, D. Y. (2019). The impact of media censorship: 1984 or brave New World? *American Economic Review*, 109(6), 2294–2332.
- Cox, G. (2009). Authoritarian elections and leadership succession, 1975–2004 (*APSA 2009 meeting paper*).
- Crawford, V. P., & Sobel, J. (1982). Strategic information transmission. *Econometrica*, 50(6), 1431–1451.
- Desai, R. M., Olofsgård, A., & Yousef, T. M. (2009). The logic of authoritarian bargains. *Economics and Politics*, 21(1), 93–125.
- Dragu, T., & Przeworski, A. (2019). Preventive repression: Two types of moral hazard. *American Political Science Review*, 113(1), 77–87.
- Duong, H. N., Goyal, A., Kallinterakis, V., & Veeraraghavan, M. (2022). Democracy and the pricing of initial public offerings around the world. *Journal of Financial Economics*, 145(1), 322–341.
- Easterly, W., & Pennings, S. (2017). Shrinking dictators: how much economic growth can we attribute to national leaders? *mimeo*.
- Edmond, C. (2013). Information manipulation, coordination, and regime change. *Review of Economic Studies*, 80(4), 1422–1458.
- Egorov, G., & Sonin, K. (2011). Dictators and their viziers: Endogenizing the loyalty–competence trade-off. *Journal of the European Economic Association*, 9(5), 903–930.
- Egorov, G., & Sonin, K. (2015). The killing game: A theory of non-democratic succession. *Research in Economics*, 69(3), 398–411.
- Egorov, G., & Sonin, K. (2021). Elections in non-democracies. *Economic Journal*, 131(636), 1682–1716.
- Egorov, G., & Sonin, K. (2023). *Why did Putin invade Ukraine? A theory of degenerate autocracy* (Technical report National Bureau of Economic Research).
- Egorov, G., & Sonin, K. (2024). The political economics of non-democracy. *Journal of Economic Literature*, 62(2), 594–636.
- Egorov, G., Guriev, S., & Sonin, K. (2009). Why resource-poor dictators allow freer media: A theory and evidence from panel data. *American Political Science Review*, 103, 645–668.
- Ellis, C. J., & Fender, J. (2010). Information cascades and revolutionary regime transitions. *Economic Journal*, 121(553), 763–792.
- Enikolopov, R., Korovkin, V., Petrova, M., Sonin, K., & Zakharov, A. (2013). Field experiment estimate of electoral fraud in Russian parliamentary elections. *Proceedings of the National Academy of Sciences*, 110(2), 448–452.
- Fearon, J. (1995). Rationalist explanations for war. *International Organization*, 49(3), 379–414.
- Fisman, R. (2001). Estimating the value of political connections. *American Economic Review*, 91(4), 1095–1102.
- Francois, P., Rainer, I., & Trebbi, F. (2015). How is power shared in Africa? *Econometrica*, 83(2), 465–503.
- Francois, P., Trebbi, F., & Xiao, K. (2016). *Factions in nondemocracies: Theory and evidence from the Chinese Communist Party* (National Bureau of Economic Research WP (27755)).
- Fukuyama, F. (1992). *The end of history and the last man*. Penguin Adult.
- Gandhi, J., & Przeworski, A. (2006). Cooperation, cooptation, and rebellion under dictatorships. *Economics and Politics*, 18(1), 1–26.
- Gandhi, J., & Przeworski, A. (2007). Authoritarian institutions and the Survival of autocrats. *Comparative Political Studies*, 40(11), 1279–1301.
- Gehlbach, S., & Keefer, P. (2011). Investment without democracy: Ruling-party institutionalization and credible commitment in autocracies. *Journal of Comparative Economics*, 39(2), 123–139.
- Gehlbach, S., & Simpson, A. (2015). Electoral manipulation as bureaucratic control. *American Journal of Political Science*, 59(1), 212–224.

- Gehlbach, S., & Sonin, K. (2008). *Government control of the media* (Working papers w0126 Center for Economic and Financial Research (CEFIR)).
- Gehlbach, S., & Sonin, K. (2014). Government control of the media. *Journal of Public Economics*, 118, 163–171.
- Gehlbach, S., Sonin, K., & Svolik, M. W. (2016). Formal models of non-democratic politics. *Annual Review of Political Science*, 19, 565–584.
- Giavazzi, F., & Tabellini, G. (2005). Economic and political liberalizations. *Journal of Monetary Economics*, 52(7), 1297–1330.
- Gitmez, A. A., & Sonin, K. (2023). *The Dictator's dilemma: A theory of propaganda and repression* (Becker Friedman Institute for Economics Working Paper (2023–67)).
- Glaessel, C., & Paula, K. (2020). Sometimes less is more: Censorship, news falsification, and disapproval in 1989 East Germany. *American Journal of Political Science*, 64(3), 682–698.
- González, F., & Prem, M. (2020). Losing your dictator: Firms during political transition. *Journal of Economic Growth*, 25(2), 227–257.
- Gorlizki, Y., & Khlevniuk, O. (2005). *Cold peace: Stalin and the soviet ruling circle, 1945–1953*. Oxford University Press.
- Gregory, P. R., Schröder, P. J. H., & Sonin, K. (2011). Rational dictators and the killing of innocents: Data from Stalin's archives. *Journal of Comparative Economics*, 39(1), 34–42.
- Greif, A. (1998). Self-enforcing political systems and economic growth: Late medieval Genoa. In R. Bates & M. Levi (Eds.), *Analytic narratives* (pp. 25–64).
- Guriev, S., & Sonin, K. (2009). Dictators and oligarchs: A dynamic theory of contested property rights. *Journal of Public Economics*, 93(1–2), 1–13.
- Guriev, S., & Treisman, D. (2019). Informational autocrats. *Journal of Economic Perspectives*, 33(4), 100–127.
- Guriev, S., & Treisman, D. (2022). *Spin dictators: The changing face of tyranny in the 21st century*. Princeton University Press.
- Herz, J. H. (1952). The problem of successorship in dictatorial regimes; a study in comparative law and institutions. *Journal of Politics*, 14(1), 19–40.
- Hoff, K., & Stiglitz, J. E. (2004). After the big bang? Obstacles to the emergence of the rule of law in post-communist societies. *American Economic Review*, 94(3), 753–763.
- Hollyer, J. R., Peter Rosendorff, B., & Vreeland, J. R. (2018). *Transparency, democracy, and autocracy: Economic transparency and political (in) stability*. Cambridge University Press.
- Jia, R., Kudamatsu, M., & Seim, D. (2015). Political selection in China: The complementary roles of connections and performance. *Journal of the European Economic Association*, 13(4), 631–668.
- Jones, B. F., & Olken, B. A. (2005). Do leaders matter? National Leadership and growth since world war II. *Quarterly Journal of Economics*, 120(3), 835–864.
- Kamenica, E., & Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6), 2590–2615.
- King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, 107(2), 326–343.
- King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, 345(6199), 1–10.
- Knight, B., & Tribin, A. (2018). The limits of propaganda: Evidence from Chavez's Venezuela. *Journal of the European Economic Association*, 17(2), 567–605.
- Knutsen, C. H. (2015). Why democracies outgrow autocracies in the long run: Civil liberties, information flows and technological change. *Kyklos*, 68(3), 357–384.
- Kokkonen, A., & Sundell, A. (2014). Delivering stability—Primogeniture and autocratic survival in European monarchies 1000–1800. *American Political Science Review*, 108(2), 438–453.
- Konrad, K. A., & Mui, V.-L. (2017). The prince or better no prince? The strategic value of appointing a successor. *Journal of Conflict Resolution*, 61(10), 2158–2182.
- Kornai, J. (1986). The soft budget constraint. *Kyklos*, 39(1), 3–30.

- Kosterina, S. (2017). Ambition, personalist regimes, and control of authoritarian leaders. *Journal of Theoretical Politics*, 29(2), 167–190.
- Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2), 245–252.
- Kurrid-Klitgaard, P. (2000). The constitutional economics of autocratic succession. *Public Choice*, 103(1–2), 63–84.
- Kydland, F. E., & Prescott, E. C. (1977). Rules rather than discretion: The inconsistency of optimal plans. *Journal of Political Economy*, 85(3), 473–491.
- Lazarev, V., & Gregory, P. (2003). Commissars and cars: A case study in the political economy of dictatorship. *Journal of Comparative Economics*, 31(1), 1–19.
- Leon, G. (2014). Loyalty for Sale? Military spending and coups d'état. *Public Choice*, 159(4), 363–383.
- Little, A. T. (2023). Bayesian explanations for persuasion. *Journal of Theoretical Politics*, 35(3), 147–181.
- Little, A. T., Tucker, J. A., & LaGatta, T. (2015). Elections, protest, and alternation of power. *Journal of Politics*, 77(4), 1142–1156.
- Lizzeri, A., & Persico, N. (2004). Why did the elites extend the suffrage? Democracy and the scope of government, with an application to Britain's "age of reform". *The Quarterly Journal of Economics*, 119(2), 707–765.
- Lohmann, S. (1994). The dynamics of informational cascades: The Monday demonstrations in Leipzig, East Germany, 1989–91. *World Politics*, 47(1), 42–101.
- Lorentzen, P. L. (2013). Regularizing rioting: Permitting public protest in an authoritarian regime. *Quarterly Journal of Political Science*, 8(2), 127–158.
- Lorentzen, P. (2014). China's strategic censorship. *American Journal of Political Science*, 58(2), 402–414.
- Luo, Z., & Przeworski, A. (2019). Why are the fastest growing countries autocracies? *The Journal of Politics*, 81(2), 663–669.
- Madsen, J. B., Raschky, P. A., & Skali, A. (2015). Does democracy drive income in the world, 1500–2000? *European Economic Review*, 78, 175–195.
- Magaloni, B. (2010). The game of electoral fraud and the ousting of authoritarian rule. *American Journal of Political Science*, 54(3), 751–765.
- Martinez-Bravo, M., Mukherjee, P., & Stegmann, A. (2017). The non-democratic roots of elite capture: Evidence from Soeharto mayors in Indonesia. *Econometrica*, 85(6), 1991–2010.
- Martinez-Bravo, M., iMiquel, G. P., Qian, N., & Yao, Y. (2022). The rise and fall of local elections in China. *American Economic Review*, 112(9), 2921–2958.
- Mattingly, D. C. (2024). How the party commands the gun: The foreign–domestic threat dilemma in China. *American Journal of Political Science*, 68(1), 227–242.
- McGuire, M. C., & Olson, M. (1996). The economics of autocracy and majority rule: The invisible hand and the use of force. *Journal of Economic Literature*, 34(1), 72–96.
- Meng, A. (2020). *Constraining dictatorship: From personalized rule to institutionalized regimes*. Political Economy of Institutions and Decisions Cambridge University Press.
- Meng, X., Qian, N., & Yared, P. (2015). The institutional causes of China's great famine, 1959–1961. *Review of Economic Studies*, 82(4), 1568–1611.
- Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12, 380–391.
- Miller, M. K. (2015). Elections, information, and policy responsiveness in autocratic regimes. *Comparative Political Studies*, 48(6), 691–727.
- Montagnes, P., & Wolton, S. (2019). Mass purges: Top-down accountability in autocracy. *American Political Science Review*, 113(4), 1045–1059.
- Murphy, K. M., Shleifer, A., & Vishny, R. W. (1993). Why is rent-seeking so costly to growth? *The American Economic Review*, 83(2), 409–414.
- Myerson, R. B. (2008). The Autocrat's credibility problem and foundations of the constitutional state. *American Political Science Review*, 102(1), 125–139.

- Myerson, R. B. (2010). Capitalist investment and political liberalization. *Theoretical Economics*, 5(1), 73–91.
- Myerson, R. B. (2015). Moral Hazard in high office and the dynamics of aristocracy. *Econometrica*, 83(6), 2083–2126.
- Naumenko, N. (2021). The political economy of famine: The Ukrainian famine of 1933. *Journal of Economic History*, 81(1), 156–197.
- Newson, A., & Trebbi, F. (2018). Authoritarian elites. *Canadian Journal of Economics*, 51(4), 1088–1117.
- North, D. C. (1981). *Structure and change in economic history*. Norton.
- North, D. (1990). *Institutions, institutional change and economic performance* Cambridge university press. Cambridge University Press.
- North, D. C., & Weingast, B. (1989). Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century England. *Journal of Economic History*, 49(4), 803–832.
- North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.
- Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Harvard University Press.
- Olson, M. (1991). Autocracy, democracy, and prosperity. In R. Zeckhauser (Ed.), *Strategy and choice*. MIT Press.
- Papaioannou, E., & Siourounis, G. (2008). Democratisation and growth. *Economic Journal*, 118(532), 1520–1551.
- Papaioannou, K., & Van Zanden, J. L. (2015). The dictator effect: How long years in office affect economic development. *Journal of Institutional Economics*, 11(1), 111–139.
- Polishchuk, L., & Savvateev, A. (2004). Spontaneous (non) emergence of property rights. *The Economics of Transition*, 12(1), 103–127.
- Przeworski, A. (2009). Constraints and choices: Electoral participation in historical perspective. *Comparative Political Studies*, 42(1), 4–30.
- Qin, B., Strömberg, D., & Yanhui, W. (2018). Media bias in China. *American Economic Review*, 108(9), 2442–2476.
- Riedl, R. B., Slater, D., Wong, J., & Ziblatt, D. (2020). Authoritarian-led democratization. *Annual Review of Political Science*, 23, 315–332.
- Rodrik, D. (2000). Institutions for high-quality growth: What they are and how to acquire them. *Studies in Comparative International Development*, 35(3), 3–31.
- Rodrik, D., & Wacziarg, R. (2005). Do democratic transitions produce bad economic outcomes? *American Economic Review*, 95(2), 50–55.
- Roemer, J. E. (1985). Rationalizing revolutionary ideology. *Econometrica*, 53(1), 85–108.
- Rosenfeld, B., & Wallace, J. (2024). Information politics and propaganda in authoritarian societies. *Annual Review of Political Science*, 27, 263–281.
- Rozenas, A. (2020). A theory of demographically targeted repression. *Journal of Conflict Resolution*, 64(7–8), 1254–1278.
- Rozenas, A., & Stukal, D. (2019). How autocrats manipulate economic news: Evidence from Russia's state-controlled television. *Journal of Politics*, 81(3), 982–996.
- Rundlett, A., & Svolik, M. (2016). Deliver the vote! Micromotives and macrobehavior in electoral fraud. *American Political Science Review*, 110(1), 180–197.
- Schwarz, M., & Sonin, K. (2008). A theory of brinkmanship, conflicts, and commitments. *Journal of Law, Economics, and Organization*, 24(1), 163–183.
- Shadmehr, M., & Bernhardt, D. (2011). Collective action with uncertain payoffs: Coordination, public signals, and punishment dilemmas. *American Political Science Review*, 105, 829–851.
- Shadmehr, M., & Bernhardt, D. (2015). State censorship. *American Economic Journal: Microeconomics*, 7(2), 280–307.
- Sheen, G. C.-H., Tung, H. H., & Wen-Chin, W. (2022). Power sharing and media freedom in dictatorships. *Political Communication*, 39(2), 202–221.

- Shepsle, K. A. (1991). Discretion, institutions, and the problem of government commitment. In *Social theory for a changing society*. Western press.
- Shih, V. (2022). *Coalitions of the weak* (Cambridge studies in comparative politics). Cambridge University Press.
- Shih, V., Adolph, C., & Liu, M. (2012). Getting ahead in the Communist Party: Explaining the advancement of central committee members in China. *American Political Science Review*, 106(1), 166–187.
- Simpser, A. (2013). *Why governments and parties manipulate elections: Theory, practice, and implications*. Cambridge University Press.
- Sonin, K. (2003). Why the rich may favor poor protection of property rights. *Journal of Comparative Economics*, 31(4), 715–731.
- Spence, M. (1973). Job Market Signaling. *Quarterly Journal of Economics*, 87(3), 355–374.
- Svolik, M. W. (2012). *The politics of authoritarian rule*. Cambridge university press.
- Tullock, G. (1971). The paradox of revolution. *Public Choice*, 11(1), 89–99.
- Tyson, S. A. (2018). The agency problem underlying repression. *The Journal of Politics*, 80(4), 1297–1310.
- Wig, T., & Rod, E. G. (2016). Cues to coup plotters: Elections as coup triggers in dictatorships. *Journal of Conflict Resolution*, 60(5), 787–812.
- Williamson, O. E. (1983). Credible commitments: Using hostages to support exchange. *The American Economic Review*, 73(4), 519–540.
- Wintrobe, R. (1990). The Tinpot and the totalitarian: An economic theory of dictatorship. *American Political Science Review*, 84, 849–872.
- Wintrobe, R. (1998). *The political economy of dictatorship*. Cambridge University Press.
- Xu, X. (2021). To repress or to co-opt? Authoritarian control in the age of digital surveillance. *American Journal of Political Science*, 65(2), 309–325.
- Zakharov, A. V. (2016). The loyalty-competence trade-off in dictatorships and outside options for subordinates. *Journal of Politics*, 78(2), 457–466.
- Zhuravskaya, E., Petrova, M., & Enikolopov, R. (2020). Political effects of the internet and social media. *Annual Review of Economics*, 12(1), 415–438.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Emergence and Failure of Democracy: Explaining Paths of Political Regime Change

11

Daniel Treisman

Contents

11.1	Introduction	238
11.2	Types of Political Regimes	239
11.3	Macro-Processes	241
11.3.1	Modernization Theory	241
11.3.2	Autonomous Institutional Evolution	241
11.3.3	Limitations of Both Perspectives	242
11.4	Micro-Processes	245
11.4.1	Why Do Rulers Democratize?	245
11.4.2	What Does Democracy Do?	248
11.4.3	Limitations of Micro-Accounts	249
11.5	How These Interact	250
11.5.1	Conditional Modernization Theory	250
11.5.2	Parameters and Equilibria	251
11.6	Disaggregating Democracy	252
11.7	Global Modernization Theory	252
11.8	Democratic Breakdown	254
11.9	Conclusion	257
	References	259

Abstract

Why do countries' political regimes change over time—from dictatorship to democracy and vice versa? Competing literatures contend that the key long-run driver is, respectively, economic development or institutional evolution. Other accounts focus on short-run microinteractions among societal actors. I discuss the strengths and limitations of these approaches and suggest ways in which they might usefully be combined. While simple modernization theory implies more cross-national convergence than we have seen, institutional path-dependence predicts too much divergence. Analyses of microinteractions overemphasize

D. Treisman (✉)

Department of Political Science, University of California, Los Angeles, CA, USA

e-mail: treisman@polisci.ucla.edu

deliberate choice and ignore copious historical evidence of confusion and error. They also fail to explain the spread of elections. New approaches posit that microinteractions trigger change while its direction is set by slow-moving structural factors (“conditional modernization theory”) or that economic development has effects at both country and global levels (“global modernization theory”). Other insights may come from disaggregating democracy into subelements, each with a distinct history and logic.

Keywords

Democracy · Dictatorship · Authoritarian · Regime · Political institutions · Economic development · Path dependence · Mistakes · Backsliding

11.1 Introduction

Why do dictatorships change into democracies and vice versa? This question has engaged philosophers since at least Plato, who saw a tendency for aristocracy to decay into timocracy, oligarchy, democracy, and ultimately tyranny. Aristotle posited regression from monarchy to anarchy, while later Polybius saw rotation among ochlocracy, monarchy, tyranny, aristocracy, oligarchy, and democracy. More recently, some writers have replaced theories of cyclical fluctuation with a belief in secular trends. Rival schools see the ultimate engine of change in either economic development or autonomous institutional evolution. Still other scholars, rejecting any long-run logic, use the tools of game theory to decompose transitions into short-run strategic interactions among individual actors.

While recent work has generated important insights, I argue that none of the main approaches has produced a completely satisfactory account. Theories rooted in economic development—which has transformed almost all countries since the Industrial Revolution—tend to predict too much convergence. Yet, those focused on path-dependent institutional evolution imply too much divergence. Micro-level accounts of the games of transition tend to show how small differences in actors or setting can produce opposite outcomes—which renders even more puzzling the striking temporal and geographical patterns of the last century. Rather than comprehensive, stand-alone explanations, each of these yields at best a partial understanding of the process. The most exciting recent work, I suggest, has explored how modernization and institutional legacies, macro-trends and micro-games, interact. In what follows, I review some themes of past debates and highlight how new approaches are opening possible directions for future research.

A major challenge in studying political change is how to use empirical evidence to evaluate theories. Scholars cannot randomly assign levels of economic development or legal institutions to states. “Natural experiments” are rare and tend to bundle multiple “treatments,” leaving the role of each unclear. Some scholars have proposed instrumental variables to capture exogenous variation in aspects of economic or institutional change; I refer to some of these below. But few credibly satisfy exclusion

restrictions. To make things worse, economic development and institutions are both complex, multidimensional concepts, each of which could plausibly influence the other, either simultaneously or consecutively. What causal links one finds are likely to depend on which aspects and indicators of each one chooses and in what temporal and geographical windows one looks. Given this, there is something to be said for beginning modestly with basic reality checks. Do the observable implications of theories at least roughly fit available evidence, whether in the form of statistical data or detailed historical records? A positive answer is hardly conclusive, but a negative one does suggest the need for a rethink.

I proceed as follows. Section 11.2 reviews common classifications of political regimes. Section 11.3 discusses recent work that traces democratization to macro-processes of economic and institutional change. Section 11.4 turns to the micro-games of political actors. Section 11.5 considers how macro- and micro-levels interact. Section 11.6 reviews arguments about international influences and how they may combine with domestic processes. Section 11.7 discusses recent work disaggregating democracy into constituent elements and examining the historical processes by which these emerge. While previous sections focus on democratization, Sect. 11.8 addresses literature on democratic breakdowns and “backsliding” to authoritarian governance. Section 11.9 concludes.

11.2 Types of Political Regimes

Political regimes are the “basic formal and informal rules for choosing leaders and policies” (Geddes et al., 2018, 5). Scholars typically divide these into two broad groups—democracies and nondemocracies (often referred to interchangeably as authoritarian regimes, autocracies, and dictatorships). Until the eighteenth century, democracy was generally understood to be a system in which all adult male citizens were members of the polity’s legislative assembly. It meant actual *rule* by the people: the adult male citizens had the right and duty to make decisions about public affairs that were legally binding for all residents. As Aristotle first noted, that almost always meant rule by the poor.

By contrast, today’s democracies are representative rather than direct. Instead of ruling in person, citizens are represented by legislative and executive branch officials. Modern democracies are defined by how those representatives are chosen: by means of elections that are regular, competitive, reasonably free, and fair and involve secret, universal, equal adult suffrage. In other words, democracy is identified with the procedures used to allocate decision-making power rather than the substance of the decisions made or the ideals the system is supposed to embody (Schumpeter, 1942; Kelsen, 1955; Bobbio, 1987, 24–5). One can distinguish between “electoral” democracies, which meet the conditions already mentioned, and “liberal” democracies, which also include constitutional constraints designed to protect civil liberties such as freedom of speech, the press, assembly, and association. Of course, some degree of such freedoms is necessary for elections to be “free and fair,” and the exact dividing lines are subject to debate. Still, considerable consensus has emerged on this basic

procedural definition: democracies are systems that allocate political authority via a certain type of election (Munck, 2023, Chap. 4).¹

Nondemocracies have been classified in a number of ways. Linz (2000) contrasted authoritarian and totalitarian versions, with the latter denoting a more complete atomization of society. Geddes et al. (2018) distinguished among monarchies and military, one-party, and personalistic regimes based on who held power. Another approach (Bueno De Mesquita et al., 2005, 41–51) is to categorize nondemocracies by the size of their “selectorate”—the set of people with the right to participate in choosing government leaders—and of the “winning coalition,” the subset of the selectorate whose favored candidates actually rule. Others have distinguished between nondemocracies with and without elections (Levitsky & Way, 2010; Schedler, 2009) and between those that rule primarily by fear and violence versus those that dominate by manipulating information (Guriev & Treisman, 2019, 2022). As discussed later, different subtypes have different probabilities of democratization.

Empirical work on regime change has used several datasets and indicators. In the past, the main sources were the Polity project and Freedom House. Oddly enough, neither measures democracy per se; Polity classifies “authority characteristics of states,” while Freedom House evaluates political and civil liberties. More recently, scholars have focused on the Varieties of Democracy database (V-DEM), which, using expert-based ratings of different regime features, divides states into four categories—“closed autocracy,” “electoral autocracy,” “electoral democracy,” and “liberal democracy.”² Another choice is the Lexical Index of Electoral Democracy, which classifies regimes based on whether they hold competitive elections and whether male and female suffrage are universal (Skaaning et al., 2015).³

What leads to the replacement of one type of regime by another? Explanations have usually focused on two levels of analysis.

¹For decisions to be “legally binding,” there must be some degree of “rule of law”—that is, laws must be reliably and fairly enforced. Again, theorists disagree on how to define the minimum required level of this. One can further divide democracies, whether electoral or liberal, into various subtypes—parliamentary versus presidential, federal versus unitary, and so on. Such differences could affect regime durability, as mentioned below.

²This refers to V-DEM’s *v2x_regime* variable. “Electoral autocracies” are regimes that hold elections that are not in practice free, fair, and multiparty, or that restrict suffrage, or in which freedoms of expression and association fall below a minimal threshold. “Electoral democracies” hold elections that are in practice free, fair, and multiparty, with universal suffrage and at least a minimal level of freedom of expression and association, but fail with regard to access to justice, transparent law enforcement, respect for rule of law, or constraints on the executive. “Liberal democracies” pass on all these dimensions.

³Details on Polity are at <https://www.systemicpeace.org/inscrdata.html>; those on Freedom House are at <https://freedomhouse.org/report/freedom-world>; V-DEM’s codebook is at https://v-dem.net/documents/24/codebook_v13.pdf.

11.3 Macro-Processes

Some see regime change as driven by long-run historical processes that shape so-called structural factors—economic, social, political, environmental, or international features of the world that individual actors take as given in the short run. Two families of arguments have been particularly influential in the last half century.

11.3.1 Modernization Theory

A first perspective—often traced to Lipset (1959)—attributes the spread of democracy to the complex syndrome of economic and social changes that have unfolded in the West since the Industrial Revolution, later extended to other parts of the globe. This syndrome has many interrelated elements: the transition from agriculture and artisanal production to factory manufacturing and ultimately manipulation of symbols in the “knowledge economy” to produce intellectual property, the associated broadening and deepening of education, the exodus from the countryside to cities, the weakening of organized religion, rising incomes, the expansion of global trade and travel, and the replacement of inherited social identities by individually chosen ones. Oversimplifying a little, we see a world of corporate status groups embedded in traditional hierarchies transformed into one of self-defining individuals with human rights. While these changes tend to occur together, there are important exceptions, and sequences can differ.

A range of posited mechanisms convert these economic and social developments into pressures for more democratic government. Changes in production drive shifts in the balance of classes. Larger, more resourceful middle classes press for political access (Fukuyama, 2014, 406; Ansell & Samuels, 2014). Economic transformations increase or decrease income and wealth inequality in ways that might foster broader enfranchisement (Acemoglu & Robinson, 2000, 2006). The transition from agriculture to industry and the knowledge economy renders assets increasingly mobile and more difficult to expropriate, reducing property owners’ insistence on monopolizing government (Boix, 2003). More educated people are harder to control and knowledge workers harder to motivate except through participatory schemes. Global integration may facilitate the spread of democratic ideas and influence.

11.3.2 Autonomous Institutional Evolution

A second approach—championed by Douglass North (1981, 1990)—sees economic, social, and political change as all driven by a deeper process of institutional evolution that is largely autonomous. Several key features characterize this process. Institutions—understood as “rules of the game in society”—exhibit increasing returns, which means that adjusting existing ones is usually more cost-effective than starting from scratch (North, 1990, 3, 94–6). Adaptation is cheaper at the margin than radical change. This generates “path dependence”—the existing matrix

of institutions reflects a history of moves rather than optimization under current conditions (Arthur, 1994). Individuals' mental models and norms coevolve with the formal institutions, creating "lock-in"—new formal institutions that do not fit with inherited modes of thought tend to be rejected. A consequence of these features is that small differences at one point in time can lead to widely divergent trajectories and outcomes later on.

One example of this approach is the recent theory of the "narrow corridor" (Acemoglu & Robinson, 2020). The key insight is that the nature of political regimes depends on the relative strength of institutions of the state and of society. Over time, countries sort into three equilibria—anarchy (strong society, weak state), despotism (weak society, strong state), and democracy (strong society, strong state).⁴ To get to democracy, the strength of society and state must grow in balance. If the ratio of state to society strength ever exceeds a certain threshold, the country gravitates to despotism. If the ratio falls below a different threshold, it descends into anarchy.

Another influential example is North et al.'s (2009) account of the emergence of "open access orders"—systems in which almost all citizens have the right to create impersonal organizations, the property rights and contracts of which are legally enforced. Such orders, which correspond to a subset of liberal democracies and contrast with the "closed access" natural state, emerge in two stages. First, three institutional "doorstep conditions" must be met: rule of law for elites, perpetually lived organizations, and political control of the military (Ibid, 26). Then elites *may* complete the transition by extending rule of law and the right to form impersonal organizations to nonelites.

11.3.3 Limitations of Both Perspectives

In their simplest forms, neither the Lipset nor the North approach can fully explain observed patterns of regime change. Modernization theory has been criticized on several grounds. Some of the worlds' richest countries are authoritarian and have been rich for decades without democratizing. There is no threshold level of income—or any other element of the syndrome—at which countries always transition to popular government (Przeworski, 2000, 137). In fact, the correlation between income and democracy has fallen since the 1970s (Treisman, 2020b). At the same time, some relatively poor countries are at least electoral—if not liberal—democracies.⁵ And many of today's liberal democracies first became such when their income was low (Stasavage, 2020, 261–62)).⁶ Finally, the idea that

⁴In the authors' terminology, "absent Leviathan," "despotic Leviathan," and "shackled Leviathan."

⁵Among V-DEM's "liberal democracies," the poorest in 2022 was Barbados, with per-capita GDP of \$17,837 (World Bank, PPP US dollars). The poorest "electoral democracy" was Niger at \$1,505. And, as if on cue, a military coup returned it to autocracy as I was writing this chapter.

⁶Ireland, Sweden, and Japan had income below \$5,000 (at 2011 prices) in their first year of liberal democracy (using V-DEM and Maddison 2020 data).

modernization forces incumbents to share power ignores their other options. Even if violent repression is too costly in an internationally connected knowledge economy, leaders may try to avoid genuine democracy by just pretending to be democratic (Guriev & Treisman, 2019, 2022). In sum, countries do not all democratize in the same way as their income rises. Basic modernization theory predicts too much political convergence.

Standard institutionalist arguments, by contrast, predict too much divergence, at least for recent decades. Increasing returns, path dependence, and the magnification of initial differences suggest that countries should move in different directions that are hard to predict from their initial economic or social conditions. North devoted a great deal of attention to explaining the historical divergence in growth and political institutions between England and Spain (North, 1981, 1991). And for centuries, the two countries were diverging. Yet recently they have almost completely converged—in both income and institutions. Since 1950, Spain’s GDP per capita has risen from 31% to 83% of the UK’s, while the gap in V-DEM liberal democracy scores has narrowed from 65 to 3 percentage points, with Spain edging into the lead.⁷

Or, to return to the “narrow corridor,” that argument saw countries sorting into three clusters—anarchy, despotism, and democracy—with those in the former two getting stuck there. In fact, in the last 70 years, almost all countries have been moving toward higher income and greater democracy. Between 1950 and 2018, 87% saw both their GDP per capita and liberal democracy scores rise.⁸ Those that dropped into the bottom 10 percentage points of the liberal democracy scale did not usually get stuck. In fact, within 20 years, 54% had climbed out again and 10% had made it into the top half of the scale. One study that looked for the corridor using measures of state and society strength derived from V-DEM found not a divergence in three directions but a common movement up the middle: between 1855 and 2015 “nearly all countries have increased both their strength of society and their strength of the state” (Murphy & O’Reilly, 2023, 1064). In short, rather than diverging, almost all states have been heading in the long run toward higher income, stronger institutions, and more democratic politics.

Some of the criticisms of modernization theory could be addressed with minor modifications. Almost all rich autocracies have massive quantities of oil and their high income is not always matched by other aspects of modernization such as education and knowledge economy jobs, so they may not be so anomalous

⁷Income using Maddison’s (Bolt & van Zanden, 2020) data 1950–2018, liberal democracy data (*v2x_libdem*) from V-DEM 13.

⁸Again, using Maddison’s (Bolt & van Zanden, 2020) income data and V-DEM 13’s *v2x_libdem*. Of 135 countries with data available on both indicators, 117 had upward moves in both. The median change was +17 percentage points on the democracy scale and +\$9,376 (2011 prices) in per capita GDP. On the Lexical Index of Electoral Democracy, 115 countries rose during this period, compared to three countries that fell and 38 that stayed the same—25 of them because they were already capped out at the highest level in 1950.

(Ross, 2012). The cross-national correlation between income and democracy surges again if the most oil-rich states are excluded (Treisman, 2020b). Singapore, however, is both wealthy, resource-poor, and nondemocratic, so it remains an exception.⁹ In Guriev and Treisman's (2022) terminology, it is the quintessential "spin dictatorship," avoiding democracy by imitating it. The existence of poor democracies suggests that even if development does drive democratization, it is not the only cause. If other factors matter as well, we should not expect all countries to transition at the same income threshold.¹⁰

Where both perspectives fall short is in explaining the timing and mechanisms of discontinuous regime change. Most institutionalist arguments predict stability or incremental evolution. A popular move has been to assign the label "critical juncture" to periods in which institutions change rapidly (see, for example, Capoccia (2016)). But giving such periods a name does not explain what causes them or help to predict their occurrence. North et al. (2009, 219, 240) are candid about their inability to forecast transition. In the West, getting to the "doorstep conditions" required "several centuries of technological, intellectual, and institutional change," after which extending the same rights to nonelites became "easier," although not necessarily desirable from the elites' perspective. As for modernization, while a reasonably strong link emerges between income or education and democratization over 20 year periods, it is hard to see a clear relationship in a single year (Treisman, 2015).¹¹ Timing and mechanisms have been the

⁹V-DEM rated Singapore an "electoral autocracy" in 2022. Its leaders contend that it is actually a democracy. However, its ruling party, the PAP, has won more than 89% of the seats in all 13 legislative elections since independence. According to Human Rights Watch, Singapore's government "uses draconian criminal laws and civil defamation suits to harass and prosecute critical voices" and allows "little freedom of assembly" (Human Rights Watch, 2023).

¹⁰As noted, finding instruments for economic development or institutional legacies is extremely hard. Boix (2011) used four instruments for development, based on the income of other countries, weighted by trade patterns or genetic distance, and simple time trends interacted with world median income. As he acknowledges, these may fail exclusion restrictions for various reasons (for instance, trade or genetic distance could affect not just income but also migration and the flow of ideas). Acemoglu et al. (2001) used mortality rates among European settlers to instrument for inclusive institutions in the early colonial era. The adequacy of the data used is debated (Acemoglu et al., 2012; Albouy, 2012). But, in any case, the propensity for European settlement captures not just the type of institutions created but the many other things European settlers brought with them—educational traditions, particular religions, and so on (Glaeser et al., 2004). It is also unclear how institutional legacies that were fixed as of the sixteenth or seventeenth century can explain the uneven, spatially dependent, and contrasting paths by which elements of democracy unfolded across countries in the nineteenth and twentieth centuries. All this nonlinear variation is hard to predict from a historically remote state variable. On the importance of European colonization patterns, see also Gerring et al. (2022), who find a strong relationship between the location of natural harbors, European colonization, and democracy today.

¹¹And even in the longer run, there are exceptions, such as Singapore.

subject of another set of arguments focused on the short-run strategic interactions of actors.¹²

11.4 Micro-Processes

While macro-processes may explain *why* democracies emerge, a micro-perspective is more helpful in accounting for *when* and *how* they do so. Micro-approaches focus on the actors involved and how power comes to be reallocated among them. At their simplest, they explore the interactions between a state's ruler and its population. At slightly greater complexity, they allow for divisions within the ruling group and the public. Such arguments differ on two key questions: how democratization occurs and what renders it stable.

How might democratization come about? The ruler might deliberately choose to introduce popular government because changed conditions have made this optimal for him. He might be forced to share or surrender power by the population or external actors. Or he might unintentionally undermine his own position, triggering regime change, despite the availability of superior options, in what I have called “democratization by mistake” (Treisman, 2020a).

11.4.1 Why Do Rulers Democratize?

11.4.1.1 Deliberate Choice

Many arguments suppose autocratic incumbents choose to give up power because they think doing so is in their own interest. For democratization to occur, Huntington (1993, 108) wrote, “some political leaders have to want it to happen or be willing to take steps . . . that may lead to it happening.” Some refer to such scenarios as “authoritarian-led” or “elite-biased” democratization (Albertus & Menaldo, 2018; Riedl et al., 2020). Why might incumbents think democracy would benefit them?

Rich rulers might believe the alternative to be a social revolution that would destroy both their political dominance and their wealth. They may seek to forestall the revolution by committing to pay high taxes, made credible by handing power to a representative of the poor, aka the “median voter” (Acemoglu & Robinson, 2006).

¹² Another question about institutional legacies is whether particular types of nondemocracies are more prone to democratize. Scholars have debated whether autocracies that hold elections are more or less likely to transition (Edgell et al., 2018; Gandhi & Lust-Okar, 2009; Lindberg, 2009). A quick look at the data confirms that electoral autocracies democratize at a far higher rate than do closed autocracies. Panel regressions including fixed effects for country and year and controlling for logged GDP per capita suggest that the probability of a V-DEM “electoral autocracy” moving up to at least “electoral democracy” in a given year is almost ten times higher than the probability for a “closed autocracy.” Of course, holding elections might be a symptom of authoritarian fragility rather than a cause. Military dictatorships are more likely to transition to democracy than personalistic, one-party, or monarchical ones (Geddes et al., 2014).

Second, democracy might emerge through a pact the ruling elite negotiates with other domestic groups that it cannot defeat on the battlefield (Rustow, 1970; O'Donnell et al., 1986, 37–9). Third, rather than negotiate with elite rivals, incumbents might try to enlist mass support against them. In Schattschneider's (1960, 16) view of US history: "The expansion of the electorate resulted from party competition for votes." Finally, incumbents might share power with ordinary citizens to motivate them to fight off foreign attackers (Ticchi & Vindigni, 2008) or to create incentives for future governments to invest in public goods (Lizzeri & Persico, 2004).

All these motivations seem logical. But which can be observed in practice? To answer this, I examined 316 cases of democratization (under eight definitions) that occurred between 1800 and 2015, looking to see in which of these the historical details plausibly fit the premises and predictions of the deliberate choice theories just mentioned (Treisman, 2020a).¹³ All the theories examined had some empirical support—just not very much. The argument about incentives for public good provision seemed "probably" or "very probably" to explain why democratic reform occurred in 2% of cases (averaging across the eight democracy measures). The argument about elite pacts had similar explanatory power in 16%. At least one of the deliberate choice arguments "probably" or "very probably" fit the historical details in 29% of cases (Ibid, 801).

A few other possible motives were not investigated in that article. Rather than a device to commit the rich to pay more, forestalling revolution, protodemocratic institutions might help them pay less. Some see Britain's 1688 Glorious Revolution settlement, which empowered parliament vis-à-vis the Crown, as a mechanism by which property owners protected themselves against royal expropriation (North & Weingast, 1989; see also Ansell and Samuels (2014)).¹⁴ An elite might also democratize to attract settlers or capital. In the early United States, Western frontier states offered a broad franchise with such incentives in mind (Keyssar, 2009, 32). And incumbents might democratize—or at least pretend to—in response to pledges of aid or threats of punishment by foreign states or international institutions. Accounting for these might increase the scope of deliberate choice arguments a bit. Still, the limited empirical support for most such arguments suggests much past thinking may have generalized from unrepresentative examples.

¹³Methodologically, this employed "congruence analysis," a technique for checking whether historical evidence matches the observable implications of theories. I first derived observable implications of each deliberate choice argument. (For instance, if elites democratize to motivate citizens to defend the state in war, democratization should occur around the time of a war—or serious threat of one—and involve empowering young men needed to fight. One might also find this motive mentioned in speeches, correspondence, or diaries of decision-makers.) Then, for each case, I searched a range of historical sources for evidence that fit the various observable implications, considering what biases each source might conceal. This allowed me to rate the strength of historical evidence for each theory.

¹⁴Of course, such shifts of power between rulers and rich elites rarely constitute even minimal democratization. Britain's regime in 1690 remained far from democratic.

11.4.1.2 Unavoidable Revolution or Foreign Intervention

If deliberate choice plays a smaller role than sometimes thought, were incumbents forced from office by unavoidable revolutions or foreign invasions? At times such loss of control cannot be prevented, even with good information and planning.

One example is Marshal Pétain's regime in Vichy France. Pétain did not plan to give up power after 1944 nor could he have avoided military defeat by individually making wiser choices. Yet, incidences of *force majeure* turn out to be surprisingly rare. I found that only 3% of cases fit a scenario of "unavoidable democratization" (Treisman 2020a, 800). In most democratizations preceded by revolution or foreign occupation, at least one feasible counterfactual course offered the incumbent better odds of surviving as an autocrat than the course he chose. The failure to take the most promising path available is best seen as a mistake.

Consider, for instance, France's King Louis Philippe, who was forced to abdicate and flee after a mob stormed his palace in 1848, initiating the short-lived Second Republic. Far from inevitable, the protests only began after the government unwisely banned an opposition banquet—and they might well have died down had troops not killed dozens of unarmed protesters. Even then, Louis Philippe's acting prime minister, Adolphe Thiers, thought the king could restore calm by marginally broadening suffrage. Louis Philippe categorically refused—and lost his throne (Wolf, 1963, 172; Senior, 1878, 6). Many other revolutionary overthrows could probably have been prevented by avoiding provocations or making minor but timely concessions. As for democracy installed by avoidable foreign intervention, one example is the self-defeating behavior of Panama's General Noriega, who accepted Cuban military aid, allowed his troops to brutally harass US soldiers in the canal zone, and then announced that Panama was "in a state of war" with the United States. Had he not taken these actions, the odds of US invasion in 1989 would have been significantly lower (see, for example, Koster and Sánchez (1990, 359)).

11.4.1.3 Mistakes

Mistakes can, of course, combine with elements of rational optimization and domestic or external pressure.¹⁵ In the high stakes, unforgiving world of authoritarian politics, an incumbent may maneuver successfully for years, only to make a bad

¹⁵In a sense, any leader who steps down without fighting to the death could be said to have "deliberately" chosen to do so. But to answer the question why Louis Philippe abdicated in 1848 by saying that he chose to do so does not penetrate very deep into the logic of the situation. A more satisfying explanation would point to suboptimal decisions he had made earlier that undermined his control. Mistakes based on incomplete or inaccurate information can be accommodated easily by standard game theory. Those based on nonoptimization under complete information fit less easily. But the striking success of behavioral economics in incorporating heuristics and biases into rationalistic accounts of consumer behavior suggests that relaxing extreme assumptions need not always lead to a dead end.

decision at some point. Poor information, miscalculation, and cognitive biases may blind him to a more promising course.¹⁶

I found that at least one of a series of mistakes had “probably” or “very probably” influenced the outcome in about 74% of cases studied (Treisman, 2020a, 804). Rulers’ mistakes might involve interactions with society, another part of the elite, or international actors. Among the most common were avoidably alienating the army or security services (26% of cases), mishandling an election or referendum (22%), and using excessive or poorly targeted repression that catalyzed opposition (22%). In 17% of cases on average, the misstep was a major foreign policy failure that provoked foreign intervention or discredited the leader (for instance, starting and losing an avoidable war).¹⁷

Incumbent mistakes are closely related to leader change. Mistakes may cause a ruler’s overthrow, by revolution, coup, or invasion. They may also weaken the ruler’s support base—in the armed forces, security services, elites, or population—so much that he has little choice but to step down. If mistakes sometimes cause leader turnover, leader turnover can also—by promoting the inexperienced—cause mistakes. And, to complicate things further, the particular leader chosen may himself be a mistake from the point of view of his predecessor, as in Spanish dictator Francisco Franco’s sponsorship of King Juan Carlos, who later dismantled his dictatorship. As we will see in Sect. 11.5.1, leader change has been found, in some conditions, to trigger democratization.

11.4.2 What Does Democracy Do?

The deliberate choice arguments noted in Sect. 11.4.1.1 conceive of democracy in different ways. Oddly, none of these bears much resemblance to the consensus definition of modern democracy in Sect. 11.2. They have almost nothing to say about voting.

Some, echoing Aristotle, see democracy as rule by the poor (Acemoglu & Robinson, 2006). Others see it as power-sharing with the (newly) wealthy (Ansell & Samuels, 2014, 15), with other elites (Rustow, 1970), or with certain nonelites, whether poor or middle class (Lizzeri & Persico, 2004; Schattschneider, 1960). The quintessential institution of modern democracy—the free and fair election—is oddly

¹⁶O’Donnell et al. (1986, 3–4) also emphasized accidents and uncertainty. Weyland (2014) argued that bounded rationality, cognitive heuristics, and loss aversion on the part of publics and activists helped doom democratization attempts in 1848. The same biases on the part of dictators can explain why democratization sometimes succeeds. Goldstone (2023), building on Brinton and others, shows that revolutions often involve a complicated series of interdependent choice points, which makes identifying the optimal strategy and predicting the outcome in advance extremely difficult in a low information environment.

¹⁷To be a mistake, an action must have had an expected payoff less than that of some other feasible action; it is not enough merely for it to have ended badly.

absent from these accounts.¹⁸ Rule by the poor need not involve voting at all—a famous name for it is “dictatorship of the proletariat.” Elite pacts can replace elections, rather than requiring them. Some models assume an electoral setting and a Downsian equilibrium, but what drives the results is the distribution of power among actors, not the machinery of leader selection.

Why might authoritarian incumbents choose not just to shift the locus of power but to introduce regular, competitive votes for political officials? One motive might be to renew leadership periodically. But preagreed rotation or elections within just the ruling group (as occurs occasionally in one-party regimes) could accomplish that. In some settings, elected bodies might tie the ruler’s hands, encouraging investment. Yet, a council of the wealthy could protect owners of capital, without any popular voting (Boix & Svolik, 2013). The argument that seems to me most promising is that elections can reduce the danger of armed rebellion by offering potential rebels a peaceful path to power (Przeworski et al., 2015). This is, in fact, a version of the democratization as power-sharing argument—instead of dividing power in every period, it is divided stochastically over time. Note that for elections to forestall revolution, they need not be competitive or fair. When potential rebels are weak or the state is strong, incumbents can get away with highly fraudulent ones, so democracy in this view requires a weak state or strong rebels. But why would incumbents prefer to share power stochastically rather than through a simple rotation agreement? Perhaps because elections can flexibly coopt new challengers before the incumbent has identified them and assigned them positions in the rotation? These ideas await further elaboration.

11.4.3 Limitations of Micro-Accounts

Like macro-approaches, micro-accounts leave many questions unanswered. Most deliberate choice arguments offer ingenious reasons why incumbents might choose to surrender power or share it with other groups but do not explain why they would do this by means of competitive elections. They also fit poorly with the historical details of most cases. And, even when they do fit, it is often unclear why incumbents would democratize at a particular time. Why today if not yesterday?

In fact, most deliberate choice micro-arguments mentioned in Sect. 11.4.1.1 implicitly invoke macro-processes such as economic development or globalization. In Acemoglu and Robinson’s (2006) argument, elites only need democracy if transitory economic shocks that facilitate revolutionary mobilization become rare. The cost of repression must also be high and inequality not too low. All of these conditions are plausibly related to modernization, which may reduce economic

¹⁸ Unless we assume that foreign states, international institutions, or domestic revolutionaries insist on elections. Of course, elections are also held in many autocracies (Gandhi & Lust-Okar, 2009; Levitsky & Way, 2010). So democracy can be thought of as a system in which the odds of outsiders winning is above some threshold.

volatility, render citizens harder to control, and generate concentrations of wealth (Krishna & Levchenko, 2013). Modernization may also strengthen the property-owning classes, seen by others as driving democratization, or the urban, middle classes, who demand public goods instead of patronage. Economic development should increase the opportunity cost of failing to resolve a civil war by elite pact. Meanwhile, the need to coopt citizens into national defense will grow as economic development elsewhere boosts the military power of rival states.

Mistakes by incumbents seem very common in the path from dictatorship. But most mistakes do not lead to regime change. And when they do, the outcome is not always democracy. Both mistakes and unavoidable revolutions and invasions often lead to more authoritarianism or state collapse. Even revolutions that start out democratic often end up something else—as in France in 1789 or Egypt in 2011. Indeed, the effects of such disruptions clearly depend on context. And that context is, in turn, determined at the macro-level.

11.5 How These Interact

The inadequacies of both macro- and micro-arguments suggest the need to combine them. We can think of macro-processes as determining a predisposition toward one kind of regime but not necessarily prompting transition at a particular time. Micro-events may then act as triggers, destabilizing existing institutions and opening the door to change. If the macro-environment remains consistent with the previous regime type, then a similar order is restored, although perhaps under new leadership. But if the macro-environment has changed—whether through economic and social development or the evolution of new informal norms and values—then the micro-disruption can translate into regime change.

11.5.1 Conditional Modernization Theory

Along these lines, several papers have suggested a conditional version of modernization theory. A country's level of economic development establishes a predisposition for a certain level of democracy. But change occurs discontinuously when a triggering factor fires. Such factors might include economic crisis (Kennedy, 2010), violent coups or revolutions (Miller, 2012), or any leader turnover (Treisman, 2015). The importance of such triggering factors could also explain why statistical evidence of a relationship between income and democracy is stronger in panels of ten or more years than in the short run: such triggers are more likely to fire at least once over a decade than within any given year (Ibid). Temporal clustering of such triggering factors after periods of economic development could help to explain waves in regime change (Treisman, 2020b).

11.5.2 Parameters and Equilibria

How can we think about such micro-macro-combinations in game theoretic terms? An influential line of argument identifies institutions with equilibria in static, noncooperative games (Greif & Kingston, 2011; Weingast, 1997). The key idea is that institutions define patterns or rules of behavior that are rational for the individual players to follow, given what they believe others will do. In this sense, they are self-enforcing. Each regime type—democracy, authoritarianism, or subtypes of these—can be thought of as a different equilibrium. Regime change, then, presupposes the existence of multiple equilibria. Transitions occur when one equilibrium is replaced by another.¹⁹

Different equilibria may correspond to different underlying conditions, captured in models by exogenous parameters. In that case, regime change comes about when conditions consistent with one equilibrium change to values only consistent with another. For example, Boix (2003, 35) sees authoritarianism as the only possibility if inequality and the vulnerability of wealth to expropriation are both high—as in nineteenth-century Latin America, where landed elites feared that democracy would lead to the seizure of their estates. If one or both of these falls to a low level, democracy becomes the only equilibrium. In today's Latin America, the great grandchildren of wealthy landowners can diversify assets abroad and secure much of their fortunes in Miami banks, reducing the risk to them of majority rule.

However, many games that involve beliefs—and, in particular, beliefs about the types or beliefs of others—have multiple equilibria for the same parameters. Thus, multiple equilibria may exist in a single round of play. In that case, regime change can be triggered without any alteration in conditions if something causes beliefs to adjust. For instance, if protests by some citizens lead others to decrease their estimates of the government's popularity, this can set off a cascade of revolt—as occurred, for instance, in the velvet revolutions of late communist Eastern Europe (Kuran, 1991; Lohmann, 1994).

As noted in Sect. 11.3.3, when the existence of a democratic equilibrium depends on both income and some other parameters, there is no reason to assume all regime changes will occur at the same income threshold. Imagine a world in which regime type depends on income and some other factor, x (perhaps, international support for democracy). Suppose that only one equilibrium exists at each combination of income and x . Both democratic and authoritarian equilibria exist at all levels of income, but at very low income, democratic equilibria exist only at very high levels of x , while at very high income, democratic equilibria exist at all but very low levels of x . Assuming that income and x are not strongly negatively correlated, we should expect to see more democracies—but still some autocracies—as income rises (Treisman, 2020b).

¹⁹One can also think of institutions as phases in a Markov equilibrium of a dynamic game, in which different static equilibria occur in different rounds, depending on the value of some state variable.

Multiple, belief-based equilibria for the same parameters could loosen the relationship between democracy and income still further. Then, in some parameter ranges, we might see countries with the same levels of income and x but different political regimes. And, even without any change in income or x , events that change beliefs or facilitate coordination—such as leader turnover—could trigger a shift from one equilibrium to another. Spain was hardly more economically developed and socially modern in 1977 than in 1974. Yet, Generalisimo Franco's death in 1975 set off a process of adjustment that tipped the country from dictatorship to democracy.

11.6 Disaggregating Democracy

Democracy, as defined in Sect. 11.2, consists of multiple elements—elected legislatures, responsible government, universal suffrage, competitive and free and fair elections, and the secret ballot, to name a few. Much empirical work on democracy seeks to define and explain a composite of these. Yet, the individual elements may originate in different periods. They may emerge from different political battles, perhaps even involving different political actors. The attempt to capture all this in a single measure and model may be misguided.

Another approach is to disaggregate. Rokkan (1968, 2009) explored many aspects of electoral institutions in a number of European states and suggested a few hypotheses.²⁰ More recently, scholars have investigated certain subelements such as the secret ballot (Aidt & Jensen, 2017; Buchstein, 2000; Mares, 2015; Teorell et al., 2017) and the reduction of electoral fraud (Mares, 2022).

Disaggregation has certain advantages (Capoccia & Ziblatt, 2010). It allows one to sidestep the issue of what exact combination of the various elements defines the lower bound of democracy and to study instead the processes that led to particular elements emerging. The same changes on a composite indicator may in different countries reflect change on different subcomponents, with different causes. Disaggregation also allows for the exploration of institutional complementarities and path dependence in the order of emergence. And, of course, one can combine study of disaggregated elements with that of aggregate indicators when the focus is on composite trends.

11.7 Global Modernization Theory

One challenge for explanations of democratization has been to integrate domestic and international influences into a single theory. On the one hand, scholars have proposed macro- and micro-arguments about paths of domestic development,

²⁰For instance, that strong “inherited traditions of representative rule” would increase the chances of what he called “early legitimization of opposition” (Rokkan, 2009, 82).

whether economic or institutional. On the other hand, global patterns of regime change exhibit strong temporal and spatial clustering (Brinks & Coppedge, 2006; Gleditsch & Ward, 2006). How to combine these dynamics of internal evolution and international coordination?

One possible connection is via learning about economic performance. Two recent papers examine how political actors assess the growth potential of different types of political regime by observing other states. Miller (2016) found that democracy was more likely to spread when democracies had better economic performance relative to autocracies. Abramson and Montero (2020) argue that incumbents learn from their own and neighbors' experience about the relationship between democracy and economic growth and then choose regimes to maximize their survival probability. Estimating a structural model based on this view, they find it fits the data. Learning—perhaps with cognitive biases—could contribute to the clustering of regime change if decision-makers in different countries learn at the same time (Weyland, 2014).

Still, learning about economic performance is probably not the whole story. Some dictators have sought to stimulate growth by copying the greater pluralism found in the prosperous West. Yet, despite China's stunning performance in recent decades, few democratic leaders switched to Beijing-style market communism. And, as noted, regime change is often not a deliberate choice of incumbents. Among other possible mechanisms, transitions might cluster because shocks such as wars and financial crises affect whole regions, accelerating leader turnover and movement to a new political equilibrium, consistent with the country's level of development (see 11.5.1; Treisman, 2020b). Or simultaneous shocks might undermine autocracies, with the experience of neighbors then determining what comes next (Houle et al., 2016). A useful focus for future work would be to develop techniques to assess the relative evidence for different mechanisms producing spatial dependence.

Another way to combine domestic and international dynamics is to globalize modernization theory. Economic development may have powerful effects not just within individual countries but also in the world as a whole. Within states, development fuels the growth of domestic NGOs and political movements. Globally, development in a critical mass of states may generate international humanitarian networks and NGOs determined to press for political change in even relatively poor countries.²¹ A similar dynamic may characterize national and international media, which constrain antidemocratic behavior both in individual countries and, eventually, worldwide (Guriev & Treisman, 2022, Chap. 7). That could explain why some less developed countries are swept up in democratic waves, a point I return to in the next section.

²¹ Democratic breakthroughs in recent decades in Poland, Ukraine, Serbia, Chile, and Ghana all involved training, technical support, and the sharing of ideas by Western governments and international NGOs with local prodemocracy activists (Stoner et al., 2013, 16–7). Others that received little Western support—in China, Algeria, and Iran—were less successful, although the context in those cases may have been less promising in other ways. A recent comprehensive review of studies evaluating the effectiveness of Western “democratic aid” confirmed that such aid has had a positive impact on democracy building around the world (Gisselquist et al., 2021).

11.8 Democratic Breakdown

I have so far discussed mostly moves from autocracy to democracy. But what about changes in the other direction? While the proportion of democracies in the world has grown dramatically over the last two centuries, from almost none to more than 50%, there have also been periods of retrenchment.²² Huntington (1993) classified these two centuries into three “waves” of democratization, with two “reverse waves” in between. And even when the balance of change was positive, some democracies nevertheless relapsed into autocracy.

Some scholars have expressed alarm about a new reverse wave that they believe began in the 2000s or even the 1990s (Diamond, 2015; Lührmann & Lindberg, 2019). The strength of the evidence is disputed, with some detecting only a little reversion in the data, especially if one focuses on measures that require less “expert” judgment (Little & Meng, 2023; Treisman, 2023). Still, some countries have certainly become more authoritarian. At the same time, the typical way in which democracies break down has changed in recent decades. In the past, military coups dominated. More recently, the favored method has been subversion by insiders—in the current jargon, “backsliding” (e.g., Maeda (2010), Bermeo (2016), Haggard and Kaufman (2021)).²³

What might account for democratic failure? Both modernization and institutional factors may help explain why some democracies are more prone to break down. Economically developed democracies are more resilient than poor ones, as virtually all studies attest (Przeworski, 2000).²⁴ And a longer history of democracy is associated with lower reversion odds, which might reflect the self-reinforcing adaptation of informal to formal institutions (O'Donnell, 1996; Svobik, 2015).

Modernization may also explain the trend from coups to backsliding. Overt repression can work against a rural, uneducated population, but its effectiveness is lower and opportunity costs greater in modern, internationally connected economies. In such settings, citizens tend to have more prodemocratic values and greater skills and resources to resist (Inglehart & Welzel, 2005). Military and police methods are harder to combine with knowledge economy innovation than with agricultural work or industrial mass production. Such considerations may explain why generals hesitate to seize power in more developed countries (Londregan & Poole, 1990)

²² 50.3% of countries were “electoral” or “liberal” democracies in 2022, up from 4.2% in 1900 (V-DEM 13's *v2x_regime*).

²³ A quick calculation with V-DEM 13 data shows that 57% of the democratic breakdowns between 1960 and 1989 coincided with a successful coup; between 1990 and 2019, only 15% did. By breakdown, I mean a fall from “liberal” or “electoral” democracy to “electoral” or “closed” autocracy on the *v2x_regime* variable.

²⁴ This remains true of the years after 2000. In 2000–2018, the rate of (V-DEM) breakdowns among democracies with GDP per capita above \$10,000 per capita was less than a quarter of the rate among those with income of \$10,000 or less. Still, the highest income level at which breakdown has occurred has moved upward—from \$12,700 (Argentina in 1976) to \$25,600 (Hungary in 2018).

and why recent incumbents have preferred to accumulate control by stealth while pretending to be democratic.²⁵

If economic development and a long history of democracy protect against reversion, why might we now see backsliding? This is precisely because recent democratizers are much less economically developed and experienced with democracy than previous cohorts. Recall the argument of global modernization theory that, as more countries develop economically, this fuels the rise of global NGOs, international institutions, and humanitarian movements that press even less developed countries to democratize. That is exactly what happened during the “third wave” from 1974 to around 2005. As everyone from Western aid donors to Amnesty International pushed for political reform, some poorer autocracies were swept into the upsurge. As a result, median income among V-DEM “democracies” fell from almost \$18,000 in 1981 to under \$10,000 in 1995. At the same time, democracies’ average age fell. In 1977, only 8 (4%) were less than 10 years old; by 1998 there were 41 (19%). Given the greater vulnerability of poor and inexperienced democracies—many of which were only marginally democratic to begin with—an uptick in backsliding was to be expected. Regressions using only income and democracy measures from before 2000 actually overpredict the rate of reversion in the next two decades (Treisman, 2023, 18).

Still, modernization and institutional arguments leave some puzzles. While most breakdowns occur in poor countries with little democratic experience, some backsliders have been richer. As noted, Hungary had GDP per capita of almost \$26,000 and it had been democratic for 28 years when V-DEM downgraded it to autocracy in 2018. So, even if the extent of backsliding is not surprising, the identity of some backsliders is.

Recent literature suggests several other macro-explanations for democratic erosion. Some argue that growing political polarization—sharp splits over policy, ideology, and partisanship between groups—fuels the rise of autocratic leaders (Haggard & Kaufman, 2021, 5–6). Yet, the logic here is not self-evident. If incumbents believe they can monopolize control indefinitely, then dismantling liberal protections may indeed make sense. But if—as in the United States and many other supposed backsliders—incumbents expect to lose at some point, such a strategy risks strengthening the tools a rival will use against them. Rather than raise the stakes, one might expect incumbents to insure against this by reinforcing checks and balances or at least to preserve the current rules while endeavoring to win under them. Some “backsliding” may actually be attempts to cheat rather than to change the rules. Accounts do not always distinguish clearly between cheating within a system and trying to demolish it.

If polarization has increased in many of the backsliding democracies, that raises the question why.²⁶ Some point to increased economic inequality. But some regions

²⁵ Another reason to favor stealth over military coups is the effort by organizations like the OAS and AU to isolate countries where military juntas seize power (Boniface, 2002; Souaré, 2014).

²⁶ V-DEM publishes an indicator of political polarization. Yet, since reliable evidence of attitudes in past historical periods rarely exists, judgments of polarization are likely to be contaminated by coders’ knowledge of subsequent events such as civil wars or revolutions. For discussion of puzzling codings on this, see Treisman (2023, 13).

with high levels of reported backsliding such as Latin America have actually recorded recent declines in inequality (Cornia, 2014). And, although inequality features in various theories, there is little empirical link to regime type except in the case of nineteenth-century landholding inequality (Ansell & Samuels, 2014; Scheve & Stasavage, 2017).

Among institutional legacies that might facilitate backsliding, one is the ease with which the law allows leaders to impose states of emergency (SOEs). SOEs—which may allow leaders to suspend checks on their power, dissolve the legislature, and censor media—are one key tool that backsliding leaders use.²⁷ Bjørnskov and Voigt (2018) studied 193 countries between 1976 and 2007 to see what factors accompanied the imposition of an SOE. Strangely enough, they found no statistically significant relationship between the legal and political difficulty of declaring an SOE and the frequency of their use. The one institutional factor that mattered was the type of democracy: leaders in parliamentary democracies imposed fewer SOEs than those in presidential or mixed ones (Ibid, 120). However, that does not mean that parliamentary regimes experience fewer democratic breakdowns; controlling for income and the recent quality of democracy, I found no significant relationship between (V-DEM) democratic failures and the type of democracy (parliamentary, mixed, or presidential) (Treisman, 2023). If leaders in parliamentary democracies are less likely to declare a state of emergency, they apparently compensate by undermining institutions more in other ways.²⁸

Some see the international environment as less favorable to democracy recently than in the 1990s, with the United States since 2001 prioritizing the “war on terror” and autocrats in Russia and China supporting dictators and coups in Africa (Diamond, 2021, 34–39). This is plausible. But although Putin and Xi certainly welcome backsliding and have aided various undemocratic regimes, their impact on deterioration in such salient cases as Poland, Brazil, Mexico, Turkey, India, and Israel seems minimal.

As with democratization, a convincing explanation for backsliding probably requires integrating micro- with macro-factors. Among micro-processes, economic shocks seem to facilitate downward as well as upward change. Lower growth predicts reversals (e.g., Svolik, 2008) as does high inflation (Kapstein & Converse, 2008). Some see immigration as important. However, while populist politicians often use xenophobic rhetoric, such appeals need not require large actual migrant inflows to be effective nor does support for populists always translate into democratic erosion. Historical data on migration flows are limited. But using UN figures

²⁷Lührmann and Rooney (2021) show that autocratization was more likely to occur in recent decades in democracies that were under a state of emergency, but since imposing a state of emergency often *is* autocratization, it is hard to see this as a separate causal factor.

²⁸Another institutional factor that might matter is the extent of political decentralization. Theoretically, the relationship between decentralization and political freedom is indeterminate (Treisman, 2007, 194–98). While strong democratic local governments might resist central backsliding, decentralization may also entrench strong *undemocratic* forces at the subnational level (Grumbach, 2022).

on the share of immigrants in national populations, there is no significant correlation between the change in the immigrant share and change in V-DEM democracy in recent decades.²⁹ Autocratization—like democratization—can result from mistakes as well as deliberate choices. Breakdowns in the mid-twentieth century, Linz (1978, 81) wrote, were “often possible only because so many of the participants were unaware of the ultimate implications of their actions and, even more often, were mistaken in their analysis of the situation.” Some of the same mistakes that usher in democracy may also feature in its collapse—for instance, incumbents may slide down the “slippery slope” of destabilizing concessions, misapply repression, or promote unreliable subordinates.

11.9 Conclusion

In recent decades, scholars have debated whether the fundamental cause of political regime change is institutional history or economic development. The two have been presented as alternatives and sometimes pitted against each other in statistical contests).³⁰ Simultaneously, political scientists have questioned whether political change owes more to “structure” or to “agency,” again casting these as rival modes of explanation. These debates have failed to produce a consensus.

That might be because these types of explanation are not alternatives so much as partial accounts that need to be combined. Neither economic development nor institutional evolution can explain all aspects of regime change. Countries’ rates of economic development are clearly driven—in part—by differences in social and political institutions. But institutions evolve to fit their environment, which, at any moment, is characterized by the outputs of past economic development—stocks of human and financial capital, technological knowledge, and social values. Whether institutions or development is primary is more a matter of perspective, time period, and explanatory strategy than of underlying truth. Past institutional differences help to explain why subsequent paths diverged. And they have been deployed brilliantly to illuminate contrasting historical trajectories, such as those of England and Spain. But these countries have since converged in a way that reflects their parallel and interconnected economic development. Structural factors determine propensities toward particular regime types, but the concrete interactions of actors are needed

²⁹ The correlations between (a) change in immigrant share and (b) change in V-DEM’s continuous index of electoral democracy (*v2x_polyarchy*) over 20-, 10-, and 5-year intervals to 2020 were, respectively, -0.09 , -0.05 , and -0.03 . Those between change in immigrant share and change in V-DEM’s liberal democracy index (*v2x_libdem*) over 20-, 10-, and 5-year intervals were -0.11 , -0.04 , and -0.01 . Across all cases of regression from democracy to autocracy since 1995 according to V-DEM, the average change in the migrant share in the preceding 5 years was 0.0, compared to $+0.38$ in cases where democracy survived (data are from the United Nations Department of Economic and Social Affairs (2020)). More sophisticated methods might turn up an effect, but immigration seems unlikely to be a major, general determinant.

³⁰ See, for instance, Acemoglu et al. (2008, 2019); Boix and Stokes (2003); Coppedge et al. (2022).

to elucidate the mechanisms of change. Neither level of analysis has so far yielded very reliable short-term predictions.

In game theoretic terms, economic development shapes the parameters, while institutions define the rules of the game—the order of moves and the set of possible actions. Both together determine what equilibria result if self-interested actors optimize correctly. Since the payoffs of different outcomes reflect parameters, no games exist without these. Nor are there games without rules. Both payoffs and rules of the game are changed over time by different, interrelated processes. But neither makes much sense without the other.

Empirically, modern democracy has spread from a few countries to much of the world over the past two centuries. This unfolding was not a simple, linear process but involved repeated contestation over multiple institutional elements. There have been many moments of regression on one or more of these. Advances and relapses have loosely synchronized into increasingly global waves. The micro-mechanisms have mixed deliberate institutional construction and self-destructive mistakes. Reforms, revolutions, and foreign interventions all played a part. The effects of such disruptions have varied with both local and global conditions. Although scholars continue to debate which conditions matter most—inequality, polarization, economic development, the business cycle, state capacity, and global integration—there is consensus that at least some of them do matter.

If looking for a “first cause” within the interwoven chains of economic and institutional evolution no longer seems so promising, many other questions might still yield to disciplined inquiry. Why have domestic processes in so many places led to consensus on a mechanism for choosing leaders—competitive elections—that was widely rejected two centuries ago? Why have elections—at least within Europe and the Americas—tended to become more equal, direct, universal, secret, free, and fair over time? How can we integrate what we have learned about domestic sources of regime change with the accumulating evidence of global processes? Does today’s backsliding represent the predictable slowing and slight reversion after a dramatic surge, or is it something more serious? Why have authoritarian regimes shifted their emphasis from violent repression toward information manipulation, and is that trend now reversing? Answers to these will probably require combining institutional legacies and modernization in an argument that connects micro- with macro-processes, while leaving some room for both rational decisions and self-defeating error.

Acknowledgments I would like to thank Claude Ménard, Mary Shirley, and Konstantin Sonin for helpful comments.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute. *This chapter was made possible in part by a grant from Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.*

References

- Abramson, S., & Montero, S. (2020). Learning about growth and democracy. *American Political Science Review*, 114(4), 1195–1212. <https://doi.org/10.1017/S0003055420000325>
- Acemoglu, D., & Robinson, J. A. (2000). Why did the west extend the franchise? Democracy, inequality, and growth in historical perspective. *The Quarterly Journal of Economics*, 115(4), 1167–1199.
- Acemoglu, D., & Robinson, J. A. (2006). *Economic origins of dictatorship and democracy*. Cambridge University Press.
- Acemoglu, D., & Robinson, J. A. (2020). *The narrow corridor: States, societies, and the fate of liberty*. Penguin.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5), 1369–1401.
- Acemoglu, D., Johnson, S., Robinson, J. A., & Yared, P. (2008). Income and democracy. *American Economic Review*, 98(3), 808–842.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2012). The colonial origins of comparative development: An empirical investigation: Reply. *American Economic Review*, 102(6), 3077–3110.
- Acemoglu, D., Naidu, S., Restrepo, P., & Robinson, J. A. (2019). Democracy does cause growth. *Journal of Political Economy*, 127(1), 47–100.
- Aidt, T. S., & Jensen, P. S. (2017). From open to secret ballot: Vote buying and modernization. *Comparative Political Studies*, 50(5), 555–593.
- Albertus, M., & Menaldo, V. (2018). *Authoritarianism and the elite origins of democracy*. Cambridge University Press.
- Albouy, D. Y. (2012). The colonial origins of comparative development: An empirical investigation: Comment. *American Economic Review*, 102(6), 3059–3076.
- Ansell, B. W., & Samuels, D. J. (2014). *Inequality and democratization: An elite competition approach*. Cambridge University Press.
- Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. University of Michigan Press.
- Bermeo, N. (2016). On democratic backsliding. *Journal of Democracy*, 27, 5.
- Bjørnskov, C., & Voigt, S. (2018). Why do governments call a state of emergency? On the determinants of using emergency constitutions. *European Journal of Political Economy*, 54, 110–123.
- Bobbio, N. (1987). *The future of democracy: A defence of the rules of the game*. University of Minnesota Press.
- Boix, C. (2003). *Democracy and redistribution*. Cambridge University Press.
- Boix, C. (2011). Democracy, development, and the international system. *American Political Science Review*, 105(4), 809–828.
- Boix, C., & Stokes, S. C. (2003). Endogenous democratization. *World Politics*, 55(4), 517–549.
- Boix, C., & Svolik, M. W. (2013). The foundations of limited authoritarian government: Institutions, commitment, and power-sharing in dictatorships. *The Journal of Politics*, 75(2), 300–316.
- Bolt, J., & van Zanden, J. L. (2020). Maddison style estimates of the evolution of the world economy. A new 2020 update. Maddison Project Database, version 2020.
- Boniface, D. S. (2002). Is there a democratic norm in the Americas? An analysis of the Organization of American States. *Global Governance*, 8(3), 365–381. <http://www.jstor.org/stable/27800350>
- Brinks, D., & Coppedge, M. (2006). Diffusion is no illusion: Neighbor emulation in the third wave of democracy. *Comparative Political Studies*, 39(4), 463–489.
- Buchstein, H. (2000). *Öffentliche Und Geheime Stimmabgabe : Eine Wahlrechtshistorische Und Ideengeschichtliche Studie*. Nomos Verlagsgesellschaft.
- Bueno De Mesquita, B., Smith, A., Siverson, R. M., & Morrow, J. D. (2005). *The logic of political survival*. MIT Press.

- Capoccia, G. (2016). Critical junctures. In *The Oxford handbook of historical institutionalism* (pp. 89–106). Oxford University Press.
- Capoccia, G., & Ziblatt, D. (2010). The historical turn in democratization studies: A new research agenda for Europe and beyond. *Comparative Political Studies*, 43(8–9), 931–968.
- Coppedge, M., Edgell, A. B., Knutsen, C. H., & Lindberg, S. I. (2022). *Why democracies develop and decline*. Cambridge University Press.
- Cornia, G. A. (2014). *Falling inequality in Latin America: Policy changes and lessons*. Oxford University Press.
- Diamond, L. (2015). Facing up to the democratic recession. *Journal of Democracy*, 26, 141.
- Diamond, L. (2021). Democratic regression in comparative perspective: Scope, methods, and causes. *Democratization*, 28(1), 22–42.
- Edgell, A. B., Bernhard, M., Mechkova, V., Altman, D., & Lindberg, S. I. (2018). When and where do elections matter? A global test of the democratization by elections hypothesis, 1900–2010. *Democratization*, 25(3), 422–444. <https://doi.org/10.1080/13510347.2017.1369964>
- Fukuyama, F. (2014). *Political order and political decay: From the industrial revolution to the globalization of democracy*. Farrar, Strauss and Giroux.
- Gandhi, J., & Lust-Okar, E. (2009). Elections under authoritarianism. *Annual Review of Political Science*, 12, 403–422.
- Geddes, B., Wright, J., & Frantz, E. (2014). Autocratic breakdown and regime transitions: A new data set. *Perspectives on Politics*, 12(2), 313–331.
- Geddes, B., Wright, J. G., & Frantz, E. (2018). *How dictatorships work: Power, personalization, and collapse*. Cambridge University Press.
- Gerring, J., Apfeld, B., Wig, T., & Tollefsen, A. F. (2022). *The deep roots of modern democracy: Geography and the diffusion of political institutions*. Cambridge University Press.
- Gisselquist, R., Niño-Zarazúa, M., & Samarin, M. (2021). *Does aid support democracy?* (Working Paper 2021/14). WIDER.
- Glaeser, E. L., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2004). Do institutions cause growth? *Journal of Economic Growth*, 9, 271–303.
- Gleditsch, K. S., & Ward, M. D. (2006). Diffusion and the international context of democratization. *International Organization*, 60(4), 911–933.
- Goldstone, J. A. (2023). You Can't always get what you want: Why revolutionary outcomes so often diverge from revolutionary goals. *Public Choice*, 1–16.
- Greif, A., & Kingston, C. (2011). Institutions: Rules or Equilibria? In N. Schofield & G. Caballero (Eds.), *Political economy of institutions, democracy and voting* (pp. 13–43). Springer.
- Grumbach, J. (2022). *Laboratories against democracy: How National Parties Transformed State Politics* (Vol. 184). Princeton University Press.
- Gurie, S., & Treisman, D. (2019). Informational autocrats. *Journal of Economic Perspectives*, 33(4), 100–127.
- Gurie, S., & Treisman, D. (2022). *Spin dictators: The changing face of tyranny in the 21st century*. Princeton University Press.
- Haggard, S., & Kaufman, R. (2021). *Backsliding: Democratic regress in the contemporary world*. Cambridge University Press.
- Houle, C., Kayser, M. A., & Xiang, J. (2016). Diffusion or confusion? Clustered shocks and the conditional diffusion of democracy. *International Organization*, 70(4), 687–726.
- Human Rights Watch. (2023). Singapore: Events of 2022. *World Report 2023* (blog). <https://www.hrw.org/world-report/2023/country-chapters/singapore#:~:text=Keynote,-A%20New%20Model&text=The%20Singapore%20government%20uses%20draconian,is%20little%20free dom%20of%20assembly>
- Huntington, S. P. (1993). *The third wave: Democratization in the late twentieth century* (Vol. 4). University of Oklahoma Press.
- Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.

- Kapstein, E. B., & Converse, N. (2008). *The fate of young democracies*. Cambridge University Press.
- Kelsen, H. (1955). Foundations of democracy. *Ethics*, 66(1, Part 2), 1–101.
- Kennedy, R. (2010). The contradiction of modernization: A conditional model of endogenous democratization. *The Journal of Politics*, 72(3), 785–798.
- Keyssar, A. (2009). *The right to vote: The contested history of democracy in the United States*. Basic Books.
- Koster, R. M., & Sánchez, G. (1990). *In the time of the tyrants: Panama: 1968–1990*. W.W. Norton & Co.
- Krishna, P., & Levchenko, A. A. (2013). Comparative advantage, complexity, and volatility. *Journal of Economic Behavior & Organization*, 94, 314–329.
- Kuran, T. (1991). Now out of never: The element of surprise in the East European revolution of 1989. *World Politics*, 44(1), 7–48.
- Levitsky, S., & Way, L. A. (2010). *Competitive authoritarianism: Hybrid regimes after the cold war*. Cambridge University Press.
- Lindberg, S. I. (2009). *Democratization by elections: A new mode of transition*. Johns Hopkins University Press.
- Linz, J. J. (2000). *Totalitarian and authoritarian regimes*. Lynne Rienner Publishers.
- Linz, J. J. (1978). *The breakdown of democratic regimes*. Crisis, breakdown and reequilibration. Johns Hopkins University Press.
- Lipset, S. M. (1959). Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review*, 53(1), 69–105.
- Little, A., & Meng, A. (2023). Measuring democratic backsliding. *PS: Political Science and Politics*. <https://doi.org/10.7910/DVN/G2SQ6Y>
- Lizzeri, A., & Persico, N. (2004). Why did the elites extend the suffrage? Democracy and the scope of government, with an application to Britain's 'Age of Reform'. *The Quarterly Journal of Economics*, 119(2), 707–765.
- Lohmann, S. (1994). The dynamics of informational cascades: The Monday demonstrations in Leipzig, East Germany, 1989–91. *World Politics*, 47(1), 42–101.
- Londregan, J. B., & Poole, K. T. (1990). Poverty, the coup trap, and the seizure of executive power. *World Politics*, 42(2), 151–183.
- Lührmann, A., & Lindberg, S. I. (2019). A third wave of Autocratization is here: What is new about it? *Democratization*, 26(7), 1095–1113.
- Lührmann, A., & Rooney, B. (2021). Autocratization by decree: States of emergency and democratic decline. *Comparative Politics*, 53(4), 617–649.
- Maddison Project Database, version 2020. Bolt, Jutta and Jan Luiten van Zanden (2020), Maddison style estimates of the evolution of the world economy.
- Maeda, K. (2010). Two modes of democratic breakdown: A competing risks analysis of democratic durability. *The Journal of Politics*, 72(4), 1129–1143.
- Mares, I. (2015). *From open secrets to secret voting: Democratic electoral reforms and voter autonomy*. Cambridge University Press.
- Mares, I. (2022). *Protecting the ballot: How first-wave democracies ended electoral corruption*. Princeton University Press.
- Miller, M. K. (2012). Economic development, violent leader removal, and democratization. *American Journal of Political Science*, 56(4), 1002–1020.
- Miller, M. K. (2016). Democracy by example? Why democracy spreads when the world's democracies prosper. *Comparative Politics*, 49(1), 83–116.
- Munck, G. L. (2023). *Reimagining the social sciences: Knowledge, epistemic change and the study of democracy*. manuscript.
- Murphy, R., & O'Reilly, C. (2023). The expansive corridor: Testing Acemoglu and Robinson (2019). *The Journal of Development Studies*, 1–16.
- North, D. (1981). *Structure and change in economic history*. W.W. Norton & Co.

- North, D. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- North, D. C., & Weingast, B. R. (1989). Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century England. *The Journal of Economic History*, 49(4), 803–832. <http://www.jstor.org/stable/2122739>
- North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.
- O'Donnell, G. A. (1996). Illusions About Consolidation. *Journal of Democracy*, 7(2), 34–51. <https://doi.org/10.1353/jod.1996.0034>
- O'Donnell, G., Schmitter, P. C., & Whitehead, L. (Eds.). (1986). *Transitions from authoritarian rule: Comparative perspectives*. Johns Hopkins University Press.
- Przeworski, A. (2000). *Democracy and development: Political institutions and well-being in the world, 1950–1990*. Cambridge University Press.
- Przeworski, A., Rivero, G., & Xi, T. (2015). Elections as a conflict processing mechanism. *European Journal of Political Economy*, 39, 235–248.
- Riedl, R. B., Slater, D., Wong, J., & Ziblatt, D. (2020). Authoritarian-led democratization. *Annual Review of Political Science*, 23, 315–332.
- Rokkan, S. (1968). The structuring of mass politics in the smaller European democracies: A developmental typology. *Comparative Studies in Society and History*, 10(2), 173–210.
- Rokkan, S. (2009). *Citizens, elections, parties: Approaches to the comparative study of the processes of development*. ECPR Press.
- Ross, M. L. (2012). *The oil curse: How petroleum wealth shapes the development of nations*. Princeton University Press.
- Rustow, D. A. (1970). Transitions to democracy. *Comparative Politics*, 2(3), 337–363.
- Schattschneider, E. E. (1960). *The Semisovereign people*. Holt, Rine-Hart and Winston.
- Schedler, A. (2009). Electoral authoritarianism. In *The SAGE handbook of comparative politics* (pp. 381–394). Sage Publications.
- Scève, K., & Stasavage, D. (2017). Wealth inequality and democracy. *Annual Review of Political Science*, 20(1), 451–468.
- Schumpeter, J. A. (1942). *Capitalism, socialism and democracy*. Harper and Brothers.
- Senior, N. (1878). *Conversations with M. Thiers, M. Guizot, and other distinguished persons during the second empire*. Hurst & Blackett.
- Skaaning, S.-E., Gerring, J., & Bartusevičius, H. (2015). A lexical index of electoral democracy. *Comparative Political Studies*, 48(12), 1491–1525.
- Souaré, I. K. (2014). The African union as a norm entrepreneur on military coups d'état in Africa (1952–2012): An empirical assessment. *The Journal of Modern African Studies*, 52(1), 69–94.
- Stasavage, D. (2020). *The decline and rise of democracy: A global history from antiquity to today*. Princeton University Press.
- Stoner, K., Diamond, L., Girod, D., & McFaul, M. (2013). Transitional successes and failures: The international–Domestic nexus. In K. Stoner & M. McFaul (Eds.), *Transitions to democracy: A comparative perspective* (pp. 3–24). John Hopkins University Press.
- Svolik, M. W. (2008). Authoritarian reversals and democratic consolidation. *American Political Science Review*, 102(2), 153–168.
- Svolik, M. W. (2015). Which democracies will last? Coups, incumbent takeovers, and the dynamic of democratic consolidation. *British Journal of Political Science*, 45(4), 715–738.
- Teorell, J., Ziblatt, D., & Lehoucq, F. (2017). An introduction to special issue: The causes and consequences of secret ballot reform. *Comparative Political Studies*, 50(5), 531–554.
- Ticchi, D., & Vindigni, A. (2008). *War and endogenous democracy* (IZA working paper 3397).
- Treisman, D. (2007). *The architecture of government: Rethinking political decentralization*. Cambridge University Press.
- Treisman, D. (2015). Income, democracy, and leader turnover. *American Journal of Political Science*, 59(4), 927–942.

- Treisman, D. (2020a). Democracy by mistake: How the errors of autocrats trigger transitions to freer government. *American Political Science Review*, 114(3), 792–810.
- Treisman, D. (2020b). Economic development and democracy: Predispositions and triggers. *Annual Review of Political Science*, 23(1), 241–257.
- Treisman, D. (2023). How great is the current danger to democracy? Assessing the risk with historical data. *Comparative Political Studies*, 56(12), 001041402311683.
- United Nations Department of Economic and Social Affairs, Population Division. (2020). *International migrant stock 2020*. <https://www.un.org/development/desa/pd/content/international-migrant-stock>
- Weingast, B. R. (1997). The political foundations of democracy and the rule of the law. *American Political Science Review*, 91(2), 245–263.
- Weyland, K. (2014). *Making waves: Democratic contention in Europe and Latin America since the revolutions of 1848*. Cambridge University Press.
- Wolf, J. (1963). *France, 1814–1919: The rise of a Liberal-democratic society*. Harper and Row.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part III

Legal Institutions of a Market Economy

Revisiting the Many Legal Institutions that Support Contractual Commitments in a Globalized World

12

Gillian K. Hadfield and Alexander Bernier

The problems of achieving third-party enforcement of agreements via an effective judicial system . . . are only imperfectly understood and are a major dilemma in the study of institutional evolution.
North (1990)

Contents

12.1	Introduction	268
12.2	The Problem of Contractual Commitment and the Selection of Efficient Enforcement Mechanisms	270
12.3	The Role of Legal Institutions in Structuring an Effective Law of Contracts	273
12.3.1	Courts and Judges	274
12.3.2	Lawyers	278
12.3.3	Collateral Legal Rules	282
12.4	Legal Institutions in a Globalized World	285
12.4.1	Private International Law	285
12.4.2	Structuring Interforum Competition for Litigation	286
12.4.3	Legal Innovations to Support the Transnational Enforcement of Contracts ...	287
12.5	Conclusion	288
	References	289

Abstract

Neoclassical economic models assume the enforcement of contracts to be costless and automatic. In practice, litigation and adjudication subject contract

This chapter is a revised and updated version of *The Many Legal Institutions that Support Contractual Commitments* published in the first edition of this *Handbook*.

G. K. Hadfield (✉) · A. Bernier
University of Toronto, Toronto, ON, Canada
e-mail: g.hadfield@utoronto.ca; Alexander.Bernier@mail.utoronto.ca

enforcement to high transaction costs. The characteristics of the local legal environment and the structure of available legal institutions affect the transaction costs of contract enforcement. Empirical and theoretical literature are beginning to disentangle the contributions that distinct formal legal institutions make in securing contractual commitments at low cost. Well-specified private law and efficient rules of procedure contribute to low-cost enforcement. So do competitive markets for legal talent and a judiciary that has incentives to produce fair and efficient outcomes. Increasingly, institutions outside of the local jurisdiction, including predictable rules of private international law, well-structured competition between courts, and innovations in online dispute resolution, determine enforcement costs. Interdependencies between domestic and transnational legal institutions suggest that well-functioning local courts and legal professions will no longer be sufficient in guaranteeing low-cost access to contract law. Interjurisdictional cooperation and transnational institutions are needed to secure contractual commitments in a globalized world.

Keywords

Contractual commitment · Law and innovation · New Institutional Economics · Private law theory · Transaction cost economics

12.1 Introduction

The problem of enforcing agreements in exchange is at the heart of economic life and has been a central topic for economic theory in the past several decades. When economists study what goes on inside the “black box” of the firm, especially under conditions of uncertainty and asymmetric information, the role of contractual commitment in economic organization comes to the fore. Much of the theory of incentives that emerged since the 1970s depends crucially on assumptions about the enforceability of contractual mechanisms designed to align the interests of principal and agent and achieve efficient production and exchange (Laffont & Martimort, 2002).

One of the fundamental contributions of transaction cost theory and institutional economics has been to focus attention on opening the “black box” of contract enforcement, drawing attention to the institutions required to achieve effective and low-cost contract enforcement. Williamson (1985) emphasizes the obstacles to perfect complete contracting in his approach to analyzing the institutions of capitalism. North (1990) places specific emphasis on understanding the costs of third-party enforcement in his analysis of the dynamics of institutional change and the differential performance of economies across time and space. Our understanding of the critical interplay between institutions and the enforcement of contracts and economic development has been substantially advanced by the work of Greif (1989, 1993), Milgrom, North, and Weingast (1990), Greif, Milgrom, and

Weingast (1994), and others on the role of coalitions, the private law merchant, and the merchant guilds in securing the commitments necessary to facilitate long-distance trade and the commercial revolution in medieval Europe. In the early days of the commercial Internet (Greenstein, 2009), attention focused on the need to develop institutional mechanisms for secure commitment—notably for the problems of identification, security, and verification in electronic transactions that in many ways recapitulated the problems of the legal vacuum facing traders in the twelfth century (Hadfield, 2004).

The idea that the effectiveness of contract law is critical to the growth of economic activity is widespread in the literature on economic development. Study of the problems of economies making the transition from socialist to market organization at the turn of the twenty-first century ultimately focused on the role of institutions necessary to support the enforcement of contracts (Greif & Kandel, 1995; Hay & Shleifer, 1998; Murrell, 2001; Johnson et al., 2002). Acemoglu and Robinson's magisterial (2013) book on *Why Nations Fail* rested its answer squarely on the absence of inclusive institutions including, importantly, those that support private contracting. Numerous studies emerged attempting to document the strength of formal contract enforcement in different settings (McMillan & Woodruff, 2000; Hendley, 2001; Murrell & Ryterman, 2001; Lee & Meagher, 2001; Pei, 2001; Williamson, 2003; Johnson et al., 2002). The World Bank launched the Doing Business project in 2002, producing cross-country and time-series data on, among other things, the costs of enforcing contracts (Djankov et al., 2003; Besley, 2015).

It is now fairly well appreciated that it is not sufficient merely to have contract law on the books to achieve the benefits of private contracting. In this chapter, we focus on the *complexity* of the operation of law in practice, and the need for a much more detailed appreciation of the multiple legal institutions at work to make contract law effective. It explores the many legal institutions that support contractual commitments by structuring an essential environment for basic contract law and by increasing the efficacy and decreasing the cost of formal enforcement of agreements. By way of background, Sect. 12.2 sets out the basic problem of contractual commitment and places formal contract law in context as one of a range of enforcement mechanisms available. Section 12.3 then provides the principal contribution of this chapter, surveying a range of legal institutions necessary to support even simple contract law. Here we examine the many institutional structures at work in the organization of courts, the judiciary, the legal profession, enforcement services, and the process of lawmaking and legal innovation. Section 12.4 considers how public and private transnational legal institutions affect the cost of enforcing agreements with ties to multiple jurisdictions. As trade and commerce become still more globalized, legal institutions such as private international law, well-structured competition between courts, and innovative dispute resolution mechanisms—that incorporate legal technologies—are required to secure contractual commitments at low cost.

12.2 The Problem of Contractual Commitment and the Selection of Efficient Enforcement Mechanisms

The problem of contractual commitment refers to the commitment necessary to support agreed-upon exchanges that take place over time. The problem is essentially this: if I act today—invest my resources or give up other opportunities— and you take the actions we agreed on in exchange for my efforts, paying me money or providing a return service, tomorrow, how can I be sure that you will in fact do as you promise and that I will not be left having spent resources I will never recoup? The solutions to this problem can all be understood in terms of how they affect the likelihood that the first-mover in exchange will not be disappointed or exploited by the second-mover. (Of course in any exchange both parties—all parties in a multi-lateral contract—may be first or second-movers or both.) The second-mover's failure to act as agreed can stem from a number of sources: there may be an obstacle to performance, there may be a lack of information about the conditions of performance having been met, and there may be a dispute about what performance was in fact promised. These are not, however, problems of *commitment*. The problem of commitment refers to the incentives on the part of the second-mover: the failure to act stems from the divergence between the second-mover's *ex ante* incentives—the incentives that led to the agreement to act—and their *ex post* incentives—the incentives that determine their behavior after the agreement has been struck and the time for performance has arrived. (Note that a commitment problem in need of an enforcement solution only arises when a gap in incentives arises, that is, when the limits of baseline norms of trust and trustworthiness are reached.)

There are many potential enforcement mechanisms available to support agreements. The list includes self-enforcement, reputation, organization, technology, and contract law. These enforcement mechanisms in one form or another all have a common feature: they seek to bring *ex post* incentives in line with *ex ante* agreements, to produce the outcome that the second-mover promised. They differ only in how they manage this shift.

Self-enforcement mechanisms and *reputation mechanisms* change the incentives of contracting agents by changing the consequences of actions. Self-enforcement mechanisms include the posting of bonds or exchange of hostages and the termination of valuable trading relationships (particularly those in which quasirents are generated by investment in assets specific to a particular relationship) (Williamson, 1983; Klein et al., 1978; Klein & Leffler, 1981; Telser, 1981). Reputation mechanisms alter the likelihood of future transactions with potential trading partners if an agent defaults on an agreement. Greif, Milgrom, and Weingast (1994) demonstrate the potentially complex structures involved in reputation mechanisms at work in medieval Europe. Reputation mechanisms fundamentally rely on institutions that capture and disseminate information about an agent's performance to a set of potential trading partners (such as coalition members) and include such diverse structures as trademark law, trade associations, and third-party certification. Hadfield and Bozovic (2016) demonstrate that formal contracting law can support reputation

mechanisms even when contracting agents eschew use of formal contract enforcement, by coordinating beliefs about what constitutes a breach.

Organizational mechanisms change the actors involved in a decision and hence the incentives that are operative ex post. The most dramatic organizational mechanism, of course, is horizontal or vertical integration: transforming a transaction across organizational boundaries to one within organizational boundaries (Williamson, 1975). Other organizational mechanisms include delegating control over corporate oversight to third parties. Delegation can take numerous different forms. It can occur through pure private ordering, wherein purchasers seek out suppliers that have self-selected into oversight from third-party accreditation bodies. Consumer preferences or purchaser-imposed contractual requirements create the incentives for suppliers to submit to third parties' audit and accreditation schemes. Examples include the third-party administration of consumer-facing quality designating conditions of production, such as the Forest Stewardship Council's oversight of environmental standards for timber and related products (McDermott et al., 2010; Gunningham & Sinclair, 2017). It can also occur where a group of suppliers police one another regarding quality standards, to preserve the price premium derived from controlled designations or other premium-generating intangibles (Ménard, 2005). In other cases, such as global food quality regulation, public law establishes compliance with well-formulated third-party certification and audit requirements as a defense against statutory liability, stimulating producers to turn toward competitive markets for private certification and audit services (Jahn et al., 2004; Henson, 2008). Information channeling to alter the information available to an agent who may be tempted to renege on performance is another organizational mechanism. Examples include sharing information with agents about how the principal's output affects the agent's payoff structure, incentivizing the agent to exert effort in a manner that maximizes both their own returns and those of the principal or investing in learning about an agent's exerted effort, to disincentivize shirking (Holmström, 1979; Holmström, 1999; Chaigneau et al., 2018).

Technological mechanisms change the costs of actions and the capacity of an agent to act in various ways. Internet transactions, for example, depend on technological solutions such as encryption to assist e-commerce providers in committing to security procedures to protect private information and the integrity of a transaction (Hadfield, 2004). Blockchain can automate the enforcement of smart contracts and reduce information asymmetries between principals and agents, or agents and other cooperating agents, in recording of information on a role-specific decentralized ledger. This eliminates incentives for agents to withhold information from the principal or from other agents, reducing the costs of sharing or reusing such information throughout the firm (Szabo, 1997; Davidson et al., 2016; Arruñada, 2020; Goldsby & Hanisch, 2023). Peer-to-peer exchange platforms reduce incentives to behave opportunistically and lower the transaction costs of transactions with unknown counterparties. For example, Airbnb, Lyft, Fiverr, and the like translate singular transactions between sellers and purchasers, which create problems of commitment, into repeat transactions between the platform and its sell-side or the buy-side counterparties. Potential future transactions incentivize players to respect

their contractual commitments, granting single-shot transactions the incentive structure of a relational contract. These platforms provide mechanisms that reduce transaction costs, including the supply of standard-form contractual language, dispute resolution services, and guaranteed recoveries for losses incurred (Nagle et al., 2020; Akbar & Tracogna, 2022).

In its most rudimentary form, *contract law* achieves enforcement by establishing a set of rules administered by a third party—generally the state—which determines when an agreement or promise is enforceable, establishes the grounds on which a breach of the agreement will be found and sets out the consequences for breach. Formal contract enforcement backs up agreements with the third party's ability to extract penalties or issue injunctions in the event a party to a contract fails to act as promised.

What enforcement mechanisms share is their impact on a first-mover's rational beliefs, at the time the bargain is struck, about the likelihood of performance: in the presence of the mechanism the first-mover attaches a higher probability to the occurrence of the event in which they receive the value of the promised performance. Formal contract law may achieve this transformation in beliefs in two ways. First, it can alter the payoff associated with reneging on a promise and thus alter the incentives of the second-mover, even if the cost of reneging arises from transactional rather than formal penalties (Hadfield & Bozovic, 2016). Second, it can provide some guarantee of compensation or court-ordered performance in the event the second-mover fails to fully respond to the risk of legal consequences.

Enforcement mechanisms can vary in their effectiveness at transforming the rational beliefs of contracting parties about the likelihood of performance. Consider a simple agreement to accept deferred payment for goods. One enforcement mechanism could be completely effective, generating a belief in the seller that full future payment will occur with probability one. Another could be incompletely effective, increasing the seller's expectation of payment from a baseline given generalized norms of trust and trustworthiness of, say, 20–50%. Moreover, the way in which expectations of payment are increased can vary: a mechanism could increase the probability of full recovery to 0.5 or it could increase the amount of certain recovery to one-half of the amount owed.

Using an enforcement mechanism is generally costly. It may require time, information collection and/or dissemination, human capital investments, or the services of others. It may make errors. It may require technology. It may require distortions in incentives or reductions in the liquidity of assets. For an enforcement mechanism to be effective for a given agreement, these costs must not outweigh the gains achieved from increased contractual commitment in that agreement. Selecting an *efficient* enforcement mechanism, or combination of mechanisms, involves an assessment of the relative cost and efficacy of the alternatives available for a given contract in a given environment. As the minimum cost of enforcement across the range of available enforcement mechanisms increases, so too do the minimum gains that must be available from contractual commitment. Put differently, in an environment with only high-cost enforcement mechanisms, only high-value contracts and

those that are supported effectively by baseline institutions such as trust or family relationships are likely to go forward.

The problem of contractual commitment from an economy-wide perspective is thus not a discrete question of whether contracts can or cannot be enforced. Rather it is a question of the *cost* of various enforcement mechanisms and the *efficacy* with which these mechanisms improve the confidence contracting parties have in the performance of their agreements. This is where attention to the institutional environment in which enforcement mechanisms operate becomes essential to understanding an economy's relative capacity to generate economic activity and growth. A self-enforcing mechanism such as a bond requires an institutional environment that recognizes and enforces the transfer of property rights; a lower cost mechanism is achieved if the institutional environment provides, for example, the organizational and legal elements necessary to establish escrow accounts or liens. Conversely, a higher-cost mechanism is produced if, for example, legal rules override or penalize the use of the mechanism. A reputation mechanism such as a trademark requires an institutional environment that protects the integrity of the trademark as a reliable indicator of the origin (producer) of a product; this is the function of trademark law. If establishing or protecting trademarks is costly, the reputation mechanism is less effective in generating economic activity.

An understanding of the many institutions that support formal contract law is thus a prerequisite to the even more complex institutional analysis necessary to appreciate the full range of enforcement mechanisms available to contracting parties, particularly those that merge features of different mechanisms (such as relational contracts which rely both on formal contract law and self-enforcement and reputation mechanisms), to assess the relative cost and efficacy of these mechanisms in different institutional settings, and to analyze the interactions between institutions that may either increase or decrease the cost and efficacy of particular enforcement mechanisms. In what follows, we survey the wide range of institutions and considerations on which such analysis should focus.

12.3 The Role of Legal Institutions in Structuring an Effective Law of Contracts

Basic contract law seems simple enough: contracting parties designate a set of actions that each will perform (deliver goods, pay money, perform work, etc.) and the law (typically) establishes a right to marshal the coercive power of the state to extract penalties (damages, injunctions, fines, etc.) in the event the actions are not performed as promised.¹ But for even this simple, although critical, enforcement

¹Contract law could also simply perform the “classification” function, making an authoritative declaration of what constitutes a breach and then leaving a successful plaintiff to rely on other enforcement mechanisms to collect compensation. For a general discussion of this role of law, see Hadfield and Weingast (2013) [Law without the State]; for a contract-specific discussion, see Hadfield and Bozovic (2016).

mechanism to be effective and relatively low-cost, a wide array of legal institutions must exist. Contract law makes its own promises to contracting parties: it promises to be available to accurately interpret the agreement the contracting parties have made, to impartially judge the performances rendered, and to reliably implement appropriate remedies. The fulfilment of these promises, and the cost of accessing them, depends on many other legal institutions and the coordination and interaction between them.

12.3.1 Courts and Judges

Contract law requires the institutions of courts and judges; this much is plain. What is less obvious are the multiple court- and judge-related legal rules and institutions that are essential to the effective and low-cost operation of courts and the judiciary.

12.3.1.1 The Organization of Courts

Courts must be accessible at relatively low cost in geographic and linguistic terms to contracting parties. They must also be accessible in legal terms, and this is a function of the legal rules governing personal jurisdiction (who may be required to appear in court and bound by its determinations), subject matter jurisdiction (what contract issues may the court adjudicate—contracts involving the government? Contracts involving foreign entities?) and standing (who has legally recognized interests in a contract dispute for purposes of invoking the work of the court—the parties to the contract? The competition authority of the government? Third-party beneficiaries of the contract?)

Delay in the resolution of contract disputes is an important impediment to the effectiveness of contract law—not only does the value of payment decrease with time but so too does the probability of recovery diminish as the potential for assets to be dissipated increases and circumstances change to make performance more difficult and/or less valuable. Thus, courts must operate effectively as organizations for contract law to function. Effective courts require personnel and resources to administer their procedures and mechanisms by which they can establish effective and low-cost internal rules and procedures (Voigt, 2016).

Most importantly, courts must be able to perform a critical role in the coordination of contract dispute processes. They require effective systems and rules for scheduling and giving notice of hearings, trials, and other meetings requiring the coordination of court personnel, court space, parties, lawyers, and witnesses. Achieving this coordination requires tools for enforcing a schedule (such as the power to sanction failures to appear with fines or legal consequences) and for assessing what are legitimate reasons for not appearing (such as inadequate notice or conflicts in obligations). These coordination functions are distinctively different in Anglo-American systems, in which there is a culmination in a single event—namely, a trial—and civil law systems, in which evidence and legal theories are developed and explored in a series of hearings that can address evidence in a piecemeal fashion

and which are not governed by a strictly sequenced presentation of plaintiff's and then defendant's case (Merryman, 1985).

Coordination also requires effective information systems for tracking cases and reliably storing documents and evidence of the actions taken by the court (such as orders that have been issued) and the parties (such as compliance with filing deadlines and procedural obligations or the submission of motions or other requests for court action). Information systems are also important for implementing procedural and jurisdictional rules coordinating the relationship between courts over time and space: Has the matter already been adjudicated in this court? Is it currently being adjudicated in a different court, whether in another region or at another level in the court system? And information systems must also be externally accessible at low cost by those who must use them to assess the validity of legal claims, to convey requests for court action, to monitor court action, and to coordinate procedures.

The implementation of court procedures also implicates a host of auxiliary service providers. Process servers ensure that notification of parties and witnesses (through subpoenas for example) takes place as required by court rules. Notaries may be required to verify signatures on contract documents, reducing the cost of procedures of proof in the court; notaries may also play a role in the drafting of specialized contracts such as corporate by-laws. The power of the court to enforce its procedures (such as appearances by subpoenaed witnesses or the production of subpoenaed documents) also depends on the availability of enforcement services from institutions such as the police. Finally, the power of a court's order—to pay damages or deliver goods for example—depends on the institutions available to identify, seize, and/or liquidate assets. These are services that may be performed by police, court personnel, or private bailiffs. The effectiveness, cost, and integrity of these services depends then on the host of institutions that govern these service providers. Private bailiffs in some economies, for example, receive law degrees equivalent to those received by lawyers, are subject to the rules established by their professional organization and potentially subject to limits on their numbers and fees set by the government.

Finally, courts must, in fact, follow their rules and procedures in a reliable way. This requires court personnel and auxiliary service providers who possess the necessary human capital—knowledge of the rules and procedures, expertise in making judgments about scheduling and information systems, and so on—and who are motivated to act in accordance with the rules as opposed to shirking or accepting bribes. As we discuss in more detail in the context of judicial corruption below, protecting against the corruption of court personnel and auxiliary services is a function of multiple institutions, including civil service compensation systems, monitoring mechanisms, professional organizations, and penalties.

12.3.1.2 Judges

Just as the institution of a court is more than a building, so too is the institution of a “judge” more than a person or public office. The institution of judging supports contractual commitments only when judges are expected to implement contract law

accurately and reliably. Multiple legal institutions must support and complement the role of the judge.

The fundamental requirement for effective judging in contract law is the accurate and faithful application of the legal rules parties relied on when making their contracts. Judges may fail to apply contract law for many reasons: they may not know the law, they may lack the human capital necessary to apply the law, they may make mistakes, they may shirk their duties, or they may intentionally act at variance with the law because they will receive private benefits such as bribes, political influence, judicial advancement, or the satisfaction of their own policy preferences (Garoupa & Ginsburg, 2015). Many legal institutions play a role in reducing the risk of judicial failure to implement basic contract law.

The human capital of judges depends most obviously on the institutions of legal education and training. Who may become a judge? What are the educational requirements? What is in the curriculum? What resources and requirements are there for ongoing judicial education? What are the requirements for practical experience or training? More subtly, however, judges' human capital depends on the organization of their work. Are courts and case assignments organized in a way such that judges develop specialized knowledge in particular areas of law or are able to transmit specialized knowledge within the court system? What resources—such as law clerks, databases, legislative updates, legal commentaries, and libraries—are available to judges to learn legal rules and develop legal reasoning? Are the reasons for decisions reached by other courts and judges recorded and disseminated?

A judge's case-specific knowledge, of both law and facts, is also a function of the organization of legal work, much of which is fundamentally determined by legal rules of procedure and the institutions of legal practice. This is not merely a matter of information transmission but also of the incentives created by these rules and institutions for the discovery and development of information that can be used by the judge (Hadfield, 1992, 2008, 2011a).

In Anglo-American legal practice, judges are largely passive with respect to the production of case-specific information. Litigants obtain evidence, interview witnesses, research the law, and develop legal reasoning about the application of the law to the evidence. The incentives for litigants to make these investments in developing the judge's human capital are based on the legal rules governing judicial practice and the exercise of judicial power. In civil law systems, in contrast, judges play a more active role in obtaining evidence and, less often, legal principles (Merryman, 1985). These differences in the regime governing the production of evidence and legal rules and reasoning have important implications not only for the incentive for the development of the judge's human capital but also for the allocation of costs between litigants and public legal institutions: litigants may bear these costs in adversarial systems more extensively than litigants in civil law systems. Unfortunately, data on the costs of litigation are very difficult to come by and comparative assessments difficult to undertake. These institutional features however are quite likely to affect the private cost of, and hence the private reliance upon, contract law.

Judicial human capital and the institutions that determine its cultivation affect the incidence of judicial error. So do legal rules and mechanisms that audit or correct

judicial error. Appeal mechanisms are more or less effective, and more or less costly, depending on whether appeals are available as of right or only with permission of the appellate court and the extent to which appellate courts defer to the fact-finding or legal conclusions of trial courts. Moreover, appeal mechanisms depend fundamentally on the assumption of the cost of an appeal by the parties in litigation; this has important implications for the selection of cases in which judicial error may be discovered. Other review mechanisms—such as bureaucratic supervision and auditing—alter the selection of cases for higher review, while also shifting the cost from private litigants to the public civil service.

In developing economies, substantial attention has been paid to the problem of “error” caused by judicial corruption, that is, the risk that judges will decide and manage cases not on the basis of legal rules and evidence but on the basis of personal rewards in the form of bribes (Bardhan, 1997; World Bank, 2000). Although selection mechanisms for judges are obviously important, corruption is a complex institutional phenomenon, not merely a failure of personal ethics (Rose-Ackerman & Palifka, 2016; Fisman & Golden, 2017). Compensation systems for judges play a role in determining judicial incentives to accept bribes: judges whose incomes have in the past depended on supplementing official salaries with payments collected directly from litigants and lawyers—whether legal or illegal—may be embedded in economic and social circumstances (where they live, the obligations they have taken on, the standard of living or social status they enjoy) that make for powerful incentives to continue accepting payments. But the relationship between compensation and corruption is complex: merely increasing salaries is insufficient in the absence of institutions that promote judicial accountability, such as independent public prosecutors, and some, but not complete, judicial independence (Ríos-Figueroa, 2012).

The incidence of judicial bribery also depends on the institutions that generate the incentives for litigants to offer bribes. Among these are some of the legal institutions we have already discussed, namely, the rules and procedures in courts that may result in substantial delays in obtaining court action or significant failures in scheduling, case tracking, or information management. Similar failures in other institutions that generate important pieces of evidence in contract disputes—such as title registries or banks—may also contribute to judicial bribery as judges are offered payments to overcome these obstacles. If required documents are unavailable or costly to obtain, litigants will have an incentive to offer bribes to induce judges to accept faulty documents or proceed with inadequate evidence. This is an incentive that can face litigants who seek justified outcomes as well as those who seek unjustified outcomes. The frequency with which bribes are offered, the belief that any outcome—right or wrong—cannot be obtained without them, and the difficulty of distinguishing between those who offer payments to bring outcomes closer to the one contract law would achieve under full information and those who offer payments to distort outcomes—all of these factors contribute to the incidence of judicial bribery and the loss of integrity in contract law.

Corruption, whether in developing or advanced market economies, can also take the form of distortions in judicial decision-making caused by political influence and the pursuit of private policy preferences. The institutions of judicial selection and

appointment (are judges elected? selected from those with expertise through a civil service process?) and removal (do judges have life tenure? can they be removed by the electorate? by administrators? politicians?) influence the likelihood that judges act on the basis of the legal rules contracting parties expect.

Finally, efforts to control judicial failures to apply contract law accurately and faithfully depend critically on institutions for detecting and sanctioning failures. Legal institutions determine whether legal decisions and reasons are written and to whom they are disseminated. Errors or corruption in written decisions made available only to the parties are less likely to be detected than are errors or corruption in decisions made available more generally. Are other judges aware of how individual judges are determining cases? Are lawyers and litigants in future cases? Are legislators and administrators? The general public? The publication of legal commentary by lawyers, law professors, and other legal experts—made available to judges and those who appoint judges—also serves the function of supporting the accuracy and fidelity of judging to the announced rules and principles of law.

12.3.2 Lawyers

Although lawyers are often derided as mere clogs on the operation of legal rules and courts, the institutions organizing the training, selection, governance, compensation, and incentives of lawyers in fact are fundamental determinants of the cost and efficacy of contract law. Lawyers, however, play a critical role in connecting litigants with law and courts and in the process of legal development (Hadfield, 2007). The institutions governing the production and allocation of legal services play a crucial role in determining the cost of accessing both contract law and the many other laws (evidence, procedure, judicial selection and conduct, etc.) on which contract law depends and on the substance of law as it develops through precedent, legislation, regulation, and practice (Hadfield, 2022; Hadfield & Heine, 2016).

Legal services can be generally thought of as falling into two types: the provision of information and advice about law and legal institutions and representation before legal bodies such as courts. The cost and quality of legal services—and hence the cost and value of relying on contract law as an enforcement mechanism—is fundamentally dependent on the legal institutions governing who may provide legal services and when legal services must be used in order to make use of law.

At one extreme, we can imagine legal services being treated no differently than any other service in a market economy: supplied by private actors at a price solely determined by market conditions. Under this pure market model, there would be no restrictions on who could give legal advice, draft legal documents, or appear in court to act on behalf of someone else. At the other extreme would be legal services provided exclusively as a public good: supplied by public actors selected by government officials and paid out of the public purse.

The pure public good model is generally rejected in a democratic regime to ensure that lawyers independent from the government are available to enforce laws restricting government power. The pure market model is rare but not unheard of:

in England and Wales, for example, with few exceptions (immigration and asylum advice, conveyancing and probate matters), anyone may give legal advice. Even under a pure market model, however, legal institutions will influence the cost and quality of legal services. Most importantly, the complexity of legal rules and procedures—and hence the necessity of specialized investments in human capital—will affect both the underlying cost of the service and the competitiveness of the market for these services (Hadfield, 2001).

In most market democracies, the organization of legal services is a mix of market and public good mechanisms, and thus there are many legal institutions that influence the cost and quality of legal services.

12.3.2.1 Legal Education

For lawyers to play an effective role in reducing the cost and increasing the efficacy of contractual commitments, lawyers must first know the relevant law. The extent of this need reflects the complexity of law and hence the demand for expertise. Legal rules prohibiting those without a law degree from practicing law serve the goal of ensuring that suppliers of legal advice and representation know the law, but they also restrict the supply of legal services and create a role for degree-granting institutions and their governance structures in determining the conditions of supply. Shepherd and Shepherd (1999), for example, argue that the cost of legal services in the United States is significantly affected by monopoly power exercised by the American Bar Association, which accredits law schools. The institutions governing legal education influence curriculum as well, with implications for the cost of human capital, the value of legal services, and access to justice. Critically, our institutions of legal education impact the capacity for law to evolve in tandem with economic innovation and technology (Hadfield, 2011b, 2017).

12.3.2.2 Professional Organizations

In most advanced market democracies, legal services are organized as a self-governing profession meaning that the state delegates to one or more professional organizations the authority to regulate the conduct of its members. This institutional structure has enormous implications for the cost and quality of legal services. The profession establishes controls over who may practice law by establishing the standards and procedures governing admission to the profession and continued authorization to practice.

Professional control over the supply of lawyers (through both initial admission to practice and suspension of the right to practice in the event of failures of competence, honesty or loyalty) is potentially an important institutional instrument for increasing the value of legal services to contracting parties. Professional control, however, also creates the risk of rent-seeking by the profession as a whole (Shaked & Sutton, 1981). Historically bar associations have, in the name of quality control, played a tremendous role in structuring the market for lawyers: restricting advertising, establishing minimum fees, requiring that legal services firms be organized as partnerships or sole proprietorships and not limited liability corporations, prohibiting the practice of law in conjunction with the practice of other professions such as

accounting or business consulting, outlawing the selling of shares (and thus risk allocation and investment) in legal outcomes, preventing the use of contingency fees, and so on. The institution of the self-governing profession thus plays an essential role in structuring the determinants of the cost and quality of legal services and hence the cost and efficacy and availability of contract law as an enforcement mechanism (Hadfield, 2022).

Numerous aspects of the organization of the legal profession influence the development of legal human capital and specialization. Professional organizations may promote the development and sharing of legal human capital through continuing legal education requirements, the organization of professional meetings and seminars, and the publication of legal reports, bulletins, newsletters, and so on. But they also may inhibit investments in and diffusion of legal human capital through organizational restrictions on the practice of law. For example, some jurisdictions impose explicit prohibitions on specialization, mandatory representation requirements that penalize a lawyer who is unable to serve a general clientele, or requirements that lawyers practice in sole proprietorships or partnerships that restrict the size of a law firm or sources of capital.

These factors take on special significance in the context of contract enforcement. Contracts are essentially products designed by lawyers. As Gilson (1984) emphasizes, lawyers in advanced market economies are “transaction cost engineers”: they assist in the development of transaction-cost reducing contractual provisions. They do this in conversation with contract law, establishing the legal meaning of the provisions they invent and that their clients implement. Whether lawyers, on net, reduce or increase the transaction costs of contracting is still an open empirical question; as Coates (2001) found, variations in corporate contracts may be best explained by “agency-costs between owner-managers and their lawyers” rather than optimal transaction-cost design.

12.3.2.3 Courts and Judicial Oversight of Lawyers

Courts and judges influence the cost of legal services by establishing the incentives facing lawyers in the day-to-day practice of their profession. An essential determinant of the cost of legal services is the coordination of the activities of lawyers, witnesses, parties, court personnel, and judges. Low-cost litigation requires, for example, documents to be exchanged when expected, evidence to be presented when the opportunity for response and cross-examination is also made available, and attendance at hearings by those required to resolve a matter. Courts authorized to sanction lawyers for failure to comply with scheduling orders appear at hearings or present the evidence necessary to decide a matter are able to control the time and hence expense of litigation more effectively than those that do not. Sanctions can include penalties expressly directed at the lawyer—such as fines or disbarment—and penalties that indirectly penalize the lawyer by imposing a loss on the lawyer’s client, the authority to enter a default judgment or dismiss an action to the detriment of a party whose lawyer fails to attend a hearing or to present evidence, are powerful weapons for courts and powerful incentives for lawyers.

Courts may also play a direct role in establishing the legal fees earned by lawyers. Courts may be empowered to award legal fees to litigating parties as a routine matter, such as under the “British” rule awarding legal fees to a prevailing party, either based on an assessment of reasonable rates by the court or based on actual expenditures. Even under the “American” rule, in which parties routinely bear their own legal fees whether they prevail in litigation or not, courts may be involved in assessing legal fees when litigation takes place under a statute or under a contract that provides for fee-shifting. Finally, courts’ management and scheduling procedures can have important, and sometimes unexpected, indirect effects on legal fees.

12.3.2.4 Norms and Practices of Judicial Reasoning

Legal practice and the organization of legal services are also significantly affected by the norms and practices of legal and judicial reasoning. Much of what happens in a court depends not on formal legal rules about procedure but the practical way in which a judge manages a case. If a judge relies on adversarial presentation of evidence and legal argument, this requires and induces investments in legal human capital by lawyers. If a judge is attentive to the decisions of other judges, this has significant implications for the cost of what lawyers do and the investments that they make in acquiring information from precedent. It also spurs the development of services to help reduce the cost of these investments, such as case reporters, journals, and bulletins.

Informal norms of legal reasoning also influence the cost of legal services (Hadfield, 2001). Legal reasoning rewards increasingly sophisticated argument about the contours of legal categories and concepts. In systems in which caselaw generates the legal principles applied in future cases, this approach to legal decision-making can generate legal complexity and ambiguity. Increased complexity and ambiguity directly raise the human capital investments necessary to provide legal services. Indirectly, more complex law contributes to the imperfect nature of competition in the market for lawyers.

12.3.2.5 Direct Government Regulation and Service Provision

Finally, the cost and quality of legal services may be influenced by the institutions of direct regulation and service provision by government. As mentioned previously, the pure public good model is generally rejected in market democracies in light of the perceived need for an independent legal profession capable of challenging government action. Still, particularly in the context of contract law where the goals are primarily focused on structuring an effective market system, there are numerous public institutions that may be involved directly or indirectly in the provision of legal services.

Government-employed lawyers may provide legal services; this is frequently the case for criminal defense work. Legal services in the context of contract law, however, may be more likely to be afforded by government funding of private lawyers through legal aid mechanisms providing assistance to lower income-contracting parties, such as consumers, employees, or small business operators. Legal services that support contracting may also be provided directly by government

agencies in the form of legal information and/or dispute resolution services: consumer complaint bureaus, labor tribunals, motor vehicle arbitration panels, and so on. Government subsidies of legal education also have an impact on the supply of legal services to contracting parties.

12.3.3 Collateral Legal Rules

Lawyers like to say that “the law is a seamless web.” They mean by this that it is impossible to deal with one legal issue—such as the enforceability of a contract—without coming within the purview of a host of other legal rules: procedural rules, property laws, economic regulations, principles of legal rule development, and so on. In this section, we examine particular sets of collateral legal rules and doctrines that influence the efficacy and cost of particular instances of contract law.

12.3.3.1 Procedural Laws

Consider the basic problem of ensuring that contract law is applied on the basis of accurately determined facts: what was promised, what was performed, and what loss was caused by a breach. Ensuring accuracy in factual determinations requires a host of procedural rules to be in place to answer questions such as the following: Who can determine facts: judges, juries, administrators, specially appointed referees, and private evaluators? What evidence is admissible? Are parties obligated to produce documents or witnesses sought by the opposing side or the judge? What are the penalties for presenting false testimony or fraudulent documents in court? What third parties can be required to present documents and testimony in court? What rules are in place to ensure that the process of fact discovery takes place in a reasonable amount of time and is coordinated at low cost?

Contract law must also be reliably applied: as an institution, it must credibly commit to apply its announced rules and procedures. This commitment depends on legal rules regulating the exercise of judicial powers. In what cases *may* a court apply contract law? In what cases *must* it apply contract law—can parties avoid contract law by pleading their case under some other set of rules such as tort law? When do particular courts (local courts, specialized courts, etc.) have jurisdiction to decide a particular contract dispute (such as one involving foreign parties)? Does the court have the authority to order parties (governments? foreign firms? state-owned enterprises?) to perform a contract? What if a judge simply ignores the law: what recourse of appeal or complaint is available?

Procedural and structural rules such as those governing evidence, discovery, jurisdiction, and so on have an impact not only on the efficacy of contract law in theory but also, very importantly, in practice. These rules play a fundamental role in determining the transaction costs of using contract law as an enforcement mechanism. Extensive pretrial discovery processes, while potentially promoting increased accuracy in fact-finding, may also give rise to costly strategic behavior and delays. A major legislative effort to promote the use of pretrial judicial management

techniques to overcome delays and reduce costs in US courts, for example, significantly *increased* litigation expenditures (Kakalik et al., 1996).

12.3.3.2 Laws Governing the Contracting Environment

Basic contract law also relies on laws governing the environment in which contracting takes place, laws that regulate information and bargaining conditions. For example, the integrity of contractual exchange relies on fraud law to punish those who have positively misled their contracting partners into a contractual relationship and who are best (from an efficiency point of view) deterred by fines and the threat of criminal punishment. This is particularly the case when contract law is limited to the awarding of damages, a remedy that may be inadequate to deter deliberate fraud and that may be no deterrence at all against those who have no assets. Laws requiring truth in advertising and other consumer protection measures support the creation of contracts that reflect more closely the deals that those exchanging contractual commitments prefer under full information (Hadfield et al., 1998).

Legal rules governing the allocation of assets to satisfy contractual obligations (particularly debt obligations) also have an important impact on the efficacy of contract law. Bankruptcy law supports the credibility of commitments by providing a basis for committing to the order in which various creditors (including those who are entitled to collect payments or damages under contracts for goods and services) may lay claim to the assets of an insolvent contracting partner. The laws enabling and governing transactions secured by collateral, deposits, bonds, and so on provide mechanisms for increasing the effectiveness of a contractual promise by reducing the cost of enforcement and/or the risk of unsatisfactory court adjudication and orders. Hansmann and Kraakman (2000), for example, illuminate the role of corporate law in allowing the owners of a firm to partition assets and make them available only to the firm as an entity (and not the owners themselves) for purposes of securing the contractual commitments of the firm.

The institutional nature of contract law also varies across different types of contract. In many cases, particular types of contracts are regulated by statutes addressed to the particular contracting environment. Contract law in civil code systems, for example, tends to separately regulate particular types of contracts: sale contracts, credit contracts, transportation contracts, agency contracts, and so on. Even in common law systems, however, with an overarching law of contract applicable to any bargained-for exchange, there are numerous laws and regulations specific to particular types of contracts and particular contracting relationships, often with a view to consumer protection and balancing perceived inequalities in bargaining power that may disrupt the efficiency of a bargained-for exchange. Insurance contracts, for example, are often heavily regulated, both with respect to terms and allowed rates, to achieve goals of efficiency and fairness. Franchise contracts may be subject to state regulation—sometimes specifically at the level of industry as is the case in the USA with respect to automobile dealerships and gasoline service stations—in response to concerns about defects in the bargaining process or the judicial interpretation of these contracts (Hadfield, 1990). Labor law

such as the US National Labor Relations Act regulates the process of collective bargaining and contract negotiation between unions and management. Other employment statutes may regulate contracts between nonunionized workers and their employers, guaranteeing minimum notice periods for dismissal for example. Consumer credit contracts are frequently subject to regulations governing required disclosures and terms; in some settings, consumer contracts are voidable to protect against overreaching sales efforts, such as in the door-to-door sales context.

Laws such as these, which regulate the environment in which specific types of contracts are negotiated, may reduce the costs of contracting by reducing the need for more costly self-protective and enforcement measures, such as costly avoidance, costly negotiation, or costly legal interpretation. They may also increase the costs of contracting and/or decrease the efficacy of contractual enforcement mechanisms, when they substitute other (public) goals—such as redistribution or political equality—for the efficiency goals of private ordering. They thus constitute an important part of the richer institutional setting in which “simple” contracts are enforced.

More complex contracts are also dependent on the features of a richer institutional environment to achieve effective and low-cost enforcement. The corporation as a “nexus of contracts” is heavily dependent on the law of corporations supplying rules governing, for example, the duties of corporate officers, shareholder rights, and the potential to achieve managerial change through takeovers. Contracts for cooperative business endeavors, such as partnerships, agency agreements, and joint ventures, are frequently supported by detailed legal rules governing when these relationships are established, what rights and obligations they confer on the participants, what activities they may undertake, how profits will be shared, and how they may be dissolved. Although much of this law is supplied as default rules in Anglo-American systems—meaning contracting parties can substitute their own privately tailored terms for the statutory terms—the development of specialized legal human capital, regularized interpretations, customized procedures, and so on have an important impact on the cost and efficacy of contract law in these settings.

Often, parties cannot draft a contract that fully specifies the obligations that they wish to create in all possible contingencies (a complete contract). Complete contracting is impracticable where the features of a good or service are too complex to carefully delineate, when conditions are likely to change making it efficient to adapt performance over time, or when the value of a contractual relationship depends on the delegation of roles and responsibilities in making future choices, and it is difficult to judge *ex post* whether the choices made were efficient or self-serving. Because parties leave their contracts incomplete for good reason, the cost and the efficacy of contract enforcement depend on access to legal institutions capable of, and willing to, fill gaps and interpret the incomplete contracts that parties write (Goetz & Scott, 1981; Williamson, 1985; Macneil, 1985). A judicial approach to interpreting vague and incomplete contracts that attempts to identify the obligation that the parties would have created had they anticipated a particular contingency may promote more efficient contracting; an approach that penalizes parties for failing to divulge information about the contingency in initial contract negotiations may produce lower-cost, effective, contracting (Ayres & Gertner, 1989).

12.4 Legal Institutions in a Globalized World

We have addressed how courts and the legal profession secure contractual commitments domestically. Since this chapter was first published, contractual agreements have become increasingly transnational and many have migrated to the online world (Hadfield, 2017), bearing diminished ties to any one specific jurisdiction. To estimate the cost of using contract law, contracting parties must increasingly consider how legal institutions in other localities would treat their agreements. In this section, we discuss how (i) private international law, (ii) competition between courts to attract litigation, and (iii) innovative providers of online dispute resolution services affect the costs and reliability of using contract law to enforce contracts.

12.4.1 Private International Law

Private international law (PIL) comprises choice of law, choice of forum, and the international enforcement of foreign judgments. Choice of law determines which nation's laws are used to interpret private law disputes with ties to multiple jurisdictions. Choice of forum determines whether the court in which a party has brought suit bears sufficient ties to the dispute to adjudicate on the claim. The international enforcement of foreign judgments defines the preconditions that foreign judgments must satisfy for domestic courts to enforce them.

The clarity of private international law rules, and their consistency across jurisdictions, influences the transaction costs of contract enforcement everywhere. If outcomes emerging from domestic courts applying PIL generate conflict, are unpredictable, or are inconsistently applied, contracting parties will need to price in the costs of additional litigation, potential duplicate proceedings, and comparative legal research (Grossi, 2011). PIL litigation increases costs for litigants but generates no additional value for them—its outcomes are purely redistributive (Spiegelman, 2006). Compatible PIL rules are therefore a “global public good,” reducing the costs of enforcing contracts that span multiple jurisdictions. One might assume that private international law would have been standardized worldwide, as all contracting parties and contract litigants stand to benefit from such standardization (through lower transactions costs). However, there are rational reasons for why this has not occurred. Litigants from a specific jurisdiction might more often find themselves to be plaintiffs, or else defendants, in litigation with transnational ties, prompting the courts or legislatures in that specific jurisdiction to adopt default rules that favor the interests of domestic actors. This can happen despite the chosen rules deviating from global standards in a manner that increases transaction costs for all cocontracting parties (Herings & Kanning, 2008). Harmonization can produce competitive advantages for legal professionals in the jurisdictions that first generated the harmonized rule that is adopted, both because it might better protect the interests of its parent jurisdiction and because local legal professionals will be better versed in its interpretation and navigation, enlarging the market for their services (Brand, 2009). Jurisdictions might therefore avoid harmonization in a bid to promote their own local rules as the eventual standard.

12.4.2 Structuring Interforum Competition for Litigation

For contract litigants that limit their horizons to the home jurisdiction in which their bargain was struck, it is the features of the local legal environment that determine the costs of using contract law to secure contractual commitments. But when contracting happens in multiple jurisdictions, the features of all available courts (and private dispute resolution fora) must be considered in estimating contract enforcement costs. Because litigants choose the most advantageous forum available, courts can compete with one another to attract litigants. Contract, corporate law, intellectual property, and bankruptcy disputes have stimulated explicit competition on the part of public courts and private arbitrators to attract cases (Romano, 1993; Coyle, 2011; Anderson, 2014; Casey & Macey, 2020).

Interforum competition can be beneficial. It can commoditize access to legal services and drive down the costs of litigation. It eliminates domestic court monopolies on dispute resolution. Competition thus reduces the prospect for administrative delays, a lack of dispute-relevant legal knowledge on the part of adjudicators, and lengthy or high-cost case management and trial processes to jeopardize litigant access to efficient adjudication.

The challenge is to benefit from competition while reducing the potential for courts to engage in self-dealing. This has been shown to occur in areas where one class of disputants with shared interests selects the forum of a dispute unilaterally. Courts can favor specific litigants or interest groups to attract future litigation from them. For example, courts that hold themselves out as magnets for patent litigation are often accused of structuring patent law and its supporting institutions to benefit patent holders (Anderson, 2014). Because patent holders are the plaintiffs in patent litigation, courts can adopt rules of case management, civil procedure, and judge assignment, among others, to incentivize patentees to select their forum over others (Anderson, 2014). Courts' incentives to do this include stimulating the flourishing of local legal markets or enhancing the reputation and prestige of the court and its adjudicators (Coyle, 2011; Anderson, 2014; Klerman & Reilly, 2015).

Numerous mechanisms can help ensure that competition to attract litigants stimulates an interforum "race-to-the-top." Forum selection rules that require litigants to choose their forum through mutual agreement or that require the forum to be determined prior to a dispute arising dissuade adjudicators from privileging the interests of initiating plaintiffs or repeat players (Rasmussen & Thomas, 1999; Casey & Macey, 2020). Organizing courts into federations enables federation members to deter opportunistic behavior on the part of other members through the vertical or horizontal coordination of sanctions. Appellate review can be used to overturn—and to inflict reputational costs on—lower courts that attempt to create jurisprudential rules that favor the targeted clientele of their adjudication services (Anderson, 2014). Courts that form part of a federation could together refuse to enforce judgments from those of its members that fail to offer basic rule of law guarantees (e.g., through the application of local rules on the enforcement of foreign judgments) (Hadfield & Weingast, 2013).

The rules that determine which bodies are permitted or required to hold themselves out as adjudicative forums also affect the consequences of interforum competition on litigants (Landes & Posner, 1979; Barendrecht & de Vries, 2005; Ware, 2013). Enabling public courts, private mediators or arbitrators, and private courts to compete for litigants creates the potential for differentiation in the offerings of each such service-provider, without jeopardizing access to conventional domestic courts provisioned as a public good. This also justifies greater specialization on the part of courts, as their offerings become tailored to the needs of specific categories of litigants, rather than each being structured for a generalist audience. Presently, commercial courts and arbitral bodies demonstrate differentiation in their geographic or jurisdictional ties, industrial domain knowledge, and area of legal expertise. Each such court appeals to groups of litigants with unique needs (Godwin et al., 2017). Last, arbitrators, special-purpose public courts, or private courts can experiment with procedural or case management innovations, technological offerings, and other deviations from standard courts practices without jeopardizing access to justice or the good functioning of public courts. Successful innovations will proliferate to other courts, stimulating rapid iteration that is not possible if adjudication services were provided as a uniform government-supplied public good (Barendrecht, 2012).

12.4.3 Legal Innovations to Support the Transnational Enforcement of Contracts

The rise of the platform economy has produced innovation in how contract disputes are heard and their outcomes enforced. In the early years of the commercial Internet, large platforms discovered the economic value of bundling their service with reliable, neutral, and cost-effective dispute resolution. eBay, an early entrant into online marketplaces, developed an online dispute resolution mechanism that has fostered widespread developments, including use for business-to-business contract disputes (Hadfield, 2004, 2002; Rule, 2020). The transaction costs of using online dispute resolution (ODR) mechanisms are often low because such mechanisms can be operated in a scalable manner across all of the different jurisdictions in which the platform-provider offers its services (Shah, 2004). Technological mechanisms—such as blockchain-based transfer of assets held on-chain or the use of encryption to preserve the confidentiality of information from those other than its intended recipients—are self-enforcing (barring technological failure) and scalable (Schmitz & Rule, 2019).

Innovations in dispute resolution and contract enforcement can take root without deliberate public institutional support—blockchain solutions in particular are driven by efforts to provide transactional infrastructure entirely outside public contracting and currencies. But to flourish, they require public institutions that at least do not prohibit their use or vitiate their resolutions. In an ideal world, public institutions support the development of online contracting technologies by creating frameworks that validate their resolutions; connect them to other publicly provided or regulated

systems, such as real property and banks; and regulate their providers to protect against fraud (Davidson et al., 2016; Tapscott & Tapscott, 2016).

The institutions that determine how domestic contract law and domestic courts will treat the decisions that ODR platforms produce affect the transaction costs of using either contract law or ODR platforms. This requires cost-effective mechanisms for deciding whether and according to what conditions the decisions that ODR platforms generate are treated as outcomes with legal significance and what balance of deference and review to accord to those decisions (Schulz, 2004). Numerous options are available to establish how law treats ODR outcomes. The default, costly, method is for producers or users to rely on legal services and litigation to enforce or contest the results that online dispute resolution platforms produce. The costs of doing so can be reduced where domestic regulators or producers of private law issue soft-law guidance that helps guide this analysis. Regulatory sandboxes, licensing bodies, and formal technical requirements could instead be implemented that lower the costs of determining the conformity of online dispute resolution outputs with domestic contract law (Schulz, 2004; Mania, 2015).

If innovative dispute resolution services operate in parallel with formal legal institutions, without a mechanism to coordinate their outcomes, this can increase the costs of using either mechanism to secure contractual commitments—users must price-in the costs of resolving conflicts between them. Furthermore, if online dispute resolution were to emerge as a lower-cost mechanism for securing contractual commitments than contract law, this could arrest central functions of court adjudication—such as that of producing public classifications of lawful and nonlawful behavior and developing precedent that further specifies law (Landes & Posner, 1979; Hadfield & Weingast, 2013).

12.5 Conclusion

In this chapter, we have documented a wide range of institutions that play a role in supporting basic contract law, potentially contributing to the cost and efficacy of this method of enforcing agreed-upon exchange. What we most need to know, however, is which of these institutions matter and how in a given environment. The challenging aspect of studying contractual commitment is the fluidity and adaptability of contracting relationships. Contracting parties have available to them a wide array of enforcement mechanisms and an even wider array of mixed mechanisms such as relational contracts that combine features of formal contracting with reputation and self-enforcement and, increasingly, private contracting platforms and technologies. These enforcement mechanisms vary in cost and effectiveness. Much empirical and theoretical research has addressed the optimal structure of domestic court systems and the contribution of distinct formal and informal institutions in enabling courts to act as a mechanism for securing credible commitments. Limited literature has started to address the role of transnational institutions in securing contractual commitments, as well as the interdependencies between formal, domestic legal institutions, and their public or private counterparts in other jurisdictions, or that operate across

multiple jurisdictions. For this reason, institutional economists need to explore in far greater detail than we have to date the wide variety of specific institutions that support contractual exchange (Coase, 1960).

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute

References

- Akbar, Y. H., & Tracogna, A. (2022). The digital economy and the growth dynamics of sharing platforms: A transaction cost economics assessment. *Journal of Digital Economy*, 1(3), 209–226.
- Anderson, J. J. (2014). Court competition for patent cases. *University of Pennsylvania Law Review*, 163, 631–698.
- Arruñada, B. (2020). Prospects of blockchain in contract and property. *European Property Law Journal*, 8, 231–259.
- Ayres, I., & Gertner, R. (1989). Filling gaps in incomplete contracts: An economic theory of default rules. *Yale Law Journal*, 99, 87–130.
- Bardhan, P. (1997). Corruption and development: A review of issues. *Journal of Economic Literature*, 35, 1320–1346.
- Barendrecht, M. (2012). Courts, competition and innovation. *Revista Forumul Judecatorilor*, 7, 44.
- Barendrecht, M., & de Vries, B. R. (2005). Fitting the forum to the fuss with sticky defaults: Failure in the market for dispute resolution services. *Cardozo Journal of Conflict Resolution*, 7(1), 83–118.
- Besley, T. (2015). Law, regulation, and the business climate: The nature and influence of the World Bank doing business project. *Journal of Economic Perspectives*, 29, 99–120.
- Brand, R. A. (2009). Competition in and from the harmonization of private international law. In K. M. Meessen (Ed.), *Economic law as an economic good: Its Rule function and its tool function in the competition of systems*. Sellier European Law Publishers.
- Casey, A. J., & Macey, J. C. (2020). Bankruptcy shopping: Domestic venue races and global forum wars. *Emory Bankruptcy Developments Law Journal*, 37, 463–507.
- Chaigneau, P., Edmans, A., & Gottlieb, D. (2018). Does improved information improve incentives? *Journal of Financial Economics*, 130(2), 291–307.
- Coase, R. H. (1960). The problem of social cost. *The Journal of Law and Economics*, 56(4), 837–877. (2013).
- Coates, J. C., IV. (2001). Explaining variation in takeover defenses: Blame the lawyers. *California Law Review*, 89, 1301.
- Coyle, J. F. (2011). Business courts and interstate competition. *William and Mary Law Review*, 53, 1915–1983.
- Davidson, S., De Filippi, P., & Potts, J. (2016). Disrupting governance: The new institutional economics of distributed ledger technology. Available at SSRN: <https://ssrn.com/abstract=2811995> or <https://doi.org/10.2139/ssrn.2811995>.
- Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2003). Courts: The Lex Mundi project. *The Quarterly Journal of Economics*, 118, 453–517.
- Fisman, R., & Golden, M. A. (2017). *Corruption: What everyone needs to know*. Oxford University Press.
- Garoupa, N., & Ginsburg, T. (2015). *Judicial reputation: A comparative theory*. University of Chicago Press.
- Gilson, R. J. (1984). Value creation by business lawyers: Legal skills and asset pricing. *Yale Law Journal*, 94, 239–313.

- Godwin, A., Ramsay, I., & Webster, M. (2017). International commercial courts: The Singapore experience. *Melbourne Journal of International Law*, 18(2), 219–259.
- Goetz, C., & Scott, R. (1981). Principles of relational contracts. *Virginia Law Review*, 67, 1089–1151.
- Goldsby, C., & Hanisch, M. (2023). Agency in the Algorithmic age: The mechanisms and structures of blockchain-based organizing. *Journal of Business Research*, 168, 114195.
- Greenstein, S. (2009). Open platform development and the commercial internet. In A. Gawer (Ed.), *Platforms, markets and innovation* (pp. 219–248). Edward Elgar.
- Greif, A. (1989). Reputation and coalitions in medieval trade: Evidence on the Maghribi traders. *Journal of Economic History*, 49, 857–882.
- Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *American Economic Review*, 83, 525–548.
- Greif, A., & Kandel, E. (1995). Contract enforcement institutions: Historical perspective and current status in Russia. In E. P. Lazear (Ed.), *Economic transition in Eastern Europe and Russia: Realities of reform*. Hoover Institution Press.
- Greif, A., Milgrom, P. R., & Weingast, B. R. (1994). Coordination, commitment, and enforcement: The case of the merchant guild. *Journal of Political Economy*, 102, 745–776.
- Grossi, S. (2011). Rethinking the harmonization of jurisdictional rules. *Tulane Law Review*, 86, 623–712.
- Gunningham, N., & Sinclair, D. (2017). Smart regulation. In *Regulatory theory: Foundations and applications* (pp. 133–148). ANU Press.
- Hadfield, G. K. (1990). Problematic relations: Franchising and the law of incomplete contracts. *Stanford Law Review*, 42, 927–992.
- Hadfield, G. K. (1992). Bias in the evolution of legal rules. *Georgetown Law Journal*, 80, 583–616.
- Hadfield, G. K. (2001). Privatizing commercial law. *Regulation*, 24(1), 40–45. Also reprinted in *New Zealand Business Law Quarterly*, 7, 287–295.
- Hadfield, G. K. (2002). Privatizing commercial law: Lessons from ICANN. *Journal of Small and Emerging Business Law*, 6, 257–283.
- Hadfield, G. K. (2004). Delivering legality on the internet: Developing principles for the private provision of commercial law. *American Law and Economics Review*, 6, 154–184.
- Hadfield, G. K. (2007). Don't forget the lawyers: Legal human capital and the role of lawyers in supporting the rule of law. *DePaul Law Review*, 55, 401–421.
- Hadfield, G. K. (2008). The levers of legal design: Institutional determinants of the quality of law. *Journal of Comparative Economics*, 36, 43–73.
- Hadfield, G. K. (2011a). Producing law for innovation. In The Kaufmann Task Force on Law, Innovation, and Growth (Ed.), *Rules for growth: Promoting innovation and growth through legal reform* (pp. 22–54). Ewing Marion Kaufmann Foundation.
- Hadfield, G. K. (2011b). The dynamic quality of law: The role of judicial incentives and legal human capital in the adaptation of law. *Journal of Economic Behavior & Organization*, 79, 80–94.
- Hadfield, G. K. (2017). *Rules for a flat world: Why humans invented law and how to reinvent it for a complex global economy*. Oxford University Press.
- Hadfield, G. K. (2022). Legal markets. *Journal of Economic Literature*, 60, 1264–1315.
- Hadfield, G. K., & Bozovic, I. (2016). Scaffolding: Using formal contracts to support informal relations in support of innovation. *Wisconsin Law Review*, 2016, 981.
- Hadfield, G. K., & Heine, J. (2016). Life in the law-thick world: The legal resource landscape for ordinary Americans. In S. Estreicher & J. Radice (Eds.), *Beyond elite law: Access to civil justice in America*. Cambridge University Press.
- Hadfield, G. K., & Weingast, B. R. (2013). Law without the state: Legal attributes and the coordination of decentralized collective punishment. *Journal of Law and Courts*, 1(1), 3–34.
- Hadfield, G. K., Howse, R., & Trebilcock, M. (1998). Information-based principles for rethinking consumer protection policy. *Journal of Consumer Policy*, 21, 131–169.
- Hansmann, H., & Kraakman, R. (2000). The essential role of organizational law. *Yale Law Journal*, 110, 387–440.

- Hay, J. R., & Shleifer, A. (1998). Private enforcement of public laws: A theory of legal reform. *The American Economic Review*, 88, 398–403.
- Hendley, K. (2001). Beyond the tip of the iceberg: Business disputes in Russia. In P. Murrell (Ed.), *Assessing the value of law in transition economies* (pp. 20–55). University of Michigan Press.
- Henson, S. (2008). The role of public and private standards in regulating international food markets. *Journal of International Agricultural Trade and Development*, 4(1), 63–81.
- Herings, P. J.-J., & Kanning, A. J. (2008). Harmonization of private law on a global level. *International Review of Law and Economics*, 28(4), 256–262.
- Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics*, 10, 74–91.
- Holmström, B. (1999). The firm as a subeconomy. *Journal of Law, Economics, and Organization*, 15(1), 74–102.
- Jahn, G., Schramm, M., & Spiller, A. (2004). Trust in certification procedures: An institutional economics approach investigating the quality of audits within food chains. In *14th IAMA conference*.
- Johnson, S., McMillan, J., & Woodruff, C. (2002). Courts and relational contracts. *Journal of Law, Economics and Organization*, 18, 221–261.
- Kakalik, J. S., et al. (1996). *Just, speedy and inexpensive? An evaluation of judicial case management under the Civil Justice Reform Act*. RAND Institute for Civil Justice.
- Klein, B., & Leffler, K. B. (1981). The role of market forces in assuring contractual performance. *Journal of Political Economy*, 89, 615–641.
- Klein, B., Crawford, R. A., & Alchian, A. A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21, 297–326.
- Klerman, D., & Reilly, G. (2015). Forum selling. *Southern California Law Review*, 89, 241–316.
- Laffont, J.-J., & Martimort, D. (2002). *The theory of incentives*. Princeton University Press.
- Landes, W. M., & Posner, R. A. (1979). Adjudication as a private good. *The Journal of Legal Studies*, 8(2), 235–284.
- Lee, Y., & Meagher, P. (2001). Misgovernance or misperception? Law and finance in Central Asia. In P. Murrell (Ed.), *Assessing the value of law in transition economies* (pp. 133–179). University of Michigan Press.
- Macneil, I. R. (1985). Relational contract: What we do and do not know. *Wisconsin Law Review*, 3, 482–524.
- Mania, K. (2015). Online dispute resolution: The future of justice. *International Comparative Jurisprudence*, 1(1), 76–86.
- McDermott, C., Cashore, B., & Kanowski, P. (2010). *Global environmental forest policies: An international comparison*. Routledge.
- McMillan, J., & Woodruff, C. (2000). Private order under dysfunctional public order. *Michigan Law Review*, 98, 2421–2458.
- Ménard, C. (2005). A new institutional approach to organization. In *Handbook of new institutional economics* (pp. 281–318). Springer.
- Merryman, J. (1985). *The civil law tradition*. Stanford University Press.
- Milgrom, P. R., North, D. C., & Weingast, B. R. (1990). The role of institutions in the revival of trade: The law merchant, private judges and the champagne fairs. *Economics and Politics*, 2, 1–23.
- Murrell, P. (2001). *Assessing the value of law in transition economies*. University of Michigan Press.
- Murrell, P., & Ryterman, R. (2001). Law works in Russia: The role of law in interenterprise transactions. In P. Murrell (Ed.), *Assessing the value of law in transition economies* (pp. 56–93). University of Michigan Press.
- Nagle, F., Seamans, R., & Tadelis, S. (2020). *Transaction cost economics in the digital economy: A research agenda* (Harvard Business School Strategy Unit Working Paper 21-009).
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- Pei, M. (2001). Does legal reform protect economic transactions? Commercial disputes in China. In P. Murrell (Ed.), *Assessing the value of law in transition economies* (pp. 180–210). University of Michigan Press.

- Rasmussen, R. K., & Thomas, R. S. (1999). Timing matters: Promoting forum shopping by insolvent corporations. *Northwestern Law Review*, 94, 1357.
- Ríos-Figueroa, J. (2012). Justice system institutions and corruption control: Evidence from Latin America. *Justice System Journal*, 33, 195–214.
- Robinson, J. A., & Acemoglu, D. (2013). Why nations fail: The origins of power, prosperity and poverty. Currency.
- Romano, R. (1993). *The genius of American corporate law*. American Enterprise Institute.
- Rose-Ackerman, S., & Palifka, B. J. (2016). *Corruption and government: Causes, consequences, and reform*. Cambridge University Press.
- Rule, C. (2020). Online dispute resolution and the future of justice. *Annual Review of Law and Social Science*, 16, 277–292.
- Schmitz, A. J., & Rule, C. (2019). Online dispute resolution for smart contracts. *Journal of Dispute Resolution*, 2019, 103–125.
- Schulz, T. (2004). Does online dispute resolution need governmental intervention-the case for architectures of control and trust. *North Carolina Journal of Law & Technology*, 6, 71–106.
- Shah, A. (2004). Using ADR to resolve online disputes. *Richmond Journal of Law & Technology*, 10(3), 1–24.
- Shaked, A., & Sutton, J. (1981). The self-regulating profession. *Review of Economic Studies*, 11, 217–234.
- Shepherd, G. B., & Shepherd, W. G. (1999). Scholarly restraints? ABA accreditation and legal education. *Cardozo Law Review*, 19, 2091–2257.
- Spiegelman, J. (2006). Transaction costs and international litigation. *Australian Law Journal*, 80, 438.
- Szabo, N. (1997). Formalizing and securing relationships on public networks. *First Monday*, 2.
- Tapscott, D., & Tapscott, A. (2016). *Blockchain revolution: How the technology behind bitcoin is changing money, business, and the world*. Penguin.
- Telser, L. (1981). A theory of self-enforcing agreements. *Journal of Business*, 53, 27–44.
- Voigt, S. (2016). Determinants of judicial efficiency: A survey. *European Journal of Law and Economics*, 42, 183–208.
- Ware, S. J. (2013). Is adjudication a public good: Overcrowded courts and the private sector alternative of arbitration. *Cardozo Journal of Conflict Resolution*, 14(3), 899–922.
- Williamson, O. E. (1975). *Markets and hierarchies: Analysis and antitrust implications*. Free Press.
- Williamson, O. E. (1983). Credible commitments: Using hostages to support exchange. *American Economic Review*, 73, 519–540.
- Williamson, O. E. (1985). *The economic institutions of capitalism*. The Free Press.
- Williamson, O. E. (2003). *Doing business in 2004: Understanding regulation*. World Bank.
- World Bank. (2000). *Anti-corruption in transition: A contribution to the policy debate*. The World Bank.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Carol M. Rose

Contents

13.1	Introduction	294
13.2	Original Acquisition and the Cost of Property Regimes	295
13.3	Rights Definitions and the Question of Optimal Standardization	298
13.4	Transfer: Alienability, Commodification, and the Mystery of Capital	301
13.5	Use/Constraints on Use: Externality and Complexity	305
13.6	Use Redux: The View from Public Law Constraints	308
13.7	Property Law, Novel Property, and Public Service	311
13.8	Conclusion	313
	Bibliography	314

Abstract

Both detractors and proponents often describe property in individualistic terms, ignoring property's character as a *social institution* that depends critically on participants' respect of others' claims. Informal property regimes are pervasive among small or specialized social groups, but neither informal property nor property law emerges unless resources become scarce. However, property law differs in important respects from informal property, generally being a formal system that aims at universal applicability. In defining rights, property law generates fixed categories so that strangers can understand the rules; similarly, property law deploys public record systems to make transactions transparent. These simpler and more formal characteristics of property law encourage transfer. To address externalities, property law limits private uses, sometimes relatively loosely and ex post, as in nuisance law, but sometimes by creating ex ante categories to stave off disputes. Public demands on private property can lead to contentious legal disputes, as in the complexities of "takings" jurisprudence.

C. M. Rose (✉)

Yale Law School, New Haven, CT, USA

University of Arizona College of Law, Tucson, AZ, USA

e-mail: Carol.rose@yale.edu

Property law can also create innovative categories largely unknown to informal property, such as intellectual property and tradable environmental rights. Finally, both informal and formal property can contribute indirectly to democratic governance by modeling the recognition of rights and by encouraging compromise through trade, with the important caveat that property institutions sometimes fail in these respects.

Keywords

Property law · Informal property · Property transfer · Rights definitions · Numerus clausus · Property restrictions · Takings · Innovative property

13.1 Introduction

Property is not for utopians. In the early and turbulent years of the sixteenth-century Reformation, radical Anabaptists in Germany seized the town of Muenster and abolished individual property as incompatible with saintliness (Cohn, 1961). Two centuries later, Jean-Jacques Rousseau took up the theme with a more secular note, declaring that “the first man to say *this is mine*” took the initial step away from the idyllic state of nature and toward cruel civil society, with all its “crimes, wars, murders, . . . miseries and horrors” (Rousseau, 1750, p. 234). Venerable religious communities as well as American utopian communities have dismissed property as representing an individualism incompatible with their ideals for harmonious societies (Mandelker, 1984; Goffman, 1961).

And yet, however imperfect it may be, property does have many defenders. Among the important benefits claimed for it are these: property enables self-definition (Radin, 1982) and a flourishing self- and social development (Alexander, 2019); property encourages independence (Reich, 1964) and autonomy (Dagan, 2021); and property together with commerce enlarges social relationships (Hirschman, 1977). Perhaps the best-known claim for property in the modern era, however, is that property encourages labor and hence individual and social wealth (Bentham, 1802; Posner, 1973).

These asserted benefits of property redound largely to the property owner; so to speak, they are the good features of property and ownership seen from the inside out, and they generally add to the critics’ individualistic portrayal of property. But an essential aspect of property looks from the outside in and in doing so contradicts its supposed centrality of individualism: property can only function if nonowners tolerate, recognize, and respect the claims of owners (Penner, 1997; Rose, 2013). Even Rousseau said so.

It is something of a mystery why nonowners or what Henry Smith (2012) calls duty holders have the self-restraint to respect others’ claims to property (Rose, 2013). To be sure, nonowners play the role of owners themselves with respect to some things and could be moved by reciprocity; but one’s own possessions are a very small raft bobbing in a sea of possessions of others, and certainly a considerable

portion of those possessions could be grabbed. Discussing this puzzle, David Hume observed that respect for the property of others does not come naturally to individuals, but he added that people are clever, and as such they will go along with something as useful as the customs of property (Hume, 1739).¹ To this, one might say, yes and no; there are certainly instances in which property breaks down, sometimes with exceedingly disruptive effects.

The outside-in perspective thus looks at property regimes as social *institutions*, in which a property regime itself is a kind of common property of a social community whose members recognize one another's claims. Different kinds of communities have varying property norms, of course. Institutional economists and political and legal scholars have studied informal (or at least nonlegal) property institutions extensively in the last several decades.² Social communities come in multiple shapes and sizes, and accordingly, informal property regimes proliferate, ranging from the systems of taking-turns on city park basketball courts (Ellickson, 1993) to the elaborate irrigation allocations studied—along with many other property regimes for common pool resources—by the late Nobel Laureate Elinor Ostrom (1990) and her colleagues at the University of Indiana's Workshop in Political Theory and Policy Analysis.

Property *law*, on the other hand, is a formal system, and it differs from informal property regimes in its general applicability and its association with institutions of government. As the following pages will argue, these characteristics are centrally important to property law. But whether informal or legal or some combination, property regimes generally include understandings about several elements: acquisition, rights definition, transfer of claims, and use/constraints on use. This chapter will discuss all these elements, largely but not exclusively with examples from American law, and in each section it will briefly describe some leading theoretical analyses of property institutions. It will conclude with some observations about new types of legally created property.

13.2 Original Acquisition and the Cost of Property Regimes

How does property get started? This turns out to be a rather complicated question. In the first few pages of his discussion of property in *Commentaries on the Law of England*, William Blackstone (1766) narrated a little potted history of the way that property began. In the early stages in which resources were plentiful, he said, there was no property. But as the earth became more populous, people began to take things for themselves. But then, he said, from mere possession *property* emerged, basically from necessity; for who would sow if later he could not reap the fruits?

¹ Like other writers in his era, Hume (1739, vol. 3 part 2) used the term “justice” to mean respect for property but described justice as deriving not from nature but from convention.

² I describe as informal any system or set of entitlements that are governed by non-legal understandings and enforced if at all by the wish to maintain esteem. See McAdams (1997).

This may strike the reader as a good story but not a very persuasive analysis, given that it says nothing about that essential issue for turning mere *possession* into *property*, that is, securing the respect and forbearance of nonowners (Rose, 1994). Some more modern economic thinkers, notably Robert Sugden (2004 [1986]), have tried to fill this gap through game theory, describing property as a hawk/dove game in which possession itself acts as a salient fact that makes others respect it as entitlement. Supposedly this is possible because the possessor normally has the superior defensive position. But Frederick Pollock's and Samuel Wright's old treatise on possession (1888) pointed out that possession alone is a remarkably weak defense against surrounding marauders. The shift from possession to property means *not having to guard* and to be able to go about one's business without wasting time, money, and psychic energy on defending what one owns, because others respect your claim.

Whether or not the gap in the story is filled, however, the basic outline of Blackstone's narrative continues to dominate economists' stories about original acquisition. Harold Demsetz (1967) wrote a much-cited article that made explicit one implication of Blackstone's story: no one bothers with property for limitlessly available resources, because property regimes are costly. As Demsetz argued, property will not emerge unless it is worth the cost. Demsetz used the example of the eighteenth-century fur trade, arguing that the sudden increase in value of beaver pelts induced Native Americans in Canada to begin marking out and managing family territories for beaver habitat. Terry Anderson and P. J. Hill (1975) built on Demsetz' cost-benefit approach to property, using several examples from European settlement in the American west to show, among other examples, how larger herds and higher demand for beef induced cattlemen to adopt increasingly complex property schemes, both informal and legal, to secure their livestock. Anderson and Hill added that lower-cost technological property protections, notably barbed wire, also changed the cost/benefit calculation toward more property.

It would seem, then, that if people create property and devise more effective acquisition and protection methods, they do so because the benefits exceed the regime costs. As James Krier has pointed out, they may not do so even then (Krier, 1992). But if they do develop property relations, costliness gives a reason why informal property regimes so often precede formal property law: law is likely to be more expensive than informal norms.

The actual law of original acquisition often aligns with the informal norms that allocate a right of property to the first one to take observable possession of a previously unowned thing. In the United States, the common law of initial acquisition is oddly fixated on wild animals and the means by which individuals turn them into property—wild animals supposedly being the prime example of things naturally owned by no one. Wild animals become property through the common law “rule of capture” or “occupancy,” a variation on first possession, according to which the property right goes to the one who first kills the animal or brings it under visible control with nets or snares (Rose, 1994). However, wild animals give a homely illustration of the instability of possession alone as a ground of property: a wild animal may escape the snare and become legally wild again and hence available for capture by a second capturer and perhaps a third or fourth after repeated escapes.

Despite some odd comparisons, common law courts have analogized other and much more valuable “fugitive” resources to wild animals, notably oil and gas as well as groundwater. Unless altered by legislation, the overlying landowner who first brings these fluids to the surface becomes the legal property owner of the portion so captured. Unfortunately, there are often many landowners atop the sources of these highly mobile substances. Obviously, this rule of first possession sets up racing behavior that can lead to the proverbial “tragedy of the commons,” to the ruin of the common resource (Hardin, 1968; ► Chap. 15; Glennon, 2002). Thus deep groundwater pumping collapses a large underlying aquifer; competing oil rigs flare off the natural gas whose pressure brings oil to the surface; and overhunting and overfishing decimate animal and fish populations.

Common law judges will sometimes approve informal acquisition practices that are considerably more refined than mere capture or possession, even with respect to wild animals, but these too can have wasteful effects on the larger stock. An example is the judicial approbation of customary practices among some nineteenth-century whalers: these rewarded the vessel whose sailors took the greatest risk by landing the first harpoon strike against one of these very dangerous animals, even if another vessel ultimately succeeded in completing the kill of the wounded prize (Ellickson, 1991). These customary rules encouraged effort and undoubtedly made the hunt more efficient, but there was a very serious cost: whalers collectively had such success that they nearly wiped out whole whale populations. Other rules of initial acquisition have the same problem: their very local efficiency may exhaust the larger underlying common pool resource.

In each case of this sort, what is needed is a property rule or other management form that does more than chose the winner in the race among competing resource takers and instead governs the larger common resource itself. This is especially important in the case of the grand, widely available but ultimately finite resources that are likely to be designated “environmental.” These resources create especially difficult obstacles for property or indeed for any other kind of management regime.

Racing and common pool losses again point to the costliness of property regimes, here the costs of switching from property regimes that manage low-level conflict between individuals, to a succession of other regimes that can confront increasingly intensive uses of larger resources with their accompanying common pool losses. Gary Libecap has extensively studied these kinds of switching costs, observing that they arise from the fact that property regimes are social institutions, subject to collective action problems among the participants. From his numerous studies of water, oil and gas, fishing, and even Amazon frontier land settlement, Libecap argues that informal or less developed property rules may be sufficient to avoid conflict at low levels of resource pressure, but that at higher levels, more sophisticated formal law is needed to staunch the ever-growing collective losses (Alston et al., 1999; Libecap, 1989). Creating or changing property *law* involves politics, however, and in the political arena, Libecap argues, incumbent beneficiaries of the existing rules exert an outsized influence. What is more, he claims, much-needed rule changes generally can only be made politically (if at all) by allowing incumbents to continue their activities or by paying them off in some other way.

Thus the issue of initial acquisition in property reveals that property regimes have several kinds of costs, with differing cost structures depending on the level of pressure on collective resources (Rose, 1991). Moreover, the initial acquisition issue also points to the necessary if distasteful distributional side payments that may be needed to advance from open access to dispute resolution to common pool management.

The topic of initial acquisition has another quite different questionable aspect. Initial acquisition implies that the acquired thing has no legitimate prior claimant. But uncomfortably often, that is a contestable claim. Initial acquisition means that possession or capture creates property by signaling to the world the possessor's or capturer's claim. But who is in that world? Are they really the world, or just some more limited property community that understands the signal and considers it legitimate? In many parts of the globe, colonists used doctrines of initial acquisition from the sixteenth century onward to dispossess indigenous communities, even though those communities understood their relationship to the land very differently. After well over a hundred years of ignoring the issue of dispossession, it has become an important if sometimes fraught issue in many settler societies, including Canada, Australia, New Zealand, and the United States, and it currently has an impact on human rights questions about the property claims of various indigenous groups. These kinds of discussions are related to a second major requirement of property law: defining entitlements.

13.3 Rights Definitions and the Question of Optimal Standardization

Any system of property must have some way to define what counts as property. Given the critical role that nonowners (or duty holders) play by respecting the claims of others, a minimal characteristic of a claim to property is that it must be recognizable to the members of the relevant property community—that is, to the persons whom one might describe as the audience for those claims. As discussed above, the common law posits “possession” as the ground of original acquisition of property, but it treats possession itself as a signal to others in the community.

Henry Smith has very usefully applied linguistic analysis and information theory to the relationships of cognition in different kinds of property communities (Smith, 2003). Smith observes that there is a tradeoff between the extensiveness of the property audience on the one hand and on the other hand the density of information that a property claim can contain. In a small community of persons who know one another, property claims may be very complex. Thus in the island fishing community described by James Carrier (1987), some individuals or groups can fish with nets, others with spears and spearguns, and all at differing times and with differing locations on and around the reef. Similarly, prior to the enclosures of the English countryside, villagers lived with complex rules about (among other things) when and where livestock could be grazed on the village commons and which persons were permitted to run their animals (Neeson, 1993).

Modern property law, of the type recognized by today's commercial societies, cannot support this level of individualized complexity. It is not that such complexity is impossible; it is rather that the most complex relationships are created by contract rather than property. Contractual obligations are personal to the contracting parties, *in personam* in lawyers' Latin. But property law is *in rem*, purporting to be "good against the world" and governing everyone's relationship to something—everyone including strangers and foreigners and others who never made any contracts about the thing in question. At the most basic level, one or more persons are the thing's owners; everyone else has a duty to avoid disrupting the thing, on the general understanding that it belongs to someone else (Penner, 1997).

Modern property law admits greater complexity than this most basic set of obligations *not to interfere*. But because of their aspiration to general applicability, legal property regimes drive property claims toward relatively simple standardized categories. With these, any set of transacting parties can build more complex relationships by mixing and matching simpler categories and can comprehend the components without undue search costs, at least with the assistance of a professional attorney or notary (Smith, 2012). Civil law property regimes deploy a "numerus clausus" for landed property—a fixed number of categories into which claims must fall in order to be considered legal property (Merrill & Smith, 2000). Civil law jurisdictions are relatively strict, for example, omitting trust relations from the numerus clausus, although they generally do permit some limited analogs (Chang & Smith, 2015). Common law jurisdictions are somewhat more relaxed about novel property categories, adopting trusts and well as several innovative easements and covenants. If these devices are reasonably observable either physically or through documents, common law property law treats such devices as "running with the land" to successors in interest, as opposed to remaining purely personal to the original parties. But common law property regimes still push arrangements into standard forms like fee simple, life estate, lease, remainder and reversion, mortgage, emblements, and so on.

Once again, individual owners and nonowners are free to make mixtures of the various categories to create more specialized arrangements among themselves about a given property. But the more these arrangements involve individually tailored details, the less they are considered property—good against the world and benefiting or binding successive owners—and the more they are treated as contracts, which benefit and bind only those who have agreed personally. Presumably persons who negotiate contracts can include multiple details and still know what is in the deal, but the same cannot be said of later purchasers of a given piece or property.

Categorization in property law continues to subdivide when more refined types of claims become more prevalent. Thus the lease is a type of property, but as these time-limited uses of property have changed functions since the nineteenth century, particularly with respect to residences, the lease category has subdivided accordingly into leases for years (i.e., a definite time), month-to-month leases, leases at sufferance, etc., each with its own sets of rules about the obligations of lessors and lessees

(Merrill et al., 2022).³ The categorization mode typical of property continues in newer forms of property as well. Intellectual property, itself almost completely a product of formal law, subdivides into the categories of patent for inventions, copyright for creative works, and trademark for commercial branding, each with its own requirements and rules. An even further example of categorization appeared a few years ago, when critics of copyright law founded what they call the “Creative Commons.” This is a method to permit free access to writings and other creative projects; but those who post their work on the Creative Commons choose among a menu of categories of access, particularly governing the type of acknowledgment that the author prefers (Van Houweling, 2008). Thus even the Creative Commons echoes the principles of the *Numerus Clausus*.

There is a reason for what seems like a property law’s obsession with classification. The various categories define the character of property transactions—who gets what with sales, purchases, loans, gifts, security interests, and so forth. Ideally, the categories are varied enough to allow for some complexity in property arrangements, but they respond to the cognitive limitations not only of the transacting parties but also of future transacting parties and their successors in interest. Not surprisingly, future parties are a matter of special concern for transactions about something so durable as land, and thus the categories of landed interests are particularly well marked. These may seem annoyingly rigid to lay people, but they are familiar to real estate professionals, and in transactions, they cut down on the need to search out idiosyncratic rights and obligations (Merrill & Smith, 2000).

Search cost reduction is a major advantage of property categories, but there are some others: the legal categories simplify transactions by giving the parties several off-the-rack options, which the parties can then customize by mixing and matching (Rose, 1999). A condominium complex is an example; each owner usually owns her unit in fee simple but has an exclusive easement on the interior walls and tenancy in common on the swimming pool and other common facilities. In addition, the limited types of property categories reduce the chance of what is called the “anti-commons”—a scenario in which the legal interests in a given asset are so numerous or so heterogeneous that it is impossible for any of the parties to collect the necessary consents to make use of it (Heller, 1991). In the case of land assembly, say for a shopping center, a developer may be faced with large numbers of interest-holders, but at least those interests fall into known categories with which she or her legal staff will be familiar.

Readers may have noticed that the categorization familiar in modernist property regimes implicitly assumes that property interests will change hands and that standard categories encourage easy transactions when transfers occur—even among strangers who may reside anywhere in the world. These transfers are the subject of the next topic.

³ Another cross-cutting but widely recognized division is that of commercial vs. residential leases.

13.4 Transfer: Alienability, Commodification, and the Mystery of Capital

Why does it matter that property can change hands? One major answer is that the possibility of transfer creates value. John Locke famously argued that labor creates value in property but that is only a part of the story. Property encourages labor, but without the possibility of transfer, only the laborer can enjoy the fruits of her labor. When she can transfer the property, however, she can focus on what others want; her labor is all the more valuable because she can reap the rewards of exchange with all possible transferees. Transfer and exchange thus accelerate property's ability to encourage effort, planning, and care and thus personal and social wealth.

I leave to one side the tremendously important physical transportation infrastructure that governments provide to encourage commerce—the roads, the railroads, the ports and air transportation hubs, and the communications systems—but no one should underestimate their addition to people's ability to transfer property. Modern property law's categorization, discussed above, is a part of the nonphysical infrastructure that encourages transfer and exchange, because limited categories are transparent to much wider arrays of potential buyers and sellers. But there are other legal instruments that also encourage transactions, and they are closely associated with property law if not necessarily a formal part of it. The rectilinear land survey of the United States and the grid street layout of many US cities might be viewed as geographical analogs of property law's *numerus clausus*; they simplify boundaries and boundary nomenclature and incidentally may make land more valuable because it is more easily transferred (Libecap & Lueck, 2011).

Recording and registration systems are key parts of the legal infrastructure for property transfer. Not only do they inform purchasers about claims against any given property, but they also allow sellers to give verifiable title information to those purchasers, while protecting third-party claimants—like mortgage banks or neighboring easement holders—from claims by uninformed purchasers. There are, of course, important differences between registration systems and recording systems. Registration systems are associated with civil law systems, whereas common law regimes often use recording rather than registration, at least for land transactions.⁴ Registration systems purport to state on the face of the registered title all the legally recognized interests on any given property, and once the process is complete, the registry also insures the registered title against unregistered claims, however sympathetic they may be. Accordingly, registry systems involve a relatively thoroughgoing front-end investigation in the case of transfer, and they are strict in recognizing only the limited property claims of the *numerus clausus*. Recording systems, by contrast, basically give a chronology of transactions. The recorder of deeds can

⁴Even in common law countries, automobiles and ships are registered, and even land is in a registry system in Australia and other British Commonwealth jurisdictions.

record claims without investigating them and certainly without insuring them, leaving it to the parties and their attorneys to investigate any incompatible interests from past transactions. The advantage of registration systems is certainty about registered title, with the disadvantages that registration is time-consuming and expensive. Recording systems make the tradeoff in the opposite direction: rapid transfer at relatively low cost upstream, but with some uncertainty downstream about possible unrecorded claims, and about how far back to trace the records.⁵

By comparison to all this public infrastructure around property transfer, the direct legal rules about transfer are relatively limited. Jurisdictions with recording systems need to rule on the validity of claims that are not in the records but are actually known to the relevant parties, and recording and registration have to make some accommodations to claims arising from long prescription.⁶ These matters complicate transfer occasionally, but modern property law generally favors easy transfer, perhaps especially in the United States. Early in the history of the new republic, state governments abolished the English “entailed fee,” which had prevented transfer of landed property outside the established family; abolition supposedly broke up aristocratic landholding, but it also allowed families to sell their old lands and buy new land further West (Priest, 2021). From the nineteenth century onward, courts have frequently stated their disfavor to “restraints on alienation,” but they have relaxed that disfavor to permit new-fangled covenants running with the land, essential for new kinds of subdivision development. Sadly, for several decades, that relaxation included racially restrictive covenants, aimed chiefly at keeping African Americans and Asians out of all-White subdivisions (Brooks & Rose, 2013).

One controversial issue about alienability emerged in the later twentieth century in connection with arguments for land titling, specifically for bringing informal land claims into the formal legal system. Peruvian economist Hernando DeSoto is the best-known advocate, arguing that formal title can unleash entrepreneurship in impoverished areas (DeSoto, 2000). According to this view, informally held lands contain a vast quantity of untapped capital; this is because unless they have title, owners cannot use their claims as security for loans. On the other hand, the argument continues, with formal title, budding entrepreneurs can fund their projects by borrowing against their holdings, thus financing new businesses and enhancing both their own and their communities’ wealth.

With the blessing of the World Bank, land titling programs have been instituted in many less developed countries, but they have met an uneven response. City dwellers in urban barrios or favelas—often rural migrants seeking to better their chances in the city—are a centrally important target group for titling. However, some commentators have been concerned that their informal *colonias* are often built in

⁵Many jurisdictions in the United States have so-called marketable record title statutes to limit the time back of record search.

⁶Jurisdictions with so-called race systems recognize only the first to record; those with “notice” systems hold actual knowledge of prior claims against the purchaser.

environmentally sensitive or dangerous locations like floodplains or steep slopes subject to landslides (King, 2003). Others claim that these migrants do not in fact borrow against their properties even with formal title, either because lenders regard their holdings as insufficiently valuable or because they themselves do not want to put their holdings at risk (Varley, 1987). Still other researchers, however, argue that land titling has different benefits, such as giving the holders and their families security to leave home to seek education or employment or a sense of stability that encourages investment and long-term planning (Field, 2007; Galiani & Schargrodsky, 2010).

Traditional communities with long-standing informal landholdings are another target for titling programs, but they present issues different from the barrios of rural-urban migrants (Rose, 2009). Lands in such communities often have quite complex and overlapping sets of entitlements, reflecting the character of property in smaller traditional groups discussed in Part 2: entitlements are information-rich and well understood by local residents but not by outsiders. Here land titling introduces a modernist form of claim that can seriously disrupt long-standing community practices. Formal titles substantially simplify traditional claims, and these now-uncluttered titles are likely to be allocated exclusively to persons with the major interest in any given holding or even simply to the headmen of the community. This pattern cuts off former lesser rights like gleaning or cultivation privileges—exclusions that can be doubling damaging to women when they fail to correct the misogyny of pre-existing customary norms (Kapur, 2011). Individual sales as well as failed loans and foreclosures can introduce strangers into the community, who may well not heed community practices and responsibilities. A particular peril for titling programs in traditional areas is that an incomplete introduction can lead to confusion about authority and open the door to advantage-taking by more sophisticated or powerful players (Fitzpatrick, 2006).

In both barrio and traditional scenarios, upon the eventual transfer of titled properties over time, the parties may not take on the expense and bureaucratic burdens of reregistration. In that case, their holdings may slip out of the formal records, as, for example, in South Africa's Cape region, where off-registry transactions have bedeviled some programs for subsidized housing (Barry & Whittal, 2016). Even under the less exacting recording systems, properties once held formally can revert to informality. Though not part of titling programs like those advocated by DeSoto, poorer agricultural areas in the Southern United States have experienced the slide to informality in what is called "heirs' property," where the children and then grandchildren of an original owner divide up the property without recording their specific claims. Here too, the occupants assume that they are rightful owners of a particular piece of land, but they can be threatened with dispossession by advantage-taking entrepreneurs who search the records and acquire some partial title from an unwitting distant relative (Mitchell, 2005).

A quite different controversy about property transfer also emerged in the later twentieth century: the question of commodification. In her 1982 article, *Property and Personhood*, Margaret Radin fanned the debate over this topic, arguing that

many matters can be classed as alienable “fungible” property but that the category of “personhood” property—particularly important to the holder’s self-conception and projects—should enjoy protections against transfer. She followed in 1987 with a second article, *Market Inalienability*, spelling out reasons for constraining alienability through sales, but not necessarily gifts (Radin, 1987).⁷ One might describe these reasons roughly as paternalistic (damage to the transacting parties) and externalities (damage to third parties).

Several of Radin’s most striking candidates for constraints on alienability are not normally subjects of property law—notably sexual services, babies, and body parts.⁸ But some others are normally within property law, particularly residential property, where Radin’s position on homes as personhood property led her to a defense of rent controls (Radin, 1986). Commentators influenced by economics generally disagree, arguing that rent control and similar tenant protections damage tenants as a class because they discourage the development of rental property, particularly for lower-income tenants. A more recent version of this kind of issue can be seen in urban gentrification, where home-improving buyers move into low-income communities, adding to neighborhood upkeep but displacing existing tenants. Tenant protections and gentrification both exemplify a dilemma that has surfaced in alienability debates since the enclosure movement of early modern England: freer alienability may enhance overall social wealth but at the cost of damage to particular individuals or communities (Bogart & Richardson, 2011; Neeson, 1993).

As in Radin’s work, paternalism and externalities continue to loom large as reasons for legal constraints on alienability, as, for example, with arguments about limiting sales of drugs and guns. Arguments for and against these constraints follow a familiar pattern: proponents assert that they are necessary to prevent some activity that is dangerous for the participants as well as third parties, while opponents argue that the constraints sweep too broadly, banning many legitimate activities and only a few illegitimate ones, which should be halted in other ways.

An interesting variant on the externality version of legal inalienability is the use of sales bans to support conservation of endangered species. Eagle feathers, radiated tortoises, rhino horn, and wastebaskets are made from elephant legs—the sale of these and many other threatened species and species parts are banned under international law and implementing national laws. The hope is that by cutting off markets, sales bans can further discourage illegal capture and killing of these animals and plants (Webster, 1997).

Externalities form only one of the issues in the property law of transfer, but in legal issues about the use of property, externalities often play a central role.

⁷An earlier foray into this topic was Richard Titmuss’ *Gift Relationship* (1971), arguing for the superiority of blood donation over sale.

⁸Active markets do exist for some body parts such as hair and blood and, in spite of legal constraints, kidneys and other organs and dead bodies (Cheney, 2006).

13.5 Use/Constraints on Use: Externality and Complexity

The great conceit of property law is that one can use one's property just as one pleases, as in Blackstone's *Laws of England's* overly-cited nostrum about property: that it gives the owner exclusive and complete dominion of the owned thing (1766). But Blackstone himself very soon followed this assertion with a huge array of ways in which others' legal claims could limit any given owner's uses. Even more generally than specific third-party entitlements, to which we will return, private nuisance law puts a general obligation on an owner to use her property "reasonably," so as not to disturb the equally reasonable uses of others.

As with "reasonableness" in tort law, the meaning of reasonable use in property law is imprecise, depending on an ex post judgment of circumstances. "Reasonable" is often coupled with "ordinary," to which Robert Ellickson has added "normal"—in Ellickson's analysis, those are property uses that deserve neither praise nor disapprobation (Ellickson, 1973). All these terms depend on surrounding circumstances, but they may reflect a rough judgment that ordinary uses are likely to have a higher collective value than individual oddities. Aside from that, such terms have a general advantage in letting the owner know that she can use her property in the same ways as others in the vicinity or in a way that does not intrude on their ordinary uses; by the same token, they cannot intrude on hers. But a neighbor with special sensitivities gets no breaks in nuisance law; one with hypersensitive ears cannot bring a nuisance action about the lawn mower or radio next door—unless, of course, those devices create an unusually high level of noise. One might note the difference with informal norms; for example, a whole neighborhood might be aware that one person is recovering from an illness and keep their noise down; but as with formal property law generally, nuisance law ignores fine points.

The usual subjects of nuisance law are actions whose effects are difficult to measure precisely and that affect indeterminate numbers of others in varying degrees. Legally recognized nuisances often move through those great common resources, the air and sometimes water; and they involve perpetrator's uses that have unusual levels of noise, dust, smoke, odors, and pollutants. Indeed, one might argue that nuisance is always about misuse or overuse of common resources, a reminder that property is never truly solitary.

Visual matters are almost never treated as private nuisances, however. In older US law, this was because an owner's garish paint job or bizarre shrubbery was considered a matter of indeterminate taste. There are some exceptions: flashing lights are a standard nuisance because they disturb sleep and may trigger mental disruptions. A curious older exception to the no-visual-nuisance rule concerned copulating horses or asses, the unshielded sight of which could be treated as a private as well as a public nuisance (Joyce & Joyce, 1906). Given the prevalence of automobiles today, horses play a much-diminished role in nuisance cases. But changes in technology also add new kinds of nuisance issues. Maureen Brady has discussed a particularly ambiguous new nuisance question: whether light projections—pictures or messages—onto the buildings of others might be actionable nuisances or perhaps trespasses (Brady, 2020).

Property law scholars frequently contrast nuisance—the owner’s unreasonable use of her own property—with trespass, in which someone else intrudes on the owner’s property. Thomas Merrill (1985) wrote an influential article arguing that nuisance and trespass can be differentiated by reference to transaction costs. Trespass, he argued, is typically a violation in which transaction costs are low—that is, a clearly identifiable boundary crossing onto a particular owner’s property, usually involving some visible physical object or the trespasser herself. The trespasser has little or no excuse for failing to seek the owner’s permission and abide by her decision, and hence the legal response can be equally hard and fast: any such intrusion is actionable, whether or not there are damages to the complaining owner. On the other hand, when an owner uses her own property in a way that purportedly creates a nuisance, she does something much less precise as to both quantity and quality; her dogs bark too loudly or her barbeque is too smoky or aromatic, and various neighbors might be affected in different degrees. In Merrill’s argument, transactions costs are higher in such cases, and a legal decision on whether some action is a nuisance or not requires a kind of all-things-considered discretionary judgment, including a showing of damage.

There is one exceptional legal constraint on what would usually be a normal or ordinary use of property. This is the so-called spite fence, a category that grabs attention because it often illustrates an entertaining and all too recognizable feud between neighbors. The archetype is the fence built directly across from the neighbor’s window, constructed to block the view. But other matters might count too, like overwatering plants at the border or installing a sculpture of a fist with a raised middle finger in the neighbor’s view. But the offended neighbor must show that the intent was malicious, and this is not easy with what would usually be considered a normal property use.

Much more detailed constraints on an owner’s use can come from the overlapping entitlements of others, to which the owner or some predecessor owner at some point agreed. Blackstone’s discussion of property law is replete with *ex ante* entitlements of this sort. They are now commonly known as “servitudes.” Since the twentieth century, they have blossomed into the kinds of complex private land use controls that now govern many real estate developments like shopping centers, office complexes, and common interest communities.

Of all these devices, easements are the simplest, and among easements, the most common is a driveway or right-of-way; these oblige the underlying owner not to interfere, but they do not require much more. A driveway or right-of-way is usually visible and attached to an adjoining lot, so that future purchasers of either lot are unlikely to be surprised by it. The less visible the control, however, and the more difficult to find the holder, the more obstacles property law has raised to permanence. Thus easements “in gross”—unattached to adjoining land—at one time could not be transferred, or easements for a view, not visible on the burdened property, had to run the gauntlet of so-called covenants running with the land. Modern systems of recording and registration make these physically invisible and unattached claims knowable to future purchasers, and accordingly, at least some older legal limitations have relaxed.

Real estate servitudes are at the heart of modern private land use controls. Many planned community residents now live with restrictions on exterior paint colors, architectural styles, and parking, among many other matters. Older limitations on these devices reveal some still-relevant concerns about claims that “run with the land.” Legibility to future owners has always been among these, and even before widespread use of record systems, common law courts required that running covenants be placed in major documents of transfer, like leases or deeds, so that they could be discovered by future purchasers (Brooks & Rose, 2013).⁹ A second concern is that obsolete restrictions might tie up a property if claims become too fragmented for future renegotiation. Traditional doctrines addressed the fragmentation issue, at least to a degree, by requiring that successors hold the same interest as the original claimants; to the same end, modern planned communities dispense with unanimous consent and consolidate decision-making authority in elected boards. A related concern is about the aggregate land value of restricted properties—that is, the burdens and benefits of restrictions should add up to a net increase in value to the properties so burdened and benefitted. The traditional doctrine roughly dealing with aggregate value is known as “touch and concern,” which one might see as requiring that the burdens and benefits of the various private controls add more value to affected properties than they subtract (Id). Despite the limitations on private land use controls and claims, however, planned communities have come in for considerable criticism as rigid, overly controlling, and potentially obsolete—and as a barrier to social and environmental goals (Ellickson, 2022; Hyatt, 1998).

Private land use controls are among the most complicated topics in modern property law. The complexity of covenant law offers another glimpse of the difference between modernist property law and informal arrangements, especially those in more traditional communities. The latter often have layers of entitlements that are vastly more subtle than those of, say, a modern condominium. For example, different persons may have separate entitlements to the berries, bark, and leaves of an individual plant (Banner, 2007); or transfers of claims might require intricate specific ceremonies that make them memorable to a local audience (Hibbetts, 1992). These information-rich mixes of informal entitlements are known to everyone in the community, even though they may baffle outsiders. Property law’s private land use controls, on the other hand, illustrate an effort to create rules that govern participants who come from anywhere—insiders, outsiders, and strangers in the future. The complexity of the covenant law governing private land use suggests the tension between highly tailored arrangements and modernist property law’s need to standardize. It is not surprising that some commentators complain that the sets of covenants for planned communities lack variety, imposing the same sets of finicky requirements on all (Hyatt, 1998).

Nuisance and private land use controls are the primary limitations on a modern owner’s uses, but the brief account in this entry leaves out several others. For

⁹This initial relationship is sometimes called “horizontal privity,” as opposed to “vertical privity” between successive owners.

example, an owner may be subject to mining rights or to “emblemments,” that is, another person’s right to cut trees or take crops. A time-bound resident like a tenant or life-estate holder cannot “waste” the property, basically meaning that she should treat the property as she would if she owned it outright. An owner also needs to give some minimal ongoing signals that inform others of her claim, because if someone else unwittingly uses all or part of her property for a sufficient time, ownership may pass to this user under the doctrine of “adverse possession,” which serves to clear away uncertain claims.

There are more of these private limits on an owner’s use, but all those discussed so far refer to relationships among private parties. Public constraints come next.

13.6 Use Redux: The View from Public Law Constraints

Individual uses of property, and particularly property in land, can have impacts on larger-scale common goods like air, water, groundwater, wildlife, or even community ambience. Land owners often see these uses of adjacent resources—say, releasing the smoke of burning leaves or drilling water wells—as part of the rights that accompany their properties. Older common law supported them to the degree that their activities used the space directly above or below their land; but air, groundwater, and wildlife are no respecters of boundaries. There is little problem where such uses are few or unintensified. But with more numerous or more intense uses, some common management is needed to protect the common good (Rose, 1991). Many long-standing informal communities have devised various ways to “govern the commons,” as Elinor Ostrom and her colleagues have attested (Ostrom, 1990). However, Ostrom stressed the importance of boundary-keeping in maintaining common pool resources, and accordingly, some form of exclusion is likely to be present in informal ways to deal with resource pressure. One of several examples that Ostrom cites comes from Swiss mountain communities, where only persons with a long family history in the village are allowed access to the pastures. More sinister examples appear in elaborate initiation rituals for membership in criminal gangs, to the exclusion of others. But benign or sinister, informal community management practices depend on intense ties and complex understandings of who is in and who is out, whereas modernist property law is open to anyone, and depends on governments to manage stressed common resources.

Governmental management indeed brings many constraints on property from public law. The relevant areas of public law have become sufficiently sophisticated that they have hived off from conventional property law, particularly migrating to environmental law and land use law, now far from their origins in nuisance law. Despite substantial achievements in these newer public law areas, there are many complaints about them aside from their impact on property ownership, particularly that they do not serve the public interest. Thus, some claim that environmental laws block economic development, while a growing number of commentators argue that land use regulations favor the residents of wealthy communities over the less

wealthy and especially stand in the way of housing for persons with low or moderate incomes (Schleicher, 2021).

Individual property owners' claims against public regulations are often framed as "takings" of property. These are likely to arise as resource uses become more intense; here, litigants' arguments tend to fall into recognizable patterns. Beach regulations can serve as an example: modern transportation has enabled many more people to reach beaches, to the annoyance of beachfront property owners. But given the increasing popularity of beaches, some regulations have required that waterfront owners permit public access and not block anyone from walking in front of private owners' properties. Suppose that Private Owner X objects. The main components of the X's "takings" complaint are overlapping claims that (a) the regulation takes away an identifiable property right (especially the right to exclude) without compensation, (b) that the regulation singles out X to bear the cost of a benefit to the public at large, and (c) that the regulation so devalues X's property that it amounts to a taking of the property itself (or sometimes a specific right, as in (a)).

The regulators' responses usually fall into the following defenses: (a) X's property did not include the asserted right (because the relevant part of the beach belongs to the public and not to X); (b) in blocking access, X was committing a nuisance; (c) the regulation is a general limitation and not something specific to X's property; (d) the regulation implicitly compensates any limitation on X's property by the reciprocal benefit X receives from the same limitations on others (X can walk the length of the beach); and (e) any loss that the regulation causes to X's property is de minimis.

There are a great many variants on these arguments; for example, X may raise a takings claim against a new regulation that prohibits the disposal of spent oil containers on her property, whereas the regulator claims that X's containers can pollute water in the underlying aquifer, a common-pool resource owned by the public and not within X's property right. Over the last several decades, the Supreme Court in the United States has attempted to formalize some takings claims, holding that a physical invasion of an owner's property is a per se taking and so is a regulation causing complete loss of the property's economic value. Formalization has not been successful, however, leading to especially contorted claims of physical invasion (Rose, 2020). The inability to pin down takings claims and defenses suggests that this topic, like nuisance, is only approachable ex post.

Scholars have come up with many different rationales for and against property takings doctrines, admirably summarized by David Dana and Thomas Merrill in their book, *Property Takings* (2002). Many of these rationales relate to the political background of regulation, for example, that takings compensation should be available to (and perhaps limited to) those who cannot defend themselves in political arenas, that developers use the threat of takings litigation to intimidate impoverished local governments, and that, contrariwise, that developers need defenses from regulators who would use them as cash cows for local projects. Other discussions have an economic or public choice resonance. An idea first floated by Lawrence Blume and Daniel Rubinfeld (1984) is that takings claims fill a gap in insurance markets. The argument is that insurers will not cover losses of property value due to

new regulations because regulations are political matters, whose probabilities cannot be predicted statistically—unlike the normal insurable subjects like auto accidents or house fires. Instead, regulatory losses fall into the category that economist Frank Knight (1921) famously designated as “uncertainty,” as opposed to the actuarially foreseeable category of “risk.” But governmental compensation for property takings, it is said, can recompense affected property owners for such losses, with funding coming from taxes that all property owners have paid. Critics of this idea have pointed to problems with the insurance analysis, particularly the issue of moral hazard—that is, if they are insured, property owners and developers will overplay their hands, knowing that they will be compensated later if regulators halt them (Kaplow, 1986). But the moral hazard critique has critics of its own, for example, that the supposed hazard has few real effects (Lee, 2019). Another takings rationale is that of “fiscal illusion”: that unless regulators are faced with compensation requirements, they will fail to take costs into account and will blithely spend other people’s money for supposed public benefit (Dana & Merrill, 2002). The fiscal illusion argument also comes under fire, however. Daryl Levinson (2000), for example, has argued that political authorities make decisions differently from market participants and that they are more likely to be influenced by the loud complaints of individual condemnees than by the muffled rustlings of diffuse multiple taxpayers. Thus they may well be motivated to compensate too much rather than too little, saddling the public with needless expense.

Readers should be aware that the above comments take up only a small sample of the voluminous theories about property takings, leaving out even some of the best known, such as Frank Michelman’s foundational article on the issue (Michelman, 1967). There are many reasons for the volatility of this topic, undoubtedly relating to the sharply differing ideas about appropriate functions of property related in the Introduction. But one fruitful line of analysis is the view that takings jurisprudence is concerned with regulatory transitions (Kaplow, 1986). Unpacking that idea more specifically, takings cases very often arise in transitional situations where regulation attempts to address the increasing marginal costs of congestion but encounters the very modernist principle that all owners should be treated in the same way (Rose, 2018). As discussed above, property owners generally use adjacent common resources in ways that are unproblematic at low levels but become increasingly costly as they grow more intense. If unchecked, those rising costs mean severe damage to common resources—drained or polluted aquifers, unhealthy air, urban gridlock, and so forth. But as also discussed above, regulation may be feasible politically only if early users are paid off, either by compensation or by permitting their continued use, while new users are blocked (Libecap, 1989). The resulting favor to old users over new encapsulates the conflict between the principle of equal treatment and the phenomenon of rising marginal costs. Arguments of equal treatment certainly are central in takings cases—“Why can’t I build a bulkhead on the waterfront? Look at all the others!” But so are marginal cost arguments—“More walls take away even more of the beach!” Given this seemingly inevitable conflict, it is no wonder that takings jurisprudence is also inherently unstable.

At least in theory, however, property itself suggests a way to overcome the equality/marginal cost dilemma, putting all common resource uses on an equal footing while dealing with increasing congestion costs. But this brings us to the next and final topic.

13.7 Property Law, Novel Property, and Public Service

With his usual pugnacious certainty, Jeremy Bentham (1802) asserted that secure property is a product of law and law alone. Bentham clearly overstated the case. Much property is informal, and much informal property intermixes with formal law or sometimes even successfully annuls it.¹⁰ But some types of property do require law. As this entry has argued, property depends on the forbearance of others, but for that reason, property claims must communicate their existence in ways that are recognizable to some relevant audience. Claims that give no perceptible real-world signals of their existence can only bridge the communication gap through legal institutions. That is the case, for the most part, for intellectual property, since intellectual innovations are intangible; it is also the case for passive uses of property like conservation easements over wildlands, which could not be known except through recording or registration systems.

A property innovation of particular interest emerged in the later twentieth century: the regulatory creation of a right entitling the holder to use some limited portion of a larger common resource or stock. An early version was New Zealand's effort to control overfishing by setting limits on the total catch but allocating tradable catch quota to individual fishers (Clark et al., 1989). One of the most celebrated programs of this sort came with the 1990 amendments to the US Clean Air Act, which created a cap on total sulfur dioxide (SO₂) emissions along with a limited number of "tradable emission rights" to be parceled out among users, notably power plants burning fossil fuels. With such a permit, the holder can either use up the permitted amount of SO₂ or alternatively sell any unused right to another user. Thus cap-and-trade puts a price on pollution and a reward for pollution reduction (Ackerman & Stewart, 1985). The program has been credited with encouraging innovation and substantially reducing SO₂ emissions, all at a lower cost than the prior command-and-control regulatory regime.

Wetland protection borrows from these ideas, as do some wildlife habitat protection programs, but these more complex resources present substantial obstacles to property approaches. Unlike SO₂, which is relatively uniform through the air, no two wetlands are alike in location, water quality, habitat characteristics, or flood control potential, among many other characteristics. In attempting to assess trades in complex resources like wetlands, regulators draw up detailed ecological criteria to compare wetlands destroyed by development to the credits that the developer can purchase for newly

¹⁰In a famous instance, the mayor of Chicago approved city residents' "dibs" system, informally allocating shoveled-out parking places to the adjacent residents who had cleared them, despite the formal rule that parking spots are open to the public (Eig, 2001).

created wetlands elsewhere. But in a trenchant analysis, James Salzman and JB Ruhl (2000) demonstrate that regulators face a constant dilemma: they can blur required details in order to encourage larger trading markets for these environmental goods—or contrariwise, they can add details that make comparisons more precise but at the cost of narrowing the markets for eligible trades. One might add that the more specific the details of these environmental trades, the less they resemble property transactions and the more they resemble contractual arrangements.¹¹

In spite of these issues, however, an intriguing aspect of tradable rights regimes is that they could in theory alleviate the marginal cost/equality conflict that haunts takings issues, at least in the environmental context. That is, if a total but limited usable quantity of a common pool resource use can be defined—for air, groundwater, wildlife—and if individual quota rights can be auctioned off to users as a kind of regulatory property, then the resource can be protected while maintaining equal costs to all bidders, with a possible subsidy to low-income users. The trick, of course, is distribution through an auction, as economists have suggested. But for the most part, auctions do not occur. In the vaunted SO₂ permitting, the initial quota went for the most part to existing polluters, just as the allocation of initial fishing permits have been allocated according to historic catch (Merrill, 2000). Tradable rights regimes thus have not overcome the new/old distributional issue that Gary Libecap identifies with political changes in property rights regimes.

Nor have they overcome other problems identified with political decisions. All these legally created property rights aim to promote some public benefit. Intellectual property should further innovation and new ideas; conservation easements and tradeable environmental rights should advance the protection of vital common pool resources. But their origin in law makes these forms of property subject to the vagaries of politics and to the associated public choice criticisms. Thus intellectual property has come under attack as a product of industry capture of legislatures, “proptertizing” innovation at the expense of the free flow of creativity (Boyle, 2003; Lessig, 2002). And thus environmental protections are criticized both as ineffective because blocked legislatively by business lobbyists or overreaching as the projects of zealotry and alarmism. It does not help a great deal when environmental protection is structured as property-like entitlements like cap-and-trade, which come in for critiques of their own, particularly on distributional grounds.

It is hardly surprising that encounters between public demands and property law have political connotations, even when public demands are packaged in new forms of legally created property. Indeed, all property law has some political element, given that any law is derived from governmental institutions. But it should not be forgotten that informal property arrangements are political too, with their own hierarchies and conflicts, even or perhaps especially in the context of allocating access to natural resources (Acheson, 1988).

¹¹ One might note a similar slippage from property toward contract in carbon markets that aim at reducing total carbon emissions. The detailed character of many deals makes them look more like contract than property. For some samples, see <https://www.ecosystemmarketplace.com/>

13.8 Conclusion

This chapter has described informal property norms as a contrasting institutional type by which one might better understand property *law*. In the real world, of course, many of the contents of property law and property norms overlap, as in the strictures against theft; and there are many institutional mixtures of the two, as in the mutual infusion of formal law and informal norms in colonial New England's boundary issues (Brady, 2019). But at the margins, informal property operates on a smaller scale and with greater particularity, density, and familiarity, while property *law* is standardized, abstract, and relatively distant. Unlike the informal property norms that surround everyday behavior, property law is less likely to govern routine activities among people who know each other. It rather acts as a background infrastructure, chiefly organizing dealings among strangers or sometimes being called into play as a set of "end-game norms" in conflicts between estranged neighbors or erstwhile business partners (Bernstein, 1996).

The chapter began with utopian rejection of property, but if property is not for utopians or saints, who is it for? One might say that property is just for us ordinary folk, the girl or fellow next door (Rose, 2007). Nevertheless, the ability to create property regimes, rather like altruism, reveals an aspect that even utopians might consider a half-step toward redemption, in that the central feature of all property regimes is the willingness of people to acknowledge the property of others. In that feature, property acts as a school for thinking that others have rights, and it hones people's skills in coming to agreement over matters of central concern while putting other matters aside. Those qualities, incidentally, are essential for democratic self-government.

Having said that, there is no denying that property communities, and especially informal property communities, are not necessarily inclusionary. People with well-understood property relations among themselves can behave with cavalier and cruel disregard of the claims of those perceived to be outsiders, completely ignoring the principle of respect that property demands as an institution (Sharfstein, 2012). In theory, modernist property *law* should extend institutional principles and practices to strangers and outsiders, but there are certainly instances in which property law has instead justified deviations from those principles and practices. This chapter has given a few examples, in the callous and self-serving disregard of indigenous people's claims (Banner, 2005, 2007) and in the sanctimonious approbation of private racial segregation (Brooks & Rose, 2013).

Some of these deviations have been at least partially rectified in recent years, suggesting that property law at least has room for self-correction and growth. But until Utopia arrives, perhaps this is all that can be expected of an institution created by and for ordinary folk.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

Bibliography

- Acheson, J. M. (1988). *The lobster gangs of Maine*. University Press of New England.
- Ackerman, B. A., & Stewart, R. B. (1985). Reforming environmental law. *Stanford Law Review*, 37, 1333–1365.
- Alexander, G. (2019). Property, dignity, and human flourishing. *Cornell Law Review*, 104, 991–1046.
- Alston, L. J., Libecap, G. D., & Mueller, B. (1999). Titles, conflict, and land use. In *The development of property rights and land reform on the Brazilian Amazon frontier*. University of Michigan Press.
- Anderson, T. L., & Hill, P. J. (1975). The evolution of property rights: A study of the American west. *Journal of Law and Economics*, 18, 163–179.
- Banner, S. (2005). *How the Indians lost their land: Law and power at the frontier*. Harvard University Press.
- Banner, S. (2007). *Colonizing the Pacific: Land, settlers, and indigenous people from Australia to Alaska*. Harvard University Press.
- Barry, M., & Whittall, J. (2016). Land registration effectiveness in a state-subsidized housing project in Mbekweni, South Africa. *Land Use Policy*, 56, 197–208.
- Bentham, J. (1802). Principles of the civil code. In *Theory of legislation* (5th ed.) (R. Hildreth, Trans.). Truebner & Co.
- Bernstein, L. (1996). Merchant law in a merchant court: Rethinking the code's search for immanent business norms. *University of Pennsylvania Law Review*, 144, 1765–1822.
- Blackstone, W. (1766). *Commentaries on the laws of England* (Vol. 2). Clarendon Press.
- Blume, L., & Rubinfeld, D. (1984). Compensation for takings: An economic analysis. *California Law Review*, 72, 569–628.
- Bogart, D., & Richardson, G. (2011). Property rights and parliament in industrializing Britain. *Journal of Law and Economics*, 54, 241–274.
- Boyle, J. (2003). The second enclosure movement and the construction of the public domain. *Law and Contemporary Problems*, 66, 33–74.
- Brady, M. E. (2019). The forgotten history of metes and bounds. *Yale Law Journal*, 128, 872–953.
- Brady, M. E. (2020). Property and projection. *Harvard Law Review*, 133, 1143–1214.
- Brooks, R. R. W., & Rose, C. M. (2013). *Saving the neighborhood: Racially restrictive covenants, law, and social norms*. Harvard University Press.
- Carrier, J. (1987). Marine tenure and conservation in Papua New Guinea. In J. Acheson & B. J. McCay (Eds.), *The question of the commons*. University of Arizona Press.
- Chang, Y.-c., & Smith, H. E. (2015). The numerus clausus principle, property customs, and the emergence of new property forms. *Iowa Law Review*, 100, 2275–2307.
- Cheney, A. (2006). *Body brokers: Inside America's underground trade in human remains*. Broadway Books.
- Clark, I. N., et al. (1989). The development and implementation of New Zealand's ITQ management system. In P. A. Neher et al. (Eds.), *Rights based fishing*. Kluwer Academic.
- Cohn, N. (1961). *The pursuit of the millennium: Revolutionary Messianism in medieval and renaissance Europe and its bearing on modern totalitarian movements*. Harper Torchbooks.
- Dagan, H. (2021). *A Liberal theory of property*. Cambridge University Press.
- Dana, D. A., & Merrill, T. W. (2002). *Property takings*. Foundation Press.
- Demsetz, H. (1967). Towards a theory of property rights. *American Economic Review*, 57(2), 347–359.
- DeSoto, H. (2000). *The mystery of capital: Why capitalism triumphs in the west and fails everywhere Else*. Basic Books.
- Eig, J. (2001). Snow causes Chicago residents of adopt dibs system for parking. *Wall Street Journal*, January 11, 2001.
- Ellickson, R. C. (1973). Alternatives to zoning: Covenants, nuisance rules, and fines as land use controls. *University of Chicago Law Review*, 40(4), 681–781.

- Ellickson, R. C. (1991). *Order without law: How neighbors settle disputes*. Harvard University Press.
- Ellickson, R. C. (1993). Property in land. *Yale Law Journal*, 102, 1315–1400.
- Ellickson, R. C. (2022). Stale real estate covenants. *William and Mary Law Review*, 63, 1831–1867.
- Field, E. (2007). Entitled to work: Urban property rights and labor supply in Peru. *The Quarterly Journal of Economics*, 122, 1561–1602.
- Fitzpatrick, D. (2006). Evolution and chaos in property rights systems: The third world tragedy of contested access. *Yale Law Journal*, 115, 996–1049.
- Galiani, S., & Schargrodsy, E. (2010). Property rights for the poor: Effects of land titling. *Journal of Public Economics*, 94, 700–729.
- Glennon, R. (2002). *Water follies: Groundwater pumping and the fate of America's fresh waters*. Island Press.
- Goffman, E. (1961). *Asylums*. Anchor Books.
- Hardin, G. (1968). The tragedy of the commons. *Science*, 162, 1243–1248.
- Heller, M. A. (1991). The tragedy of the anticommons: Property in the transition from Marx to markets. *Harvard Law Review*, 111, 621–688.
- Hibbetts, B. J. (1992). “Coming to our senses”: Communication and legal expression in performance cultures. *Emory Law Journal*, 41, 873–960.
- Hirschman, A. (1977). *The passions and the interests: Political arguments for capitalism before its triumph*. Princeton University Press.
- Hume, D. (1739). *A treatise of human nature*. Clarendon Press.
- Hyatt, W. S. (1998). Common interest communities: Evolution and reinvention. *John Marshall Law Review*, 11, 303–395.
- Joyce, J. A., & Joyce, H. C. (1906). *Treatise on the law governing nuisances*. Mathew Bender.
- Kaplow, L. (1986). An economic analysis of legal transitions. *Harvard Law Review*, 99, 509–617.
- Kapur, A. (2011). “Catch 22”: The role of development institutions in promoting gender equality in land law: Lessons learned in post-conflict pluralist Africa. *Buffalo Human Rights Law Review*, 17, 75–116.
- King, W. (2003). Illegal settlements and the impact of titling programs. *Harvard International Law Journal*, 44, 433–471.
- Knight, F. H. (1921). *Risk, uncertainty and profit*. Houghton Mifflin.
- Krier, J. E. (1992). The tragedy of the commons, part two. *Harvard Journal of Law and Public Policy*, 15, 325–347.
- Lee, B. A. (2019). Uncompensated takings: Insurance, efficiency, and relational justice. *Texas Law Review*, 97, 935–989.
- Lessig, L. (2002). The architecture of innovation. *Duke Law Journal*, 51, 1783–1801.
- Levinson, D. J. (2000). Making governments pay: Markets, politics, and the allocation of political costs. *University of Chicago Law Review*, 67, 345–420.
- Libecap, G. D. (1989). *Contracting for property rights*. Cambridge University Press.
- Libecap, G. D., & Lueck, D. (2011). Land demarcation and the role of coordinating property institutions. *Journal of Political Economy*, 119, 426–467.
- Mandelker, I. L. (1984). *Religion, society, and utopia in nineteenth century America*. University of Massachusetts Press.
- McAdams, R. H. (1997). The origin, development, and regulation of norms. *Michigan Law Review*, 96, 338–433.
- Merrill, T. W. (1985). Trespass, nuisance, and the cost of determining property rights. *The Journal of Legal Studies*, 14, 13–48.
- Merrill, T. W. (2000). Explaining market mechanisms. *University of Illinois Law Review*, 2000, 275–298.
- Merrill, T. W., & Smith, H. E. (2000). Optimal standardization in the law of property: The numerus clausus principle. *Yale Law Journal*, 110, 1–70.
- Merrill, T., Smith, H. E., & Brady, M. E. (2022). *Property: Principles and policies* (4th ed.). Foundation Press.

- Michelman, F. I. (1967). Property, utility, and fairness: Comments on the ethical foundations of "just compensation" law. *Harvard Law Review*, 80, 1165–1258.
- Mitchell, T. W. (2005). Destabilizing the normalization of rural black land loss: A critical role for legal empiricism. *Wisconsin Law Review*, 2005, 557–614.
- Neeson, J. M. (1993). *Commoners: Common right, enclosure and social change in England 1700–1820*. Cambridge University Press.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
- Penner, J. E. (1997). *The idea of property*. Oxford University Press.
- Pollock, F., & Wright, R. S. (1888). *An essay on possession in the common law*. Clarendon Press.
- Posner, R. A. (1973). *Economic analysis of law*. Little, Brown and Company.
- Priest, C. (2021). *Credit nation: Property laws and institutions in early America*. Princeton University Press.
- Radin, M. J. (1982). Property and personhood. *Stanford Law Review*, 34, 957–1015.
- Radin, M. J. (1986). Residential rent control. *Philosophy and Public Affairs*, 15, 350–380.
- Radin, M. J. (1987). Market-inalienability. *Harvard Law Review*, 100, 1849–1937.
- Reich, C. A. (1964). The new property. *Yale Law Journal*, 73, 733–787.
- Rose, C. M. (1991). Rethinking environmental controls: Management strategies for common resources. *Duke Law Journal*, 1991, 1–38.
- Rose, C. M. (1994). *Property and persuasion: Essays on the history, theory, and rhetoric of ownership*. Westview Press.
- Rose, C. M. (1999). What government can do for property (and vice versa). In N. Mercuro & W. J. Samuels (Eds.), *The fundamental interrelationships between government and property*. JAI Press.
- Rose, C. M. (2007). The moral subject of property. *William and Mary Law Review*, 48, 1897–1926.
- Rose, C. M. (2009). Invasions, innovation, environment. In B. D. Barros (Ed.), *Hernando de Soto and property in a market economy*. Routledge.
- Rose, C. M. (2013). Psychologies of property. In *Philosophical foundations of property law*. Oxford University Press.
- Rose, C. M. (2018). The equality norm meets the evolution of property rights in the law of "takings". *Social Philosophy & Policy*, 35, 149–172.
- Rose, C. M. (2020). Rations and takings. *Wisconsin Law Review*, 2020, 343–362.
- Rousseau, J.-J. (1750, reprinted 1950). Discourse on the origin and the foundations of inequality among men. In *The social contract and discourses* (G.D.H. Cole, Trans.). Dutton.
- Salzman, J., & Ruhl, J. B. (2000). Currencies and the commodification of environmental law. *Stanford Law Review*, 53, 607–694.
- Schleicher, D. (2021). Exclusionary zoning's confused defenders. *Wisconsin Law Review*, 2021, 1315–1372.
- Sharfstein, D. J. (2012). Atrocity, entitlement, and personhood in property. *Virginia Law Review*, 98, 635–690.
- Smith, H. E. (2003). The language of property: Form, context, and audience. *Stanford Law Review*, 55, 1105–1191.
- Smith, H. E. (2012). Property as the law of things. *Harvard Law Review*, 125, 1691–1726.
- Sugden, R. (2004 [1986]). *The economics of rights, cooperation and welfare*. Palgrave Macmillan.
- Titmuss, R. M. (1971). *The gift relationship: From human blood to social policy*. Random House.
- Van Houweling, M. S. (2008). The new servitudes. *Georgetown Law Journal*, 96, 885–950.
- Varley, A. (1987). The relationship between tenure legalization and housing improvements: Evidence from Mexico city. *Development and Change*, 18, 463–481.
- Webster, D. (1997). The animal smugglers: The looting and smuggling and hoarding of impossibly precious feathered, scaly and wild things. *New York Times*, February 16, 1997.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Lee J. Alston and Bernardo Mueller

Contents

14.1	Introduction	320
14.2	The Role of Property Rights	321
14.3	The Determinants of Property Rights: Some Background	322
14.4	The Demand for Property Rights	324
14.5	The Role of the State: Supply of Property Rights	328
14.6	Comparative Frameworks for Understanding Property Rights	331
14.7	Concluding Remarks	336
	References	336

Abstract

We focus on the state's crucial functions in defining, interpreting, and enforcing property rights, introducing a supply and demand framework to understand the conditions under which state-enforced property rights evolve from extant property rights. We examine the dynamics of how these rights adapt in the face of technological change, shifting preferences, demographic pressures, and other significant shocks that influence the efficiency of resource use. Property rights matter because they determine resource use. To be meaningful, property rights need to be enforced. While rights holders can initially enforce these rights, increasing competition for the resource prompts state intervention to reduce enforcement costs. Central to our discussion is the influence of the institutional context and the power dynamics among stakeholders, which critically shape the provision, enforcement, and transformation of property rights. Through a

L. J. Alston (✉)

Department of Economics, Indiana University, Bloomington, IN, USA

National Bureau of Economic Research, Cambridge, MA, USA

e-mail: ljalston@iu.edu

B. Mueller

University of Brasilia, Brasilia, Brazil

e-mail: [bmuellet@unb.br](mailto:bmueller@unb.br)

discussion of four different approaches from the New Institutional Economics literature, we explore the political economy of the emergence, enforcement, and evolution of property rights and their role in economic development.

Keywords

Property rights · Institutional change · Resource use · Commons · Political economy · State enforcement · De facto rights

14.1 Introduction

Property rights determine the incentives for resource use. Property rights consist of the set of rights to use and transfer resources. Property rights range from an absence of property rights under open access to a fully specified set of private rights. By open access, we mean that anyone can use the asset regardless of how their use affects the use of others. A full set of private property rights consists of the following: (1) the right to use the asset in any manner that the user wishes, generally with the caveat that such use does not interfere with somebody else's property right, (2) the right to exclude others from the use of the same asset, (3) the right to derive income from the asset, (4) the right to sell the asset, and (5) the right to bequeath the asset to someone of your choice. In between open access and private property rights are a host of commons arrangements. Commons arrangements differ from open access in several respects. Under a commons arrangement, only a select group is allowed access to the asset, and the use rights of individuals using the asset may be circumscribed. For example, a societal group, a village tribe or homeowner's association, may allow its members to place cattle in a common pasture but limit the number of cattle that any member may put on the commons.

One role of the state is to define, interpret, and enforce property rights. Definition of property rights is a legislative function of the state. Interpretation of property rights is a judicial function of the state. Enforcement of state defined property rights is a function of the executive. All three functions entail costs, and for this reason, some rights may be left by the state as open access. Moreover, many assets have multiple dimensions, and it is costly for the state to define property rights over all valuable dimensions. As such, some attributes may be either *de jure* or *de facto* left as open access. Individuals and groups have incentives to expropriate the use rights over attributes that the state leaves as open access.¹ Property rights also exist in the absence of the state defining property rights. Individuals or groups rights frequently claim property rights because though the state does not define the property rights, the rights still hold value for individuals or groups to claim and defend.

In many situations, individuals or groups use violence as a strategy to capture property rights. From the vantage point of societies, violence is wasteful and can be a

¹ For full detailed treatment of property rights see Barzel and Allen (2023) and Alston et al. (2018).

motivating force for the state to enforce property rights. Violence or threats of violence may also result when the state attempts to redistribute property rights.

In this essay, we present a framework for analyzing the determinants of the emergence and evolution of property rights. We start with the proposition that some actors must perceive that they can benefit from a change in the status quo set of property rights. Or as put by Demsetz (1967), “property rights arise when it becomes economic for those affected by externalities to internalize benefits and costs.” The Demsetz view of property rights has been termed by Eggertsson (1990) as the naïve view of property rights, as it ignores social and political institutions through which demands are filtered.

We stress that to understand the evolution of property rights, it is necessary to carefully examine the interplay between “demanders” and “suppliers.” History is replete with examples of conditions where the potential net gains from a change in property rights is not sufficient to prompt change because the costs of making all the side payments to parties with veto power dissipate the potential gains. The poor performance of economies throughout history suggests that the failure to specify and enforce a more optimal set of property rights is ubiquitous.²

Our purpose is to analyze how a country’s institutions determine how property rights evolve and whether this outcome will come through cooperation, conflict, or intermediation by the State. Together with noninstitutional factors such as the homogeneity of the population and relative endowments, institutions determine which groups have the ability to block change and whether it is possible to “buy out” the political gatekeepers through side-payments. In the same manner, institutions can facilitate cooperation by providing low cost means to make credible commitments.

In Sects. 14.2 and 14.3, we briefly discuss the role that property rights play in resource use and provide some background on the determinants of property rights. In Sects. 14.4 and 14.5, we develop a demand and supply framework for understanding the evolution of property rights, with special emphasis on the difficulties in changing property rights. While a demand for property rights emerges whenever the extant rights fail to internalize externalities, the supply of property rights is linked to political forces. In Sect. 14.6 we explore case studies in property rights through the lenses of political power (Engerman & Sokoloff, 2002; Acemoglu et al., 2001; North et al., 2009; Alston et al., 2012). In Sect. 14.7, we present some concluding remarks.

14.2 The Role of Property Rights

Property rights matter because they determine resource use. The more secure property rights are to the individual or group, the greater the incentive to maintain the value of the asset. Furthermore, more exclusive rights increase the incentive to

²This was the premise of North and Thomas (1973).

improve the value of the asset by investment, for example, in the case of land, this may entail the removal of rocks and stumps or using fertilizers. Having the incentive to invest may not be sufficient to include investment if individuals or groups are “cash poor.” In this situation, the ability to invest is aided if assets can be used as collateral to secure a loan. In developed countries, land has served as collateral for centuries.

Allowing sales as a property right may improve resource allocation in two ways: (1) allowing sales helps signal scarcity value and (2) markets enable those who value the asset most to purchase the asset. Of course, we need to be careful to note that by value economists include the ability to pay which historically and today is limited by the degree of development of mortgage markets. Unfortunately, in many parts of the world, mortgage markets are not well developed and investment suffers.

To be meaningful, property rights need to be enforced. One of the critical roles of the state is to enforce property rights. Enforcement by the state typically lowers self-enforcement costs which raises the value of the asset directly but also via the incentive for increased investment. A further impact of state enforcement is that asset holders can reallocate their labor from defending their asset to household or market production.

14.3 The Determinants of Property Rights: Some Background

So far, we have discussed the role of property rights in a static world. But, over time several factors affect the scarcity of resources. Scholars studying property rights have typically looked at cases where changes in technology, population, or preferences alter scarcity value. When resources become more or less scarce, the current property rights regime may entail dissipation of the rental stream from the asset. The losses that ensue create incentives for those involved to change property rights to a form more suited to the new reality. An example of this is from unitization of oil across Western states in the USA (Libecap, 1989). Unitization meant that the oil field was drilled as if there was one owner and the profits shared according to land owned over the field. As the losses from too many oil wells and too rapid drilling mounted, some states promoted unitization to limit losses. The requirements for unitization of a field varied greatly with Texas requiring unanimity and Wyoming a simple majority. The abilities of individuals, groups, and states to alter property rights in response to changes in scarcity go a long way toward explaining the economic growth and decline of nations. This is what Douglass North described as “adaptive efficiency.”

By examining how property rights change in response to the exogenous factors of technology, population, and preferences, scholars have derived insights for a theory of the emergence of property rights and, more broadly, institutional change and economic growth. Demsetz (1967) pioneered the empirical study of endogenous property rights development with his work on the introduction of property rights over beavers among First Nations in Eastern Canada. Demsetz argued that greater specificity and enforcement of property rights emerge as a response to the potential gains from having more secure rights to resources. Some early papers that followed

in Demsetz's wake studying the emergence of property rights to resources in the USA. US West are Anderson and Hill (1975) with respect to land rights, Dennen (1976) with respect to cattle, Libecap (1978) with respect to silver, and Umbeck (1981) with respect to gold in California.

Subsequent literature has analyzed contemporary cases of the evolution of property rights. Alston, Libecap, and Mueller (1997, 1999a, b, 2000) analyzed the evolution and effects of property rights in the Brazilian Amazon; Ensminger (1997) examined property rights arrangements in Kenya; Besley (1995) looked at the impact of property rights on land use in Ghana; Feder and Feeny (1991) examined property rights to land in Thailand; and Migot-Adholla et al. (1991) studied the impact of property rights in sub-Saharan Africa.

Recent research on property rights has been shaped by the credibility revolution in empirical economics that imposed a shift toward more rigorous research designs and methods, aiming to establish causal relationships and improve the reliability of empirical findings (Angrist & Pischke, 2010). This has prompted a more meticulous research design approach when evaluating the effects of property rights. In this work, property rights, along with broader institutions, frequently take the role of the primary variable of interest, the treatment effect, or the source of exogenous variation that provides the randomization needed for identification. Consequently, research on property rights gained popularity in applied work. While this shift toward more rigorous empirical scrutiny is commendable, it many times overly concentrates on what can be measured, sidelining, or overlooking important contextual detail and intricate institutional nuances or ignoring important issues where no natural experiment or other source of external variation are available.³

When modern methods of causal inference have been used not as the objective itself, but as tools for answering larger questions, this research has helped advance our understanding of the impact and dynamics of property rights. Dippel, Frye, and Leonard (2020), for example, study how limitations on property rights affect land use and investment. They analyze the historical policy of allotments in some American Indian reservations in the early twentieth century that restricted transfer and thus use for collateral and incentives to invest or to consolidate plots to achieve economies of scale. To measure these impacts, they take advantage of the natural experiment that resulted because the policy of allotment was not uniformly applied. On some reservations, individuals could receive their own parcels (it was allotted) but on other reservations land was held in common. Dippel et al. found that the land with restricted property rights had lower agricultural productivity, lower land values, and lower levels of investment in improvements from 1974 to the early twenty-first century, compared to neighboring plots held with full property rights. Importantly, the careful research design provides confidence that the estimated effect was indeed due to the property right restrictions and not to other confounding factors.

³See examples in Abad and Maurer (2021) and Kelly (2019) in the context of persistence studies, which often involve property rights, such as the Banerjee and Iyer's study of the legacy of the colonial land tenure in India (Banerjee & Iyer, 2005).

Another example of careful research design is Assunção et al. (2020). The authors estimate the impact of a credit policy by the Brazilian Central Bank in 2008 that conditioned the concession of subsidized credit to farmers in the Brazilian Amazon to proof of compliance with legal titling and environmental regulations. To identify the policy's impact, they used a difference-in-difference strategy comparing the amount of deforestation in counties along the border of the Amazon region with counties on the other side of that border and thus ineligible to receive the same subsidy yet comparable in many other ways. The results show that this policy, which limited property rights by curtailing the use of land to obtain credit, substantially reduced deforestation. A similar strategy using national borders as an identification device was used by Burgess, Costa, and Olken (2018) to estimate the impact of other government policies on deforestation. The identification came from national boundaries being exogenous to the actors.

Although we have focused here on examples related to property rights to land and natural resources, the same approach applies to other markets where scarcity makes property rights relevant, for example, Coase (1959) examined the broadcast spectrum and Mueller (2002) analyzed the property rights arrangements over Internet domain names.

14.4 The Demand for Property Rights

In this section, we present a framework for analyzing how the demand for property rights arises and may lead to the evolution of property rights. Alston, Libecap, and Schneider (1996) developed this general framework to analyze the demand for secure property rights over land in the Brazilian Amazon. Underlying the analysis is the notion that the potential rent generation from more secure property rights increases as the resource becomes scarcer. The difference between the rental streams from an asset with more as compared to less secure property rights generates a “demand” for secure property rights. In Fig. 14.1, the horizontal axis measures the relative scarcity of a given resource (from right to left), and the vertical axis measures the net present value that accrues to the owner of that resource. Line AH shows that the net present value (NPV) of the resource increases as it becomes scarcer. In the case of land, the measure of scarcity could be the distance of a plot of land to a market center, as transportation costs are often the main determinant of land value.

At point H, land is so far from the market center that the economic return given the transportation costs to market is zero. The segment BCEH represents the net present value of land under a commons arrangement.⁴ The value C represents the opportunity cost of the marginal laborer. As such, point G represents the economic

⁴We could further segment line CE into the return from a commons versus open access arrangement. The losses from an open access arrangement would increase as one moves toward greater scarcity.

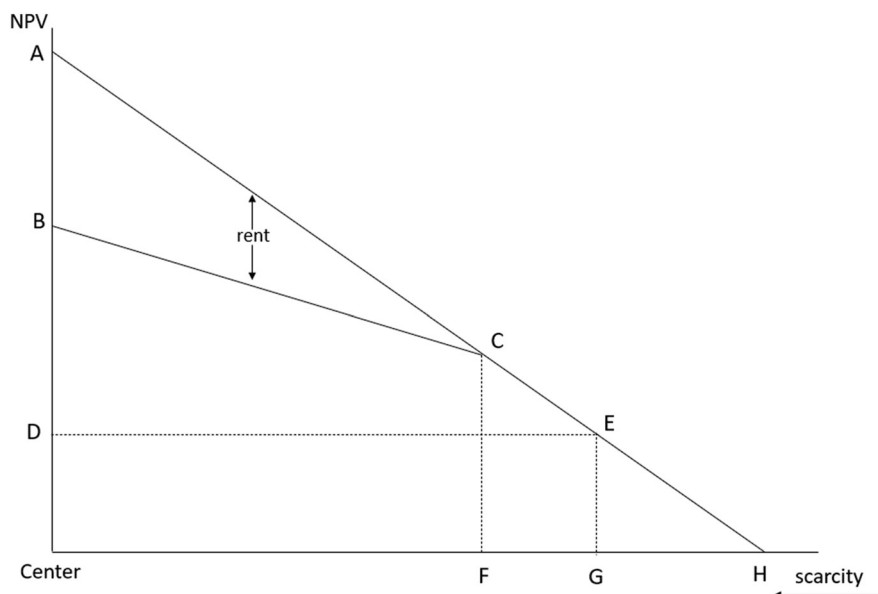


Fig. 14.1 The demand and evolution of property rights

frontier where, provided costs of migration are low, it becomes worthwhile for labor to migrate to the frontier. In our model, distance is the frontier, but it could be technological, for example, broadcasting on previously unused frequencies at the spectrum frontier.

At points between G and F, property rights are not formally defined or enforced, but this does not affect the return to the resource given that it is still abundant relative to the competition for it. As the net present value increases new users arrive yet, they can get access to the resource without detracting much from the use of those who were already there. At this stage resource, users will tend to be relatively homogeneous and nonstate property rights arise that are sufficient to arbitrate the existing competition. Any potential disputes are easily defused as accommodation yields higher expected returns than confrontation. Squatting prevails yet the absence of government-enforced private property rights does not impose significant costs.⁵ Note that the emergence of property rights at this point is already a case of institutional change.

At points to the left of F, the returns to the resource have risen and start to attract an ever-growing number of groups and individuals. This new migration typically

⁵ An example of informal arrangements includes Cattlemen's associations in the nineteenth century US West (Dennen, 1976). See Alston, Harris, and Mueller (2012), Anderson and Hill (2002), Eggertsson (1990), Ostrom (1990), and Umbeck (1981) for accounts of local groups allocating resources under "common" arrangements. See Smith (2000) for an analysis of "semicommons" arrangements.

brings heterogeneous settlers with differing amounts of wealth or human capital, nationalities, cultures, or objectives. The nonstate set of property rights that developed can no longer cope with the increased competition for the resource. It becomes necessary to expend effort, time, and money to assure continued possession of the resource and the income derived from it. This may involve incurring costs to exclude others or the costs from suboptimal uses. It may also include the costs of lobbying for changes to state-defined property rights. At some point, it becomes beneficial in the aggregate to have officially defined and enforced property rights. The pie-shaped area ABC represents the increased value of land with secure formal property rights versus the next best commons arrangement for property rights. ABC is the potential rent that forms the basis for the demand for property rights.

In our exposition, we used distance as the proxy for scarcity, but we could also use fertility of the soil, population density, or anything that generates competition for the asset as an alternative measure of scarcity.⁶ The framework is flexible to allow for changes in technology, preferences, environment, or new market opportunities. For example, if the demand for the output of the land increases exogenously, the divergence between the rental streams may emerge at point E instead of at C.

The increase in the net present value of the resource may not rise in a smooth and continuous manner as depicted in Fig. 14.1 but rather in discrete jumps. Nevertheless, the same logic holds. The shape of the present value curve will depend on the nature and characteristics of the change that affects the resource's relative scarcity. The main sources of change are technological innovations, changes in relative factor and product prices, and changes in the size of the market. An example of technological change affecting the returns to a resource is the invention of barbed wire that allowed nineteenth-century cattlemen in the US West to confine their cattle at relatively low cost, thereby increasing the return to selective breeding as well as better stocking practices (Anderson & Hill, 1975; Hornbeck, 2010). The effect of the opening of new markets to land is illustrated by the shift in comparative advantage to sugar production in nineteenth-century Hawaii that made it profitable to privatize land (La Croix & Roumasset, 1990). Libecap (1978) examines the legislative response by Nevada from 1858 to 1895 to secure the rights of claim holders to the potentially lucrative silver deposits in the Comstock Lode. To extract ore from the Comstock mine required considerable investments which in the absence of secure property rights would not have been forthcoming.

Many of the early studies on the evolution of property rights simply assumed that as the area ABC became sufficiently large property rights would emerge. This notion has been termed the naïve theory of property right, as it does not analyze the collective action problems of the politics that determine the supply of state-defined property rights (Eggertsson, 1990, p. 250). We will turn to the supply side in the next section, but here we want to delve in more depth into the determinants of ABC—the differential value of the asset under secure property rights versus the next best alternative set of property rights.

⁶The framework accommodates any force that either increases or decreases demand.

There are at least four incentives which lead to the dissipation of rents if secure state-defined property rights are not supplied at the optimal time: incentives to usurp property rights from the existing holder, incentives to defend, incentives to lobby for state defined property rights, and incentives for suboptimal use of the resource.⁷ Efforts to usurp take place when individuals or groups perceive an expected gain from taking the asset away from the current holder. Efforts include time, money, and violence. The more insecure the property rights of the current holder, the greater the likelihood that the redistribution transpires.⁸ Although the new holder may increase the value of the asset, the efforts to gain the asset are wasteful relative to a world where secure state property rights were already assigned or relative to the potential costs of assignment through the political marketplace.

Insecure property rights may also lead to dissipation through efforts by the current claimant to defend his asset against potential claimants. In the case of land, this may include fencing the plot, patrolling it, or hiring security guards. It may also include an otherwise nonoptimal allocation in labor supply: the claimant may spend more time on her plot and less labor in the market in order to maintain property rights (Field, 2007). The efforts to defend, together with the efforts by others to usurp, often lead to conflicts, which is one of the most wasteful forms of rent dissipation as the resource itself and human life may be lost in the process. Alston, Libecap, and Mueller (1999a, 2000) argue that the problem of land conflicts in Brazil since the 1988 Constitution resulted from conflicting legislation that created uncertainty over property rights to land. While the Constitution contains a beneficial use requirement for all land, which provides legal justification for squatters to occupy unproductive properties, the Civil Code, which is separate from the Constitution, allows the titleholder to request an eviction of squatters. Further dissipation comes in the form of violent conflict from titleholders and squatters.

Most dissipation arises from suboptimal use of the asset. For example, to the extent that deforestation represents beneficial use, claimants may deforest prematurely, which not only increases their private costs but may also entail social costs in terms of global warming or reduction in biodiversity.⁹ Claimants may also alter their cropping decisions because of tenure uncertainty. Without a secure claim, farmers are more likely to plant annual crops rather than permanent crops (Alston et al., 1999b; Besley, 1995; Feder & Feeny, 1991; Place & Hazell, 1993; Migot-Adholla et al., 1991). Investment is central to economic growth, and because property rights insecurity can skew investment in less than optimal ways, insecure property rights are a major impediment to a country's prosperity (Galiani & Schargrodsky, 2012; Leonard & Parker, 2021).

⁷Of course state-defined property rights are never perfectly enforced and often enforced poorly such that other de facto rights dominate.

⁸Insecure property rights may also reduce the value of the resource to the usurper; however, one would expect this effect to be smaller than the effect on the probability of successful appropriation.

⁹Allen (2002) argues that owners may purposefully reduce the value of their asset to lower enforcement costs.

When property rights are insecure, claimants may also invest too much or invest prematurely in hopes of strengthening their claim to the asset. For example, Anderson and Hill (1990) argue that homesteaders in the nineteenth-century US effectively paid for land, which was granted for free but required beneficial use, by bearing the costs of premature development. Homesteaders moved westward prematurely and also invested in the land earlier than they would have in the absence of the investment stipulations in the Homestead Act. Alston, Libecap, and Mueller (1999b) found evidence of the same kind of behavior in the Brazilian Amazon, as did Besley (1995) in Ghana.

Insecure property rights can also lead to underinvestment because insecure tenure may limit the ability of the claimant to invest, by preventing the holder from using the resource as collateral to secure a loan from a formal creditor. In addition, insecure property rights decrease the extent of the market thereby reducing the likelihood that the asset will be in the hands of the person who values it the most. In short, with insecure property rights society may not exploit all the gains from trade.

So far, we have examined some of the demand side determinants of property rights and indicated the impact of property rights on resource use. What is missing is a better understanding of how the demand side determinants of property rights get filtered through a country's political institutions.

14.5 The Role of the State: Supply of Property Rights

The early literature on property rights focused on cases where the changes in the scarcity of a resource lead to more precision in property rights at the optimal time—point C in Fig. 14.1. Though we have no way to truly gauge optimality, the rich countries of the world stand out in their protection of property rights. More secure property rights go a long way toward explaining economic prosperity. Wealthier countries have been able to solve a coordination problem in which the political actors refrain, particularly during crises, from acting in their short-run interests.¹⁰ More broadly, the issue can be couched in respect for the rule of law. We have some institutional proxies for the rule of law, such as independence of the judiciary or a constitutional court but fundamentally the backbone of the rule of law is a belief mechanism by the citizens and political elites that they will abide by the judgment of an independent third-party arbitrator. A set of universally shared beliefs in a system of checks and balances is what separates democracies with respect for the rule of law from the more recent rise and behavior of populist democracies.¹¹ Beliefs are at the

¹⁰Weingast (1997) highlights difficulties in establishing the rule of law, which is a broader set of rights than property rights. Nevertheless, the state with the support of the major political actors must solve the “time inconsistency problem” in which there will always be times when the people in power have an incentive to abridge property rights or erode the rule of law.

¹¹By a populist democracy, we mean that those in power were legitimately elected but hold views either to the far left or right of center.

heart of why some constitutions are a constraint on behavior while others are flagrantly ignored.

The protection of an existing set of property rights is easier than changing property rights. History is replete with examples of societies failing to change property rights in response to changing scarcity. The reasons for such institutional failures lie in the difficulties of compensating actors who can veto changes to property rights. Most changes in status quo property rights harm some people in a society. In a world of homogenous participants, for example, squatters in the Amazonian frontier or cattlemen in the nineteenth-century USA, it is relatively easy to establish property rights because all parties see themselves benefitting with more exclusivity. But when the parties involved are more heterogenous, some will see themselves losing from a change in the status quo set of property rights and will expand efforts to resist change to more precise state defined property rights. Yet, as scarcity increases, there is still pressure to establish more exclusivity. In the absence of third-party specification and enforcement, violence may be the least cost method for reducing the dissipation that would otherwise result.

We need a better understanding of the political and economic transaction costs associated with the state establishing or changing property rights that are more conducive to better economic performance, especially when it becomes obvious that the existing laws and regulations are not fostering economic growth. In many scenarios, special interests are able to either enact property rights legislation or block legislation so that they reap the gains. Yet society is made worse off by such activity. The question is why can't "we," the citizens or consumers, buy out special interests?¹² There are several possible explanations for why the state does not change property rights in lockstep with scarcity. Here we focus on three aspects.

1. Informational problems abound such that citizens are unaware of the policy moves that would improve on the status quo.
2. Even when aware, there are serious collective action problems which changing property rights entails.
3. Insecurity in "political" property rights prevents society at large from making the necessary side-payments in the political arena that would change property rights.

We will explore each in turn.

Given rational ignorance, it may be that many citizens are simply unaware of property rights arrangements that would improve societal welfare. For example, under the Homestead Act in the USA, settlers could acquire property rights to 160 acres of unoccupied federal land by residing on and "improving" the land. These homestead plots turned out to be economically too small and promoted externalities associated with wind erosion. Even after the great dust bowl of the 1930s, plots remained small through subsidies from the federal government. Why

¹²For many societies, the poor economic performance is explained by corrupt governments, who are more or less stealing from their own citizens. Here we focus on issues besides corruption.

did the federal government not move to reallocate land or at least not interfere with consolidation through markets? It appears that the answer rests with the information available to citizens and their beliefs in the virtues of small landholdings. This is coupled with the efforts of local politicians to maintain a population base (Hansen & Libecap, 2004).

Alternatively, people may be aware of the dissipation associated with the status quo arrangement of property rights, but it is in no one's self-interest to mount an organizational campaign to change the existing regulations. This is the classic collective action problem developed independently but almost simultaneously by Buchanan and Tullock (1965) and Olson (1965). The collective action problems are particularly acute in situations entailing multiple governments across international boundaries, for example, overfishing in international waters or global warming. The difficulties for international property rights are twofold: specification and enforcement. Specification is difficult because knowledge or beliefs about the state of the world differ (e.g., global warming), but even if beliefs are the same, preferences can vary across countries because of income (e.g., the USA versus Mexico) or simply tastes (e.g., the USA versus Germany on green issues). Collective action problems occur in both representative democracies and in dictatorial regimes. We have instances of both types of regimes not specifying and enforcing property rights at what would appear to be beneficial times. For example, the US squandered considerable oil reserves in the early twentieth century and Indonesia mowed through a large stock of their tropical hardwoods in the latter part of the twentieth century.

A third factor affecting the lack of emergence of state-defined property rights to assets is what we will term insecurity in political property rights. It may be that individuals are aware and willing to organize but there is no "market" for the emergence of property rights. Suppose that the winners from a status quo policy have the political power to veto or allow policy changes. Given their power, they would be foolish to acquiesce to policy moves that made them worse off, even if it was wealth enhancing. But they would allow such a policy move if they were compensated. The case of the Landless Peasants' Movement (MST) in Brazil is an example. The MST had devised a successful way to force the government to grant them land through the invasion of underused private properties. This strategy worked because a majority of voters sympathized with land reform objectives and interpreted the invasions as shirking by the government on land redistribution efforts. One could imagine a land reform policy that would pay off the group with land and subsidies to not invade, thus economizing on all the dissipation and violence involved in the invasion-based policy. Yet, this wasteful strategy of redistribution through invasions lasted for several decades (Alston et al., 1999b, 2000).

Why is it that we generally do not allow such side-payments? One answer is that transparent side-payments would undermine the legitimacy of the organization, whether the organization is the MST, a union, or a government. If the current property rights arrangement may be viewed as inferior to an alternative, but people "believe" that they should not pay to move to a better property rights arrangement, the result is institutional lock-in. A case involving the sale of water in the 1990s illustrates the difficulties in changing the status quo. The Imperial Valley Irrigation

District, which is a governmental unit that has jurisdiction over water, entered a contract to sell some of their water to the city of San Diego. The Imperial Valley Water District has property rights to water that are subsidized by the US taxpayers. As such they can sell water at prices higher than they pay. Interestingly, the members of the Imperial Water District imposed upon themselves that they would only sell water that they have conserved through better irrigation technologies. The interesting question is why didn't they fallow all of their land and sell their entire water allocation? We speculate that they were concerned about the political fallout that could have resulted in the district losing their current subsidy. In short, it appears as if they have secure property rights to the rental stream of water but not the clear "political" property rights to the stock.

Another factor promoting the insecurity of political property rights falls under the rubric of credible commitment. In representative democracies, politicians face the demands of constituents who may be harmed or benefited from a rearrangement of property rights. The demands of the majority of voters may not coincide with a more optimal arrangements of property rights, and politicians cannot commit to making side-payments over time to compensate the losers. Authoritarian regimes are subject to similar problems associated with catering to populist demands. A good example of this was the infringement in property rights by Peron in Argentina in the late 1940s. Peron imposed rent and price controls in the Pampas, the most fertile and productive agricultural producing area in Argentina. The punitive arrangement in property rights lead to a decline in investment which affected growth in the long run (Alston & Gallo, 2010).

A more cynical view of political behavior suggests that we do not want to encourage paying for changes in property rights because it would promote the creation and maintenance of nonoptimal property rights in order to be paid to move to a more optimal situation. Campaign finance and corruption around the globe may be testimony to special interests trying to "bribe" politicians to maintain or change property rights. In some instances, politicians may use part of the contribution to make side-payments (Norlin, 2003).

14.6 Comparative Frameworks for Understanding Property Rights

In this section, we describe four influential papers that have used property rights analysis—and institutional change more broadly—to explain countries' differential developmental paths in the past centuries. The frameworks under consideration are Engerman and Sokoloff (2002), Acemoglu and Robinson (2012), North, Wallis, and Weingast (2009), and Alston, Harris, and Mueller (2012). These four approaches are related and compatible, but each has a different focus and highlights different aspects of how property rights emerge, evolve, and impact society. In common, they have the emphasis on the political impacts of property rights on economic performance. This section differs from the earlier sections with its emphasis on economic development more broadly rather than the earlier sections dealing more exclusively with

property rights. Here we merge the literature on property rights with that on institutions.¹³

Engerman and Sokoloff (2002) laid the groundwork for what evolved into the field of persistence studies, aiming to empirically demonstrate how current institutions, beliefs, and outcomes are shaped by historical phenomena, often spanning centuries (Voth, 2021). They exploit the discovery and colonization of New World economies since the sixteenth century to understand why different countries adopted different institutions and how these choices affected comparative development. They argue that the choice of institutions imposed by colonizers on each colony was not determined by whim or culture but rather crucially constrained by the factor endowments unique to each locality.

Factor endowments are the climatic and geographical conditions in each colony, e.g., land fertility, suitability for various crops, ruggedness, rainfall, and transportation costs. The most important factor endowment, however, was the availability of labor, whether recruited or coerced into production. Given the exogenous distribution of factor endowments across different colonial enterprises, causal statements can be derived regarding the impact of institutions and property rights on economic performance. Colonies that were well suited for growing sugar cane, for example, such as Brazil and Caribbean Islands, gravitated to large plantations and slavery. This was not driven by culture or beliefs but rather by the economic nature of sugar production. Conversely, regions lacking a preexisting population for coerced labor or unsuitable climatic conditions for high-cost slave-dependent crops, such as the Northern USA, witnessed the emergence of more egalitarian economic activities, such as family farming.

The key to Engerman and Sokoloff's argument is that although the initial choice of institutions and property rights can be treated as exogenous, subsequent institutional change was crucially constrained by the early institutional choices. Where the initial conditions lead to the emergence of activities that induced economic and political inequalities, this produced elites that monopolized both wealth and power. In these cases, the choice of laws, constitutions, and policies established and enforced property rights for land ownership, banking, immigration, education, and voting *inter alia* that privileged the elites at the expense of the population in general. In the long run, this path of development led to initially high but unequal and unstable growth, for example, Argentina was one of the top ten wealthiest countries in the world from the late nineteenth century until the mid-twentieth century. Contrarily, where the initial conditions induced activities that produced more equal social and political structures, the early inclusive institutions and property rights fostered laws and policies that attracted more migration and strengthened the egalitarian nature of institutions, leading to greater economic growth and prosperity through greater incentives for innovation and entrepreneurship. Although it had

¹³ Interestingly, Douglass North took the path from focusing on property rights as the major determinant of economic development to moving to institutions as the key driver. Here we focus on the connection between property rights and institutions.

always been known that history matters, Engerman and Sokoloff reintroduced the interest in economic history to economists more generally by highlighting how persistent past choices can be sustained and by showing how these effects can be quantified and studied through the channels of institutions and property rights.

These arguments were taken up and popularized by Acemoglu, Johnson, and Robinson (2001) in their hugely influential paper on colonial origins (for which they won the Nobel Prize). Through its innovative use of settler mortality data as an instrumental variable for past institutions, this paper was a pioneer of the credibility revolution in econometrics and showed how even in areas where data is not easily available, careful empirical tests can uncover causal relations, in this case the impact of institutions on long-run growth. From this initial work, these authors developed a framework that was subsequently applied in a long series of papers to a wide variety of historical and current events and relations (summarized in Acemoglu and Robinson (2012)). The dependent variable in this framework is economic performance, which varies immensely across societies. The proximate determinant of performance are economic institutions, which include, for example, property rights to enter a market or to access credit. Their focus, however, is on political institutions and property rights, which ultimately determine who has the right to set or change the economic institutions. In general, for those countries that failed to achieve economic development, insecure property rights were the key determinant. The source of political institutions, in turn, is political power that can be determined in a *de jure* manner, such as by a constitution, and by *de facto* factors, such as when groups have physical or military strength, or some other characteristic that allows them to overcome collective action costs and thus organize more effectively.¹⁴

The key prediction of this framework is that where political institutions are more inclusive and egalitarian, there will be forces for similarly inclusive economic institutions, which are more conducive to secure property rights and subsequently investment and long-run economic development and prosperity. Conversely, where the configuration of *de jure* and *de facto* political power leads to extractive political institutions, inefficient economic institutions emerge and persist over time, even when clearly better ways to organize the economy are available. Like Engerman and Sokoloff, they emphasize the path-dependent nature of early property rights. Political power allows elites to choose economic institutions that are not conducive to long-run development in general. The stranglehold of political power also allows elites to lock in political institutions that assure perpetuation of those groups in power and continued control over institutions and property rights. In an idealized Coasean world, these inferior outcomes could be transacted away, with those harmed by the status quo purchasing the property rights that impede the economy from moving toward greater efficiency. However, as noted in the previous section, political transaction costs and the inability to make credible commitments make such transactions hard to carry out, locking in poor outcomes for much of the world.

¹⁴Drug gangs throughout Latin America currently have considerable *de facto* political power.

North, Wallis, and Weingast (2009) provide a different perspective on the role of property rights by focusing on the perennial threat of violence faced by societies through history. Wherever wealth and prosperity are produced, there is a threat of violence as usurpers from within or without the group seek to appropriate those assets by force. For economic progress to emerge, it is necessary to organize society in such a way that keeps those who have the capacity to use violence from opting to do so. While weaker groups can simply be destroyed or contained by force, groups that have greater capacity for violence must be provided with motives to hold back from using violence. That is, not turning to violence must be made incentive compatible for those groups that have the capacity for predation and conflict. The authors note that over history most societies that managed to overcome the problem of violence did so by forging institutions and property rights, that is, a social order, that contains violence by providing rents to powerful groups in a grand pact where each group refrains from force and theft.

Such a pact is often fragile, as any changes in relative power requires the terms to be renegotiated, otherwise stronger groups might opt to simply turn back to violence. North, Wallis, and Weingast call societies that achieve this type of social order “limited access societies” because the peace and stability is achieved by creating rents and privileges that by definition accrue to limited groups in society. In these societies, there are strong property rights for members of the elite but not for the majority of the population. While these rents and privileges are inefficient and distort the economy, they are the glue that creates an environment where economic activity can take place and generate prosperity, despite the latent threat of violence.

Most countries through history and still today are limited access societies, which the authors also call Natural States. A few countries, however, have managed to transition to a fundamentally different social order, the “open access order,” in which access to economic and political rents are achieved not through violence but through competition in both economic and political markets. Competition for rents provides incentives for innovation and entrepreneurship and assures that even when rents are achieved, there are forces that lead to their dissipation, spreading the gains more broadly.

How do countries transition from limited to open access orders? Given the few instances in which this has happened over history, it must be a difficult process. North, Wallis, and Weingast identify three “doorstep conditions” that a country must reach to be poised to successfully make this transition. The first is to enlarge the rule of law and assure that property rights accrue to greater swaths of society rather than just to a small elite. This enlarges the talent pool and the scale of markets. The second condition is to foster a state of perpetuity, that is, a permanence of organizational forms, both governmental and economic, despite the changes in individuals who occupy positions within them. This means that the property right to the power of office belongs to the office and not to the individuals that occupy those positions. The third doorstep condition is centralized control over violence and power in the state. Contrary to Natural States, open-access societies do not rely on a fragile balance of violence capacity to achieve peace. Even if a country reaches these three conditions, a transition proper is not assured. For current elites to willingly

expand economic and political access, it must be in their interest to do so—it must be incentive compatible—so that historical conditions must be just right to induce a successful transition. There is nothing automatic that ensures that countries that reach the doorstep conditions will transition into open access orders. Making the final leap is elusive and may require a window of opportunity that may be hard to predict or identify.

The final property rights framework we describe here recognizes the persistent effect of initial property rights highlighted by Engerman and Sokoloff and Acemoglu and Robinson. However, Alston, Harris, and Mueller (2012) show that additional factors related to the political context can influence whether those initial conditions persist or if other paths are followed. By analyzing the historical frontier land processes in Australia, Brazil, and the USA, they recognize two important possible mismatches that influence how property rights evolve and if conflict and rent dissipation is likely.

The first mismatch is that the frontier process is often simultaneously characterized by both *de facto* and *de jure* property rights, held by different groups. While the early settlers acquire *de facto* rights through use and possession, there is no guarantee that as rents increase those rights will be automatically recognized as *de jure* rights as presumed in the Demsetz model. The political process in many instances can decide to grant *de jure* rights to other groups that are not those that currently hold the land. In the three countries analyzed, the early settlers were larger claimants, instead of, as is often the case, individuals with lower opportunity costs. This was so because the early economic opportunities presented by factor endowments in these three cases required capital and experience: pastoralism in Australia, nineteenth-century coffee in Brazil, and ranching in the USA. In all three cases, at some later point, the government chose, for political reasons, to specify the right to smaller settlers. In all countries, politicians catered more to voters than to the wealthy because voters in general kept the politicians in power.

This creates the second mismatch, between who specifies the property right and who enforces the right. When the government both specifies and enforces the right, the probability of conflict and rent dissipation is low, as both enforcement and specification are *de jure*. When specification and enforcement are done by the individual or group that possesses the land, conflict and dissipation are also less likely, despite the *de facto* nature of those rights. It is when specification is *de jure* yet enforcement is left to the possessor that a mismatch arises. Because the government can be more influenced by political pressures at the center than those at the frontier, it may opt to concede official rights to groups other than those that hold the land, for example, the Brazilian government pursuing an aggressive land reform policy to appease urban voters' sympathy toward small farms. Because enforcing property right redistribution is extremely costly and contentious, the specification is often not adequately enforced. With two groups holding different expectations of rights for the same land, the result, as in the three case studies, was violence and rent dissipation. This fourth framework, as the ones before it, stresses how inextricably linked the supply of property rights is to political forces.

14.7 Concluding Remarks

In this essay, we presented a framework for analyzing the determinants of property rights. We can conceptualize the forces for changing property rights as the lost rent from a different set of property rights. In relatively homogeneous societies, the supply of property rights may come from the participants themselves. Examples include the codes established in mining camps or the rules established by Cattlemen's associations or the norms accepted by squatters. The supply of state-defined property rights typically emerges from an increase in the heterogeneity of the participants or an increase in the inherent rent of the asset causing a "race for property rights." The state generally has a comparative advantage in violence and hence better capabilities than private actors for the specification and enforcement of property rights. Enforcement by the state is never complete because it would be prohibitively costly in money and intrusion were it to attempt to do so.

The comparative advantage of the state in protecting property rights begs the question: if the state can protect citizens from stealing from one another, what protects the state from stealing from its citizens? A short answer is very little; over time and across space, many states have plundered their constituents to satisfy their self-interest. But this has not been the case in more developed countries in the world. In this essay, we suggest that the answer ultimately rests in the development of a set of beliefs by the citizens and political elites that they all will be better off in the long run by abiding by the rule of law. Day to day this may not be difficult; the stress comes during times of crises. A more definitive answer to this question is beyond the scope of this essay but it is surely the holy grail of many political scientists and economists.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute

References

- Abad, L. A., & Maurer, N. (2021). History never really says goodbye: A critical review of the persistence literature. *Journal of Historical Political Economy*, 1(1), 31–68.
- Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Crown Business.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5), 1369–1401.
- Allen, D. W. (2002). The Rhino's horn: Incomplete property rights and the optimal value of an asset. *The Journal of Legal Studies*, 31, S359–S358.
- Alston, L. J., & Gallo, A. A. (2010). Electoral fraud, the rise of Peron and demise of checks and balances in Argentina. *Explorations in Economic History*, 47(2), 179–197.
- Alston, L. J., Libecap, G. D., & Schneider, R. (1996). The determinants and impact of property rights: Land titles on the Brazilian frontier. *The Journal of Law, Economics, and Organization*, 12(1), 25–61.

- Alston, Lee J., Gary D. Libecap, Bernardo Mueller. 1997. "Violence and the development of property rights to land in the Brazilian Amazon" in *Frontiers of the new institutional economics*. Nye, J.V.C and J.N. Drobak, Academic: pp. 145–163.
- Alston, L. J., Libecap, G. D., & Mueller, B. (1999a). A model of rural conflict: Violence and land reform policy in Brazil. *Environment and Development Economics*, 4, 135–160.
- Alston, L. J., Libecap, G. D., & Mueller, B. (1999b). *Titles, conflict, and land use: The development of property rights and land reform on the Brazilian Amazon frontier*. The University of Michigan Press.
- Alston, L. J., Libecap, G. D., & Mueller, B. (2000). Land reform policies: The sources of violent conflict, and implications for deforestation in the Brazilian Amazon. *Journal of Environmental Economics and Management*, 39(2), 162–188.
- Alston, L. J., Harris, E., & Mueller, B. (2012). The development of property rights on frontiers: Endowments, norms, and politics. *The Journal of Economic History*, 72(3), 741–770.
- Alston, E., Alston, L. J., Mueller, B., & Nonnenmacher, T. (2018). *Institutional and organizational analysis: Concepts and applications*. Cambridge University Press.
- Anderson, T. L., & Hill, P. J. (1975). The evolution of property rights: A study of the American west. *Journal of Law and Economics*, 18(1), 163–179.
- Anderson, T. L., & Hill, P. J. (1990). The race for property rights. *The Journal of Law and Economics*, 33(1), 177–197.
- Anderson, T. L., & Hill, P. J. (2002). Cowboys and contracts. *The Journal of Legal Studies*, 31, S489–S514.
- Angrist, J. D., & Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2), 3–30.
- Assunção, J., Gandour, C., Rocha, R., & Rocha, R. (2020). The effect of rural credit on deforestation: Evidence from the Brazilian Amazon. *The Economic Journal*, 130(626), 290–330.
- Banerjee, A., & Iyer, L. (2005). History, institutions, and economic performance: The legacy of colonial land tenure systems in India. *American Economic Review*, 95, 1190–1213.
- Barzel, Y., & Allen, D. W. (2023). *Economic analysis of property rights* (3rd ed.). Cambridge University Press.
- Besley, T. (1995). Property rights and investment incentives: Theory and evidence from Ghana. *Journal of Political Economy*, 103, 903–937.
- Buchanan, J. M., & Tullock, G. (1965). *The calculus of consent: Logical foundations of constitutional democracy*. University of Michigan Press.
- Burgess, R., Costa, F. J.M., & Olken, B. A. (2018). *Wilderness conservation and the reach of the state: Evidence from national borders in the Amazon* (National Bureau of Economic Research Working Paper 24861).
- Coase, R. (1959). The Federal Communications Commission. *The Journal of Law and Economics*, 2(1), 1–40.
- Demsetz, H. (1967). Towards a theory of property rights. *American Economic Review*, 57(2), 347–359.
- Dennen, R. T. (1976). Cattlemen's associations and property rights in the American west. *Explorations in Economic History*, 13, 423–436.
- Dippel, C., Frye, D., & Leonard, B. (2020). *Property rights without transfer rights: A study of Indian land allotment* (No. w27479). National Bureau of Economic Research.
- Eggertsson, T. (1990). *Economic behavior and institutions*. Cambridge University Press.
- Engerman, S. L., & Sokoloff, K. L. (2002). Factor endowments, inequality, and paths of development among new world economies [with comments]. *Economia*, 3(1), 41–109.
- Ensminger, J. (1997). Changing property rights: Reconciling formal and informal rights to land in Africa. In *The frontiers of the new institutional economics*. Academic Press.
- Feder, G., & Feeny, D. (1991). Land tenure and property rights: Theory and implications for development policy. *World Bank Economic Review*, 3, 135–153.

- Field, E. (2007). Entitled to work: Urban property rights and labor supply in Peru. *Quarterly Journal of Economics*, 122(4), 1561–1602.
- Galiani, S., & Schargrodsky, E. (2012). Land property rights and resource allocation. *Journal of Law and Economics*, 54, S329–S345.
- Hansen, Z. K., & Libecap, G. D. (2004). Small farms, externalities, and the Dust Bowl of the 1930s. *Journal of Political Economy*, 112(3), 665–694.
- Hornbeck, R. (2010). Barbed wire: Property rights and agricultural development. *Quarterly Journal of Economics*, 125(2), 767–810.
- Kelly, M. (June 3, 2019). The standard errors of persistenceB28. Available at SSRN: <https://ssrn.com/abstract=3398303>
- La Croix, S. J., & Roumasset, J. (1990). The evolution of property rights in nineteenth-century Hawaii. *The Journal of Economic History*, 50(4), 829–852.
- Leonard, B., & Parker, D. P. (2021). Fragmented ownership and natural resource use: Evidence from the Bakken. *The Economic Journal*, 131(635), 1215–1249.
- Libecap, G. D. (1978). Economic variables and the development of the law: The case of western mineral rights. *Journal of Economic History*, 38(2), 399–458.
- Libecap, G. D. (1989). *Contracting for property rights*. University of Cambridge Press.
- Migot-Adholla, S., Hazell, P., Blarel, B., & Place, F. (1991). Indigenous land rights system in sub-Saharan Africa: A constraint on productivity? *World Bank Economic Review*, 5, 155–175.
- Mueller, M. (2002). *Ruling the root: Internet governance and the timing of cyberspace*. MIT Press.
- Norlin, K. (2003). *Political corruption: Theory and evidence from the Brazilian experience* [Ph.D. dissertation, University of Illinois].
- North, D. C., & Thomas, R. P. (1973). *The rise of the western world: A new economic history*. Cambridge University Press.
- North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.
- Olson, M. (1965). *The logic of collective action*. Harvard University Press.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
- Place, F., & Hazell, P. (1993). Productivity effects of indigenous land tenure systems in sub-Saharan Africa. *American Journal of Agricultural Economics*, 75, 10–19.
- Smith, H. E. (2000). Semicommon property rights and scattering in the open fields. *Journal of Legal Studies*, XXIX(1), 131–170.
- Umbeck, J. (1981). Might makes right: A theory of the formation and initial distribution of property rights. *Economic Inquiry*, XIX, 38–59.
- Voth, H. J. (2021). Persistence—myth and mystery. In A. Bisin & G. Federico (Eds.), *The handbook of historical economics* (pp. 243–267). Academic Press.
- Weingast, B. (1997). The political foundations of democracy and the rule of law. *The American Political Science Review*, 91(2), 245–263.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



State Regulation of Open-Access, Common-Pool Resources

15

Gary D. Libecap

Contents

15.1	Introduction	340
15.2	The Tragedy of the Commons and the Role of the State	343
15.3	State Regulation with Homogeneous Lobby Parties: Western US Mineral Rights	347
15.4	State Regulation with Less Homogeneous Parties: US Oil and Gas	348
15.5	State Regulation with Heterogeneous Lobby: US Fisheries	351
15.6	State Regulation of the Common Pool: Concluding Remarks	355
	References	357

Abstract

I explore the origins and impact of state regulation of the common pool (open access), one of societies' most challenging problems. Where resource values are low, group size is small, and users are homogeneous in cost and objectives, informal common-property solutions are feasible. Where these conditions are not met, state intervention can formally recognize and enforce group restrictions, recognize private property rights and exchange, or impose regulatory restrictions on access, inputs, and/or output. Government intervention is the outcome of interaction among self-interested lobby groups, politicians, and bureaucratic officials. Where the user lobby is influential and similar in resource goals, politicians and agencies respond with straightforward group or individual economic property rights. Where the lobby has heterogeneous objectives and fragmented influence, politicians and bureaucratic officials have more latitude in devising regulatory responses. The resulting property rights regimes and/or regulations incompletely address economic rent dissipation. Even so, these second-best arrangements are improvements over pure open access.

G. D. Libecap (✉)

National Bureau of Economic Research, University of California, Santa Barbara, CA, USA

e-mail: glibecap@bren.ucsb.edu

© The Author(s) 2025

C. Ménard, M. M. Shirley (eds.), *Handbook of New Institutional Economics*,

https://doi.org/10.1007/978-3-031-50810-3_15

339

Keywords

Common pool · Common property · Property rights · Open access · Lobbying · Self-interested behavior · Mineral rights · Oil and gas · Fisheries · Prorationing · Individual transferable quotas · Unitization

But still another inquiry remains; one often agitated by the more recondite Nantucketers. Whether, owning to the almost omniscient look-outs at the mast-heads of the whaleships, now penetrating even Behring's straits, and into the remotest secret drawers and lockers of the world; and the thousand harpoons and lances darted along all continental coasts, and so remorseless a havoc; whether he must at last be exterminated from the waters, and the last whale, like the last man, smoke his last pipe and then himself evaporate in the final puff. (Melville, 1922, p. 425)¹

15.1 Introduction

The passage from Herman Melville's 1851 novel, *Moby Dick*, describes the global race to exploit the enormous, rich whale stocks that roamed the world's oceans. Whales were *open-access* resources, subject to intense competitive exploitation. They provided valuable oil, baleen, and ambergris. Prior to electricity, candles from whale-oil wax and whale-oil lamps were the most sought-after sources of illumination. A single large whale could yield up to 120 barrels of oil, and they were harvested in the tens of thousands by whaling vessel crews annually over the nineteenth and early twentieth centuries. By the 1930s, competitive slaughter had depleted once seemingly inexhaustible whale populations, driving up costs and lowering catches and returns. Moreover, petroleum had displaced whale oil in most markets. The industry and whaling ports like Nantucket and Martha's Vineyard fell into economic decline.

This competitive "race to fish" illustrates the *Tragedy of the Commons* (Hardin, 1968) and the economic and social losses inherent in it. With open access and no constraints on entry, whales were harvested at a rate and to a level far beyond that which would have maximized economic returns. Had whales been "owned" in some manner, whaling would have been curtailed (not halted) far sooner. Stocks would not have been so depleted, and economic losses would have been reduced, and indeed, some whaling would have continued at a diminished rate from some ports. There would have been efforts to conserve the stock and to explore new markets.

As indicated in this case, open-access, competitive exploitation can be incredibly damaging to valuable resources and the human populations that depend upon them. Even though human wealth, resource rents, and stocks are at stake, open access often seems to be ineffectively addressed across time and space. Devising and assigning

¹I am grateful to Jim Smith who brought this passage to my attention. I also thank the editors Mary Shirley and Claude Ménard for their helpful comments.

access and use rights of some type privately or within a group, recognized by the state, is the most straightforward way of confronting the problem (Scott, 1955). Informal controls are observed for many local common-property resources when the resource is of limited value and the costs of restricting entry are low as argued by Ostrom (1990) and Schlager and Ostrom (1992). Local group arrangements, however, may not easily scale up when the resource is very valuable and large, the parties are numerous and heterogeneous, and new entry remains feasible (Cox et al., 2010). Under these circumstances, more formal individual property rights or regulatory controls that rely upon the police power of the state are more effective in addressing open-access.

Effective mitigation of common-pool losses requires a property right of some type—the granting, recognition, or refining of group rights or tradable, durable private ownership rights (Arnason, 2007). If comprehensive, property rights internalize benefits and costs in exploitation and eliminate externalities (Cheung, 1970). Alternatively, regulation can be implemented that involves limiting entry, assigning input and related capital controls, restricting exploitation times, and/or setting output restrictions that may include issuing short-term use privileges. Depending on its structure, regulation may or may not sufficiently internalize net benefits in private entry and production to avoid a race to exploit. It still may be a second-best improvement over pure open access (Homans & Wilen, 1997).

With all government responses, there are distributional consequences. Whether through the assignment of private property rights, enforcement of common property, or imposition of state controls, effective remedies to open access require that some parties be denied. Political reaction to distributional outcomes inherent in open-access restrictions involves mobilization of lobby groups. Competition among these groups focuses on limiting common-pool losses and distributing the associated costs and benefits. Politicians react to and depend upon the support of key constituents and the regulatory bureaucracy requires the backing of political mentors and outside parties for mandates and budget authorizations. What on the surface might be viewed as a public-good response to open access, underneath it is molded by competing lobbying interests, political objectives, and constraints (Stigler, 1971, 1975; Peltzman, 1976; Karpoff, 1987).

For these reasons, an understanding of government regulation of common-pool resources requires examining the composition and organization of constituents and their desired political actions. This information, in turn, provides background for assessing how observed regulations come about and why they take the forms that they do—even those that appear on their face to be inefficient. Welfare conclusions, however, require caution because the transaction costs of lobbying and political responses are real and minimizing those costs along with open-access losses must be considered in aggregate assessments.

This chapter describes observed institutional differences and provides hypotheses as to why they occur. The open-access problem is outlined and followed by discussion of the political economy of regulation. Three cases are presented to illustrate how variation in industry structure, resource characteristics, and heterogeneity of interests in combating open access determines outcomes.

The empirical cases examined in this chapter are drawn from the USA but likely are representative of general patterns. The cases begin with US mineral rights where the potentially harmed industry was uniform in seeking remedies, there were no competing outside interests, and representatives turned to the state for refinement of privately negotiated individual mineral rights. Open-access losses were minimal, the legal property rights arrangements adopted persisted, and the mining economy expanded. Indeed, the USA produced more under the observed secure property rights regime than its resource endowments alone might have predicted (David & Wright, 1997).

A second case involves US oil and gas producers, who also faced severe common-pool costs. The industry, however, was more heterogeneous among large and small producers, the nature of their holdings (oil or natural gas), and the losses they faced. Some parties benefitted from open access or were disproportionately harmed by proposed solutions. Distributional differences across firms created conflicting incentives for addressing the open-access problem. These differences, in turn, affected the formation of cohesive lobby groups and regulatory responses from politicians and agency officials. With competition among lobby groups, regulations were delayed and took forms to mitigate both distributional concerns and open-access losses. These combined objectives led to regulatory regimes that seem inexplicable without detailed examination of industry, political, and bureaucratic positions. Ultimately open-access was addressed in a complex fashion, the industry expanded with limits on production spillovers across producers, and overall, there were aggregate net economic gains.

The third case involves US fisheries, where the affected parties and resource stocks have been far more fragmented than in the other cases, leading to differential positions on various property rights/regulatory options. Moreover, fisheries have been viewed broadly as public resources or common property, raising the political transaction costs of assigning well-defined private property rights where they might have been useful (Allen, 2000). Regulatory agency officials and outside parties aligned with them have played an integral role not found in mining or oil and gas. Lobbyists from small and large fishers, sports, and commercial fishing groups, their suppliers, processors, and marketers, as well as marine biologists, other academics, and consultants, have sought potentially competing arrangements. While more formal property rights in fisheries were implemented in New Zealand and Iceland, in the USA, comparatively weaker use rights were adopted slowly and in a more limited manner across fisheries. Further, the eight US regional fishery management councils that draft and administer management plans are composed of multiple voting and nonvoting members. These include federal and state regulatory agents and other government officials, as well as those nominated by state governors, drawn generally from industry, academics, and environmental NGOs.

Although the statutory aim of US fishery laws was collaboratory management, such a heterogeneous mix has made common-pool regulation more political, costly, and uncertain than it might have been. Moreover, when fisheries are treated as a public resource, rather than primarily as an economic one, biological objectives, such as maximum sustainable yield, rather than maximum economic yield, become

regulatory objectives. When biological objectives dominate economic goals, open-access fishery controls will not necessarily be consistent with economic efficiency or wealth maximization.²

15.2 The Tragedy of the Commons and the Role of the State

Open-access or common-pool problems are long-standing ones across all societies and times. When there are few constraints on access and use of a valuable resource, agents have an incentive to rush to capture immediate rents and to overexploit. The rule of capture dominates. There is no waiting for less wasteful ways of producing, conservation, or rising prices. The resource cannot be distributed to more productive agents because there are no property rights underlying exchange to do so. There can be violent confrontation among competitors (Umbeck, 1977; Alston et al., 2000). With sufficient entry by enough parties that are relatively uniform in cost, the total rental value of the resource is wasted, and all parties eventually are made worse off.³

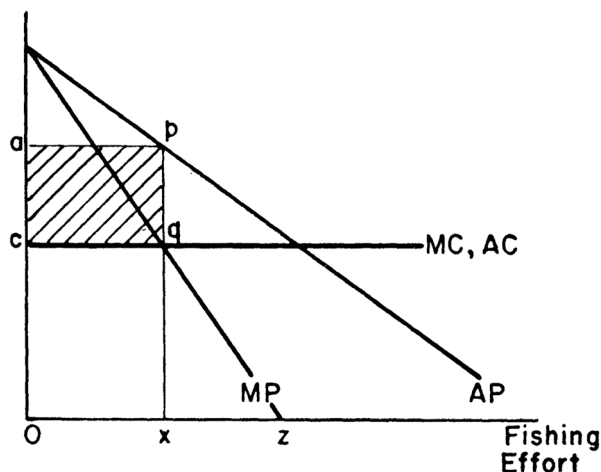
H. Scott Gordon (1954, p. 124), who studied open-access fisheries, described the problem: “There appears then, to be some truth in the conservative dictum that everybody’s property is nobody’s property. Wealth that is free for all is valued by no one because he who is foolhardy enough to wait for its proper time of use will only find that it has been taken by another. . . The fish in the sea are valueless to the fisherman, because there is no assurance that they will be there for him tomorrow if they are left behind today.” Similarly, Garrett Hardin (1968, p. 1244) in his famous article in *Science* outlined the problem with herders and an open-access meadow: “Therein is the tragedy. Each man is locked into a system that compels him to increase his herd without limit—in a world that is limited. Ruin is the destination toward which all men rush, each pursuing his own best interest. . .”

Economists, including Frank Knight (1924), H. Scott Gordon (1954), Anthony Scott (1955), and Steven N.S. Cheung (1970), analyzed the problem of the commons in their classic articles. Using open-access fisheries to define the issue, Gordon (1954, p. 130) discussed the motivation and effects of infinite entry by homogeneous fishers, operating under the rule of capture. Following his Fig. 15.1, reproduced below, entry occurs so long as the private marginal costs of access and harvest are

²As described below, one political and academic aim of fishery management that has raised the transaction costs (Allen, 2000) of assigning property rights for more effective fishery management has been the effort to protect small fishing communities from industry consolidation and competition from fishers with greater access to capital, larger vessels, and reduced labor requirements. This “equity” problem also exists in agriculture and elsewhere in the economy as economies of scale, price shifts, and new technologies change the location and nature of production. Moreover, excess entry and overcapitalization in the industry means that effective open-access controls require reduction in labor and capital inputs, all of which impacts fishing communities.

³Johnson and Libecap (1982) outline how some resource rents are maintained when the parties are heterogeneous in cost. Low-cost entrants earn rents and outcompete less-productive parties, reducing their expected returns, and limiting overall entry.

Fig. 15.1 Open-access exploitation. (Source: Gordon (1954, p. 130))



less than or equal to average returns for all parties. Continued exploitation eventually dissipates all economic rent. Alternatively, rental values are maximized when productive effort or entry is limited to point, x , an outcome possible with property rights to the resource.

Not all resources are so affected. Low-valued ones do not attract enough entry and production to generate much waste, so that open access alone is not sufficient to lead to an economic problem. But as resource values rise or exploitation costs fall, the potential for loss increases. Accordingly, parties have an incentive to bargain to assign and enforce group or private property rights to limit access and use and with the latter institution to facilitate trade. In the case of common property, after examining multiple settings, Elinor Ostrom (1990) found that local collective action to control access and production could reduce the losses of the common pool. Such collective action was feasible when user groups were relatively small, homogeneous in cost, and resource objectives, costs and benefits were shared proportionately, and the resource was of sufficiently low value to limit competitive entry (Cox et al., 2010).

Under these circumstances, it can be relatively easier for a small group of similar people who have a history of interaction with one another to gather and interpret information about the resource's status and to agree upon the types of uses and constraints necessary to conserve it. They also can accept the distribution of the costs and benefits (and ultimately, of wealth and political influence) within the community that is inherent in any informal definition and assignment of internal property rights or use privileges. Community management of regional, typically small-scale, agricultural irrigation water, pastures, or inshore fisheries provides examples of successful mitigation of the losses of the commons (Ostrom, 1990; Schlager & Ostrom, 1992).

Scaling up, however, can be challenging for these local arrangements. They rely upon stability in the group's social hierarchy, small membership numbers, and

homogeneity of goals for bargaining and often in maintaining the use of known technologies and capital equipment. New production methods and capital may change the scale of productive units, labor requirements, and the knowledge required for operation, distribution, and marketing. All these factors, introduced by markets, can be disruptive. Common-property arrangements typically do not support market trades and significant reallocation that can undermine local cohesion and traditional production structures and methods. Rising prices, new markets, and changes in economic processes and units invite entry by more numerous, dissimilar parties. When this occurs, community agreements break down. Local social values and political ties are eroded by new pecuniary opportunities. Additional parties may have access to new resource information, markets, and output opportunities previously unknown to past members. Negotiations within and across groups of resource users become more costly. Additional claims to resource rents must be resolved and more parties must be excluded. These conditions raise the transaction costs of agreement on any type of property rights (Allen, 2000).

When these conditions arise, it becomes challenging to reach a policy consensus and to enforce agreement. Parties who have adapted well to open-access and local arrangements may anticipate that proposed property rights or regulations will make them worse off relative to the status quo (Johnson & Libecap, 1982). They see few benefits from the new regime and attempt to block or seriously modify it. Other parties perceive that planned property rights remedies of any type will harm them unless compensated, but compensation is difficult to measure and agree upon. Outside observers who typically do not bear the losses of open access also can oppose the introduction of property rights and shifts in production technologies. They are attracted to historical, informal community arrangements and oppose the rise of more market-based, perhaps less-picturesque, institutions. Further, a well-defined property rights regime reduces the need for bureaucratic/political management, an arena where nonowners or nonusers may lobby to achieve their particularistic policy objectives at low cost. As Hannesson (2006) points out, inefficiencies create their own constituencies.

Equity concerns can be a major impediment to the assignment of more formal property rights, particularly as resource values rise and undermine previously effective informal controls. Formal property rights demonstrate ownership, with those excluded opposing them, even if their actions result in broader open-access losses. Their private injustices are more apparent to them than are aggregate losses. This situation is illustrated in Hardin's solution (1968, p. 1247) to the tragedy of the commons, which is coercive regulation of individual behavior—"mutual coercion, mutually agreed upon" to escape the "horror of the commons." He notes, but does not develop, the critical problem of distributional outcomes that are not acceptable to key parties. He asserts that "injustice is preferable to total ruin." Despite Hardin's assertion, total ruin is *not* so obvious to all parties in many common-pool settings. Injustice is *not* acceptable to them, but their proposed remedies can weaken the entry and harvest constraints needed to address open access.

Despite all of this with much at stake, private agents have an incentive to negotiate to address the common pool by establishing more complete constraints

on resource access and use through the definition and assignment of property rights. As resource values rise, stricter and more precise property rights of some type—negotiated private or group rights or regulatory-generated privileges—are required to control competitive open-access rental losses. State recognition and enforcement initially can involve settling disputes among existing parties and policing trespass or other infringements by outsiders to common property or informal private property regimes. With state acknowledgment, further internal negotiations beyond enforcement can take place to refine common property or private property rights in a manner outlined by Demsetz (1967) and empirically examined by Libecap (1978).

In the assignment of property rights and determination of production levels, there are economic optimal rates of exaction for both renewable and nonrenewable resources. These are based upon time preferences, interest rates, costs, projected resource prices, technologies, product substitution, as well as capital and technology trajectories as supplies diminish (Dasgupta & Heal, 1974; Hannesson, 1986). Optimal extraction rates, however, require stable property rights or other controls on entry and production. Absent property rights, resource use is governed by open-access imperatives, not long-term value maximization.

Further, optimal exploitation rates are unlikely to be consistent with popular notions of “sustainability.” Sustainability is an equilibrium, engineering concept, not an economic one. Economic markets are not equilibrium institutions. They are designed to incorporate new information on prices, costs, and technologies and reallocate resources or change output incentives. *This is their key contribution to economic welfare.* As such, markets are inherently destabilizing, affecting rates of production and conservation beyond any sustainability notion.⁴

Demsetz (1967) focused on the emergence of economic property rights that could be exchanged in markets, with the state having a restricted role in formal rights definition and enforcement. In potential open-access conditions, the introduction of property rights institutions advance welfare because they explicitly enlist user incentives for resource conservation and when tradable, facilitate the redeployment of valuable assets as economic conditions change. With trade, both the holder of the property right and society benefit as the resource moves to higher valued uses. Trade equalizes marginal willingness to accept with marginal willingness to pay in reallocation so that any market adjustment is Pareto improving. Further, owners are motivated to invest in and conserve the resource and to self-enforce conservation. Resource assets are valuable only so long as they are not depleted through excessive, noneconomic exploitation.

Even so, formal property rights solutions to open access are observed in a surprisingly narrow set of cases. In many cases, state regulation rather than private property rights is implemented through explicit, not-exchangeable access restrictions

⁴Differences in these concepts arise, for example, in fisheries where biologists and population ecologists often support harvest at maximum sustainable yield (MSY) not maximum economic yield, to maintain the stock at a targeted level. Economic exploitation, however, may be less or more than MSY depending upon costs, interest rates, and expected prices.

and authorized time periods, such as exploitation/production periods or seasons, input constraints, and output or harvest limits. Because they are imposed and not tradable, these restrictions may not generate clear user incentives to protect, conserve, and exchange the resource. Moreover, the longevity and nature of regulatory controls may be uncertain due to shifting political coalitions.

Politicians and agency officials have shorter-time horizons than most resource exploitation possibilities. They generally are not owners and may not benefit from transferring the resource across time. Further, in the absence of property rights and markets, such transfers are very difficult. Finally, when costs and benefits of resource use are not internalized by agents under regulation, compliance may be limited. Despite such drawbacks, observed regulatory arrangements can advantage particular parties, including state officials and their constituencies. Narrow benefits accrue while imposing generalized opportunity costs. Nonetheless, broad user regulation may be a second-best improvement over open access albeit with lost opportunities (Homans & Wilen, 1997).

15.3 State Regulation with Homogeneous Lobby Parties: Western US Mineral Rights

The case of Western US mineral rights clearly outlines the process of institutional change for property rights that can occur to address open access and the role of government in that progression as resource values rise, technology changes, and the number of competitors increases. It also makes clear how a homogeneous user lobby group could secure beneficial property rights refinement that advanced resource values. In the case at hand, the resort to the state was to more effectively address open access and not to redistribute mineral rents.

Nevada and the fabulously rich Comstock Lode (1858–1895) provide a laboratory for analyzing Demsetz's (1967) hypotheses about the emergence of property rights and subsequent adjustments as costs and benefits change. At the time of major ore discoveries in 1858, population numbers were small, the mineral-bearing region was narrow, and there were no existing mineral rights, informal or formal. Initially some 100 prospectors worked along with Comstock with few conflicts over marginal land under common-property regimes or mining camps. Things changed dramatically when valuable new silver ore discoveries were made in 1859 and the number of claimants to resource rents jumped to 20,000 by 1861. Libecap (1978) follows the 39-year period of major ore discovery and production when there was rapid legal change regarding mineral rights. Mine owners were the major advocates, and there was general support among the population that benefitted from growth in the mining industry. Relative to the rest of the population, mine owners were a relatively small and organized group whose interests generally coincided, making them an effective lobby (Olson, 1965, pp. 53–65).

To avoid competitive open-access losses, mining camps were created at first in 1859 to outline common-property local rules for claiming and working mineral land with enforcement by local courts and miners' groups. Even so, conflicts over

ownership of subterranean ore deposits grew, threatening the ability of mining companies to secure funds for new technologies for ore extraction as production shifted to capital-intensive, deep underground excavation. By 1861, the newly formed Nevada Territorial Government and later in 1864, the state government enacted legislation at the behest of leading mining companies to clarify the nature of mineral claims and subsurface boundaries that was particularly important if there were parallel ore veins. Territorial, and later state, courts adjudicated disputes over claim perimeters, property abandonment, and taxation. There is little indication of serious disagreement among the leading mining companies in lobbying the legislature or in turning to the judiciary. Between 1861 and 1895, mineral rights in Nevada became precisely specified and secure though litigation and legislation (Libecap, 1978, pp. 349–350). Production grew so that by 1889 the Comstock region had produced over 12% of all the gold and silver ever produced in the USA (Libecap, 1978, p. 354).

Moreover, the unified industry and responsive politicians (who also often were mine owners) led to the successful enactment of the US Mining Laws of 1866 (14 Stat. 251) and 1872 (17 Stat. 91–96) that recognized existing private mineral claims on federal lands and authorized further private exploration and production. This was the *last* major federal lands legislation that endorsed easy private claiming of the federal estate.

15.4 State Regulation with Less Homogeneous Parties: US Oil and Gas

Subsurface oil and gas reservoirs are potential common-pool resources when multiple parties competitively extract migratory hydrocarbons from a subsurface reservoir. Under the rule-of-capture in the USA, ownership of subterranean oil is obtained only upon extraction. The open-access problem occurs because ownership of mineral rights above a reservoir often is fragmented, with many firms obtaining access leases from surface owners to competitively capture the same oil. Under these circumstances, firm owners have incentives to maximize the economic value of their leases, rather than that of the whole reservoir. They competitively drill and drain, including the oil of their neighbors, to increase their private returns, even though these actions reduce the overall value of the reservoir. Capital costs are driven up with excessive investment in wells, pipelines, and surface storage, and production costs rise with too-rapid extraction. This practice leads to the premature venting of natural gas or other fluids that help drive the oil to the surface. Total oil recovery is reduced.

In early US production, private common-property remedies at the oil field level were hindered by the number of parties on each reservoir, differences in size (usually number and scope of production leases), variation in reservoir characteristics (quantity of hydrocarbons below lease lines, mixture of oil or natural gas, and geological conductivity), and asymmetric information about lease values. The number of competitive producers on Oklahoma and Texas fields in the 1920s and 1930s

suggests the bargaining problem. On the Yates, Hendrick, and Seminole fields, there were, respectively, 16, 18, and 40 different lease holders and on the huge East Texas field over 1000 (Libecap & Wiggins, 1984, pp. 89–94).

A possible solution was early consolidation of production rights through the buy-out of all leases and thereby eliminate competitive withdrawal. The early or “gusher” stage of field development with its fury of production and waste, however, was the time of least information about the value of individual holdings, limiting the possibilities for exchange (Wiggins & Libecap, 1985).

Another option was to prorate group-regulated production across all leases on a field according to an agreed-upon formula, such as surface acreage (Libecap & Wiggins, 1984). Under prorationing, a total monthly or annual field output was to be set and divided among the producers. Where producers were relatively homogeneous in size, lease holdings, or in subsurface deposits, these common-property prorationing agreements could be reached. Such conditions described smaller oil fields, but on larger, more heterogeneous reservoirs with mixed distributions of oil, gas, and geological conductivity, the parties could not reach agreement. The most valuable lease holders often in the center of the deposit were reluctant to agree with prorationing rules that did not recognize their preferential deposits, while more marginal lease holders, who relied more upon draining neighbors, resisted rules that differentially limited their drainage ability.

The most complete common-property remedy to open access, absent outright purchase, was unitization that placed the management of the reservoir under a single firm, while granting other producers shares in the net revenues (Wiggins & Libecap, 1985; Libecap & Smith, 2003). Two issues complicated negotiations. One was selection of the unit operator and the other was the share formula to divide the net proceeds of unit production among the various parties. Agreements often were not forthcoming until late in a reservoir’s productive life, when enough common information had emerged about the nature and value of both the reservoir and individual leases.

If unit contracts were to succeed, they had to award each party a *fixed* share of production and costs, making them residual claimants to reservoir-wide rents. Once unitized production ensued, the distribution of hydrocarbons and production would change permanently. If such uniform shares could not be agreed to, then either no unit was formed or less-effective partial unit contracts were concluded. The latter required that producers in the partial unit drill wells along boundaries with other, nonunit producers to limit oil migration and drainage by injecting natural gas to block internal oil flows. Hundreds of such costly fencing wells might be required, depending on field size and number of partial units.

In their analysis of unitization negotiations on seven fields in the USA, Wiggins and Libecap (1985, pp. 377–383) found that on average it took 6 years for unit contracts to be completed. During the process, many parties became discouraged and dropped out of unitization efforts. The bargaining problem was so widespread that even as late as 1975 only 38% of Oklahoma and 20% of Texas production came from unitized fields (Libecap & Wiggins, 1985, p. 702).

Similarly, the parties on the immense Prudhoe Bay field of Alaska could not reach common agreement on the value of their respective leases to assign cost and revenue shares to a complete unit. As a result, the reservoir was partitioned in 1977 into two potentially competitive units, “oil rim,” where most parties held oil and “gas cap,” where most parties held natural gas. A unit operator was selected for each partition, and there were separate allocations of production costs and benefits among the parties in each partition or subunit, even though they covered the *same* reservoir.

Conflicting motivations for oil and gas production controls were inherent in the arrangement, resulting in serious waste, causing the field to prematurely decline in productivity. Low-cost oil production required injection, not sale, of natural gas. Removal of natural gas left some oil deposits stranded in subsurface formations. No remedy was reached until 1999, *22 years after the Prudhoe Bay discovery*, when company mergers and consolidation of holdings finally completely unitized the reservoir (Libecap & Smith, 2003).

In face of these bargaining problems within common property, oil and gas producers in Texas, Oklahoma, Louisiana, and Alaska lobbied state legislatures and regulatory agencies for constraints on competitive production. Some took their competitors to court to reduce open-access losses. Because of their differences, however, firms formed competing lobby groups that affected how politicians and regulatory agencies responded. In Texas, state-enforced control on production and prorationing of output within each field was the most common political action. Regulations limited the drilling of wells, controlled their spacing, and constrained output according to assigned production or prorationing quotas.

In lobbying the state for regulation, the position of larger producers conflicted with the objectives of smaller lease holders. The latter were more numerous and politically influential and were able to secure preferential regulations. Small, high-cost producers (often called “independents”) were in almost every Texas county, and they had close ties to local politicians. Further, they were serviced by local oil-field supply firms and they hired local labor. These firms were comparatively homogeneous and effectively organized for political action as the Texas Independent Producers Organization (TIPRO). Larger producers (the so-called, majors) often were headquartered out of state and hence viewed as “foreign.” Moreover, they were located on the largest, most productive fields, which were in fewer parts of the state. They often had internal oil-field supply support and brought in their own labor. For these reasons, despite their wealth and size, the major oil firms were less politically effective in designing oil production regulation in Texas than were the independents.

These political differences molded state regulatory policies. The Texas Railroad Commission set monthly statewide production levels and allocated the total among regulated wells. The production rules were applied uniformly to all fields, even though each oil field had a unique physical configuration and optimum production potential. This approach raised production costs relative to what might have occurred with alternative regulatory designs. Small firms dominated on older, less productive fields that larger firms had exited. Further, the numerous and politically influential owners of high-cost wells on all fields were exempted from production controls altogether (Libecap & Smith, 2003).

Similarly, in terms of unitization, large firms attempted to gain political acceptance in state legislatures of forced unitization rules. The policy mechanism was to define a majority of producers on each field who had to agree to unitize the entire reservoir and then force the holdouts into the unit. Compulsory unitization laws were adopted in most states, but in Texas, political resistance by small firm owners blocked enactment. Small firm owners lobbied to protect the production advantages they held under existing prorationing regulation (Libecap & Wiggins, 1985). Opposition also occurred where oil and natural gas holdings differed among lease holders, and proposed unit agreements did not fully compensate those with more natural gas. Especially for smaller producers, the formula for natural gas value conversion to oil value was critical because they could not offset losses across leases as effectively as could larger firms (Libecap & Smith, 2001).

15.5 State Regulation with Heterogeneous Lobby: US Fisheries

Migratory ocean and inshore fisheries historically have been characterized by open access and a race to fish. Since the colonial period when stocks were extremely plentiful and the source of low-cost food, wild fish have been held as legal public property. They were common property by law and exposed to easy entry by fishers along vast ocean, river, and lake shores. US inshore species were once so abundant that early residents harvested great quantities of crabs, oysters, clams, sturgeon, herring, and shad. More remote fishing fleets harvested cod off New England and sardines in Monterey Bay, California. The Pacific sardine fishery was at one time the largest fishery in North America, peaking in the 1930s and 1940s and described by John Steinbeck in his 1945 novel, *Cannery Row*. By the 1950s, the fishery was in collapse due to natural population fluctuation and intense overharvest. Similarly, Atlantic cod harvests that peaked in the late 1960s fell sharply by the 1980s with fishing periodically halted. Not all US fish stocks suffered as dramatically, but patterns were similar for many fisheries where there was easy entry and the absence of property rights (Scott, 1955).

With falling harvests, incomes, and fish stocks, fishers and fishery biologists mobilized politically to lobby their state and federal government representatives and regulatory agencies for relief. Local US sports and commercial fishing industries and communities were important political constituencies at least through the 1990s, and politicians responded to excessive harvest by denying access to those with the least political influence—noncitizens. In the case of ocean fisheries, the adoption of 200-mile exclusive economic zones, beginning in 1976 by the US and other countries, meant that foreign fishers could be restricted, reducing harvest pressure.⁵ The USA adopted the Fishery Conservation and Management Act of 1976 (Public Law

⁵The Law of the Sea Convention of 1982 authorized 200-mile exclusive economic zones.

94–265) to define US fishery access and regulation to address open access.⁶ Fisheries outside the 200-mile limit were subsequently placed under various international agreements where compliance and outcomes were variable. Within the 200-mile exclusive economic zone, domestic fishers, however, were not so constrained and increased their fishing pressure in response.

In the USA, a frenzy of investment in new boats and equipment, often subsidized by federal and state governments, soon replaced the excluded foreign fishers (Christy, 1997). With harvests and stocks falling within the 200-mile zones, political pressures mounted once again. Industry differences based on vessel size and home port, primary target species, and techniques and gear type—purse seine, trawl, long-line, sport, and commercial fisheries—fragmented political efforts. Had the divisions only been among heterogeneous fishers as described by Karpoff (1987), buyouts or compensating regulation might have been feasible, as occurred among oil and gas producers. Unlike oil and gas, however, fishing attracted the attention of marine biologists, other academics, and environmental NGOs. They formed an influential, potentially competing constituency, at least with some segments of the fishing industry, with a focus on biological stock preservation, rather than on the economic profitability of user groups.⁷ For these reasons, along with industry divisions, the direct impact of fishery interests was diluted, affecting the ultimate design of US fishery regulations.

As stocks, harvests, and incomes fell in the 1970s and 1980s, the initial regulatory response was to differentially control access or harvests of domestic fishers with more limited political influence. For example, in the Gulf of Mexico, shrimp fishery analyzed by Johnson and Libecap (1982) inshore bay shrimpers and offshore Gulf shrimpers competed to place differential constraints on one another.⁸ Absent property rights, the less-controlled segment, however, expanded with initial rising harvests. As catch and incomes-per-unit-of-effort ultimately declined, not only did lobbying of state and federal fishery agencies increase, but internally, officials called for reduced pressure on biological stocks.

Additional limited-access controls, such as briefer fishing seasons, as well as gear and vessel-size restrictions were implemented. Short but uniform fishing periods were politically popular because they appeared to be equitable and were comparatively easy to define and monitor. Better fishers still outcompeted their less-skilled or smaller competitors so that as seasons were further reduced, the smaller, less-capitalized vessel segment was particularly affected. In addition to seasons, there were equipment restrictions that disproportionately harmed fishers who had adopted new techniques, added capital, and were more productive. While such regulations

⁶See *A Legislative History of the Fishery Conservation and Management Act of 1976*. Government Printing Office, October 1976.

⁷There does not seem to be recent analysis in the literature of voter characteristics for fishing communities or environmental NGO members to assess relative political influence. The link between income and voting is described in Tollison and Willett (1973).

⁸General fishery regulation problems are described in Johnson and Libecap (1982), Karpoff (1987), and Wilen (1988).

were popular with small vessel owners, they reduced the overall efficiency of the industry. Grainger and Parker (2013) describe the distributional conflicts that affected the regulatory response and made the industry less efficient.

Because the regulatory structure did not make fishers residual claimants of the rental value of the stock, fishers continued to compete on unregulated dimensions, dissipating the economic value of the fishery. Common-property regimes that might have constrained some competition faced nonmember vessels entering from beyond community limits as well as differential adjustments among members to rising prices and technologies. These heterogeneities reduced the effectiveness of any local controls as well as government regulations. As costs rose and harvest incomes fell, some smaller, less-efficient fishers left the industry (Johnson & Libecap, 1982; Grainger & Parker, 2013). Those who remained were larger with more capital-intensive vessels and crews. They were more homogeneous in expected long-term fishery prospects and anticipated benefits from effective fishery management, such as assignment of property rights. Property rights could mold incentives for investment in and exploitation of the fishery over time, improving both profitability and stocks.

Small-scale fishers, regulatory agency personnel, and academics, however, were not always supportive of the highly capitalized, more industrial fishing firms. Their efficiency potentially posed greater harm to fish stocks absent property rights, alarming fish biologists. Moreover, that industry segment which often was vertically integrated had less need for government services in information, measurement, or enforcement. Finally, larger, more capital-intensive fleets often were headquartered at larger ports and not older, more traditional communities. Even so, small, more traditional fishers, who sought protection via regulation, however, generally were not as efficient as their larger competitors and less able to generate wealth and value in the industry.

For these reasons, there was no cohesive lobby as there was with mineral rights, for widespread adoption of some type of property rights. The fragmented lobby included actual fishers, regulatory officials, and academics. The latter favored policies they believed would provide more regulatory control over the biological stocks and small-scale practices in the fishery. Larger, more capital-intensive fishers sought opportunities for consolidation and investment in new distribution and marketing channels.

The Pacific Northwest halibut fishery's record illustrates the efficiency problems arising from regulation of the heterogeneous industry when there were no property rights. The relevant data are provided in Table 15.1 from Grafton, Squires, and Fox (2000) for the British Columbia fishery. The stocks are migratory and the fleets similar, only partitioned by national ocean boundaries between the USA and Canada.

Following general fishery regulations in 1980, entry was limited in British Columbia to those who were granted fishing licenses. The maximum number of vessel licenses was set at 435. Additional gear restrictions and minimum fish size rules also were adopted. Even so, among the licensed vessel owners, a fishing derby ensued as fishers competed by adding vessels, crews, and times spent fishing. The number of vessels rose quickly from 333 in 1980 to the limit of 435 by 1988, when total harvest peaked at 12,859,562 pounds, up 128% from 1980. The halibut stock

Table 15.1 Season length, number of fishing vessels, and total catch in the British Columbia Halibut Fishery

Year	Season length (days)	Number of active vessels	Total catch (pounds)
1980	65	333	5,650,447
1981	58	337	5,654,856
1982	61	301	5,524,783
1983	24	305	5,416,757
1984	22	334	8,276,152
1985	22	363	9,587,902
1986	15	417	10,240,471
1987	16	424	12,251,086
1988	14	435	12,859,562
1989	11	435	10,738,715
1990	6	435	8,569,367
1991	214	433	7,189,273
1992	240	431	7,630,198
1993	245	351	10,560,141
1994	245	313	9,900,958
1995	245	294	9,499,717
1996	245	281	9,499,717

Source: Grafton et al. (2000)

declined, forcing regulators to reduce the allowable season to protect the stock. By 1990, the season had shrunk to 6 days from 65 days in 1980. Karpoff (1987, p. 191) notes that in the US portion of the fishery, the season was 4 days in 1984.

Homans and Wilen (1997) modeled a regulated fishery with entry controls and an aim to protect the stock and then empirically compared its predictions to Gordon's (1954) open-access model using data from the Pacific Northwest halibut fishery. They find that regulation actually increased the amount of capital attracted to the fishery because of a higher regulated biomass and no effective constraints on vessel capital. The only regulatory option was to shorten the fishing season, raising capital redundancy and cost.

Moreover, the value of the harvest declined. With a very short fishing season, the catch had to be stored frozen for the rest of the year, denying consumers higher-valued fresh fish. This unsatisfactory situation led to a new regulatory approach, the introduction of use rights, Individual Transferable Quotas (ITQs) in British Columbia in 1991 and in Alaska in 1995. Through quota exchanges, gradually the number of vessels declined, the stock rebounded, and the season was extended, reaching 245 days by 1993.

The ITQs implemented in modified form in the Pacific Halibut fishery were described almost 20 years earlier by Christy (1973). They were an incentive-based, property rights regulatory alternative in fisheries (Buck, 1995). Total allowable catch (TAC) was to be set with shares or individual transferable quotas (ITQs) granted fishers as a property right to the flow of harvest. With a property rights system, there were potentially self-enforceable controls on entry and compliance.

Further there were individual incentives to protect and invest in the stock. The value of fishery ITQs depended upon long-term stock viability. Additionally, if ITQs could be traded easily, the number of fishing vessels could be reduced voluntarily, leaving the most productive fishers in the industry. Trade would increase overall output-per-unit-of-effort. The remaining, more efficient fishers could better access capital markets for additional investment, new product development and delivery.

Hannesson (2006) describes a general pattern of moving from uncontrolled entry to prescriptive government controls to the adoption of property rights with their efficiency advantages. There is documented success of ITQ catch shares as property rights in improving stocks and fishery incomes (Arnason, 2007, 2008; Costello et al., 2008). Despite all of this, the political process and lobbying by differential constituencies across fishing countries, however, led to varying implementation of property rights regimes.

ITQs were adopted in New Zealand in 1986 and in 1991 in Iceland, two nations that depended critically on their fisheries and conservation of the stock. Political opposition among constituencies in the USA, however, resulted in them being placed on hold with more abbreviated rollout that included controls on the distribution and design of ITQs. In 1996, the USA placed a six-year moratorium on the introduction of additional ITQs. ITQs typically were distributed via grandfathering (Anderson et al., 2011), and small vessel owners, often from more remote communities, were concerned that they would be gradually eliminated from the industry through exchange.

To address these distributional concerns, constraints were placed on trading to limit vessel consolidation in some fisheries and to retain a small-vessel, community-based fishery (Singh & Weninger, 2015). Some academics were critical of granting of ownership windfalls to a common property, public resource (Bromley, 2009). Unlike New Zealand and Iceland, ITQs were deemed expressly not a property right in the USA but a privilege and potentially revocable without compensation (Creed & McCay, 1996). These constraints weakened the property right granted and its efficiency benefits. Grainger and Costello (2014) compare dividend price ratios (lease price/sales price) and find that the ratios were significantly higher in the USA with riskier sales and more reliance upon leases, than in New Zealand where clearer property rights existed. It is informative how long a more effective property rights response to open access took through the political process in light of constituent heterogeneity. Slow adoption exceeds even that found in US oil and gas exploitation, extending costs and welfare losses.

15.6 State Regulation of the Common Pool: Concluding Remarks

Theory and research regarding collective action to regulate common-pool problems suggest that it comes when (a) there is broad consensus on the aggregate benefits to be gained, (b) the parties perceive positive net gains from agreement, and (c) they are homogeneous with respect to bargaining objectives and in the distribution of the

costs and benefits to be incurred. Agreements reached under these conditions tend to be self-enforcing because it is in the interest of all parties to ensure success. Collective action may also achieve its objectives if the parties are heterogeneous with respect to the net gains from cooperation if (a) the spread in net gains among the parties is not too great, (b) there is little uncertainty as to the consequences of agreement, and (c) side payments and exchange are feasible to compensate those parties that may bear more costs or receive fewer benefits.

In the cases examined in this chapter, the political processes of securing state responses to the common pool have been complex, influenced by heterogeneity of the lobby parties. Heterogeneity has raised the costs of collective action and molded the ultimate results. Mineral rights agreements involved the most uniform parties and were the most complete. Oil and gas producers who encountered open-access losses were differentially affected so that their lobby efforts were disjointed, delaying and molding the regulations ultimately adopted. In fishing, the parties have varied even more radically. As a result, although the mitigation of open-access losses is understood and documented, reforms toward more secure property rights have been limited.

The problem of devising and implementing clear incentive-based user systems that internalize externalities and equate marginal social costs and benefits in use, conservation, and exchange is exacerbated when the resource is defined normatively in some cases as a common property. As described by Ostrom (1990) and Schlager and Ostrom (1992), there clearly are settings where homogeneous groups manage resources effectively as common property to avoid the losses of the common pool. In settings where Ostrom's conditions for successful collective action are not met (Cox et al., 2010), broader common-property designation likely raises the political transaction costs of designating private property rights as a more effective solution.

This condition also describes Western US federal lands, often referred to as "public lands." These are vast resources with numerous, heterogeneous users that contain valuable timber, grazing, minerals, oil and gas deposits, recreational opportunities, and critical ecosystem/species preservation locations. These uses often conflict, but there is no easy common-property mechanism to resolve competition, avoid open access in some cases, and raise overall resource values. Instead, advocate lobbying, however, molds political allocation and management. As a consequence, well-defined, secure, tradable user rights are less common than are contentious, potentially short-term regulatory controls (Libecap, 2018).

Across the three cases examined here, the regulatory institutions have been long-standing with important efficiency and wealth-generating implications. Some have incorporated clear property rights institutions with welfare improvements, whereas others have relied upon less-effective arrangements. Understanding why this is so and assessing the economic, welfare, and resource impacts requires attention to bargaining among the affected parties, the collective action costs involved, and the roles and incentives of politicians and agency officials.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute

References

- Allen, D. W. (2000). Transaction costs. In B. Bouckaert & G. De Geest (Eds.), *Encyclopedia of law and economics* (Vol. 1). Edward Elgar.
- Alston, L. J., Libecap, G. D., & Schneider, R. (2000). Land reform policies, the sources of violent conflict, and implications for deforestation in the Brazilian Amazon. *Journal of Environmental Economics and Management*, 39, 162–188.
- Anderson, T. L., Arnason, R., & Libecap, G. D. (2011). Efficiency advantages of grandfathering in rights-based fisheries management. *Annual Review of Resource Economics*, 3, 159–179.
- Arnason, R. (2007). Fisheries self-management under ITQs. *Marine Resource Economics*, 22(4), 335–451.
- Arnason, R. (2008). Iceland's ITQ system creates new wealth. *Electronic Journal of Sustainable Development*, 1(2), 35–41.
- Bromley, D. (2009). Abdicating responsibility: The deceptions of fisheries policy. *Fisheries*, 34(6), 280–302.
- Buck, E. H. (1995). *Individual transferable quotas in fishery management*. Environment and Natural Resources Policy Division. Congressional Research Service, September 25, 1995 95-849 ENR.
- Cheung, S. N. S. (1970). The structure of a contract and the theory of a non-exclusive resource. *Journal of Law and Economics*, 13(1), 49–70.
- Christy, F. T. Jr. (1973). *Fisherman quotas: A tentative suggestion for domestic management* (Law of the Sea Institute, Univ. Rhode Island, Occ. Pap. No. 19, June).
- Christy, F. T., Jr. (1997). Economic waste in fisheries: Impediments to change and conditions for improvement. In E. K. Pikitch, D. D. Huppert, & M. P. Sissenwine (Eds.), *Global trends: Fisheries management*. American Fisheries Society.
- Costello, C., Gaines, S. D., & Lynham, J. (2008). Can catch shares prevent fisheries collapse? *Science*, 321(5896), 1678–1681.
- Cox, M., Arnold, G., & Tomás, S. V. (2010). A review of design principles for community-based natural resource management. *Ecology and Society*, 15(4), 38–52.
- Creed, C. F., & McCay, B. J. (1996). Property rights, conservation, and institutional authority: Implications of Magnuson Act reauthorization for the mid-Atlantic region. *Tulane Environmental Law Journal*, 9(2), 245–256.
- Dasgupta, P., & Heal, G. (1974). The optimal depletion of exhaustible resources. *The Review of Economic Studies*, 41, 3–28.
- David, A. P., & Wright, G. (1997). Increasing returns and the genesis of American resource abundance. *Industrial and Corporate Change*, 6(2), 203–245.
- Demsetz, H. (1967). Towards a theory of property rights. *American Economic Review*, 57(2), 347–359.
- Gordon, H. S. (1954). The economic theory of a common property resource: The fishery. *Journal of Political Economy*, 62(2), 124–142.
- Grafton, R. Q., Squires, D., & Fox, K. J. (2000). Private property and economic efficiency: A study of a common-pool resource. *Journal of Law and Economics*, 43, 679–713.
- Grainger, C. A., & Costello, C. J. (2014). Capitalizing property rights insecurity in natural resource assets. *Journal of Environmental Economics and Management*, 67(2), 224–240.
- Grainger, C. A., & Parker, D. P. (2013). The political economy of fishery reform. *Annual Review of Natural Resource Economics*, 5, 369–386.
- Hannesson, R. (1986). The effect of the discount rate on the optimal exploitation of renewable resources. *Marine Resource Economics*, 3(4), 319–329.
- Hannesson, R. (2006). *The privatization of the oceans*. MIT Press.
- Hardin, G. (1968). The tragedy of the commons. *Science*, 162, 1243–1248.
- Homans, F. R., & Wilen, J. E. (1997). A model of regulated open access resource use. *Journal of Environmental Economics and Management*, 32(1), 1–21.
- Johnson, R. N., & Libecap, G. D. (1982). Contracting problems and regulation: The case of the fishery. *American Economic Review*, 72(5), 1005–1022.

- Karpoﬀ, J. M. (1987). Suboptimal controls in common resource management: The case of the fishery. *Journal of Political Economy*, 95, 179–194.
- Knight, F. H. (1924). Some fallacies in the interpretation of social costs. *Quarterly Journal of Economics*, 38, 582–606.
- Libecap, G. D. (1978). Economic variables and the development of the law: The case of western mineral rights. *Journal of Economic History*, 38(2), 338–362.
- Libecap, G. D. (2018). *Federal Lands, opportunity costs, and bureaucratic management* (NBER Working Paper 24705). <http://www.nber.org/papers/w24705>
- Libecap, G. D., & Smith, J. L. (2001). Regulatory remedies to the common pool: The limits to oil field unitization. *Energy Journal*, 22(1), 1–26.
- Libecap, G. D., & Smith, J. L. (2003). The economic evolution of petroleum property rights in the United States. *The Journal of Legal Studies*, 31(2, pt. 2), S589–S608.
- Libecap, G. D., & Wiggins, S. N. (1984). Contractual responses to the common pool: Prorating of crude oil production. *American Economic Review*, 74(1), 87–98.
- Libecap, G. D., & Wiggins, S. N. (1985). The influence of private contractual failure on regulation: The case of oil field unitization. *Journal of Political Economy*, 93, 690–714.
- Melville, H. (1922). *Moby dick or the white whale*. Dodd, Mead and Company.
- Olson, M. (1965). *The logic of collective action*. Harvard University Press.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
- Peltzman, S. (1976). Toward a more general theory of regulation. *Journal of Law and Economics*, 19(2), 211–240.
- Schlager, E., & Ostrom, E. (1992). Property rights regimes and natural resources: A conceptual analysis. *Land Economics*, 68(3), 249–262.
- Scott, A. D. (1955). The fishery: The objectives of sole ownership. *Journal of Political Economy*, 63, 116–124.
- Singh, R., & Weninger, Q. (2015). Cap-and-trade bycatch management with costly avoidance and stock uncertainty. *Marine Resource Economics*, 30(1), 97–119.
- Stigler, G. J. (1971). The theory of economic regulation. *Bell Journal of Economics and Management Science*, 2(Spring), 3–21.
- Stigler, G. J. (1975). *The citizen and the state: Essays on regulation*. University Chicago Press.
- Tollison, R. D., & Willett, T. D. (1973). Some simple economics of voting and not voting. *Public Choice*, 16(Fall), 59–71.
- Umbeck, J. (1977). A theory of contract choice and the California gold rush. *Journal of Law and Economics*, 20(2), 421–437.
- Wiggins, S. N., & Libecap, G. D. (1985). The influence of private contractual failure on regulation: The case of oil field regulation. *Journal of Political Economy* 93(4), 690–714.
- Wilen, J. E. (1988). Limited entry licensing: A retrospective assessment. *Marine Resource Economics*, 5(4), 313–324.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Legal Institutions and the Development of Financial Systems Revisited

16

Thorsten Beck and Ross Levine

Contents

16.1	Introduction	360
16.2	Legal Theories of Financial Development	362
16.2.1	Law, Enforcement, and Financial Development	362
16.2.2	The Historical Development of Europe's Legal Systems	362
16.2.3	The Spread of Europe's Legal Systems	366
16.2.4	From Legal Origin to Finance: Political and Adaptability Mechanisms	368
16.2.5	Skeptical Views Regarding the Law and Finance Theory	370
16.2.6	Alternative Views Regarding the Development of Financial Systems	372
16.3	Empirical Evidence on Law and Finance	374
16.3.1	Legal Origin and Financial Development	374
16.3.2	Investor Protection Laws, Corporate Finance, and Financial Development	377
16.3.3	Law and Finance Theory's Political and Adaptability Mechanisms	380
16.4	Conclusions	382
	References	382

Abstract

This chapter reviews the finance and law literature, which focuses on the role of legal institutions in shaping financial development. Considerable research finds that (1) in legal systems that enforce private property rights, support private contractual arrangements, and protect investors' legal rights, savers are more willing to finance firms, and financial markets more efficiently allocate capital, and (2) the different legal traditions that emerged in Europe over previous centuries and were spread internationally through conquest, colonization, and

T. Beck (✉)
European University Institute, Fiesole, Italy
e-mail: thorsten.beck@eui.eu

R. Levine
Stanford University, Stanford, CA, USA
e-mail: rosslevine@stanford.edu

imitation help explain cross-economy differences in investor protection, the contracting environment, and financial development. We discuss alternative explanations of financial development and weigh the evidence.

Keywords

Financial development · Legal institutions · Corporate finance · Legal origin · Legal adaptability · Political structure

16.1 Introduction

A burgeoning literature finds that financial development exerts a first-order impact on long-run economic growth. King and Levine (1993a, b) and Levine and Zervos (1998) show that financial development is a good predictor of economic growth. Beck, Levine, and Loayza (2000) and Levine, Loayza, and Beck (2000) show that this relationship is not due to omitted variable bias or reverse causation. At the microeconomic level, Demirgüç-Kunt and Maksimovic (1998) and Rajan and Zingales (1998) find that financial institutions are crucial for firm and industrial expansion. While disagreements remain, the bulk of existing evidence points to a strong finance-growth nexus, especially for middle-income countries (Levine, 1997, 2004, 2024; Popov, 2018).

The finding that financial development influences economic growth raises critical questions, such as why some countries have growth-enhancing financial systems, while others do not, and why some countries have developed the necessary investor protection laws and contract enforcement mechanisms to support financial institutions and markets, while others have not.

The law and finance theory focuses on the role of legal institutions in explaining international differences in financial development (e.g., the pioneering work of La Porta, Lopez-de-Silanes, Shleifer, and Vishny (1997, 1998, 2000a), henceforth LLSV, and the review in Levine (2005)). The first part of the law and finance theory holds that in countries where legal systems enforce private property rights, support private contractual arrangements, and protect investors' legal rights, savers are more willing to finance firms, and financial markets flourish. In contrast, legal institutions that neither support private property rights nor facilitate private contracting inhibit corporate finance and stunt financial development.

The second part of the law and finance theory emphasizes that the different legal traditions that emerged in Europe over previous centuries and were spread internationally through conquest, colonization, and imitation help explain cross-country differences in investor protection, the contracting environment, and financial development today. Legal theories emphasize two interrelated mechanisms through which legal origin influences finance (Hayek, 1960). The “political” mechanism holds that (a) legal traditions differ in the priority they attach to private property rights vis-à-vis the rights of the State and (b) the protection of private contracting rights forms the basis of financial development (LLSV, 1999). The “adaptability” mechanism

stresses that (a) legal traditions differ in their formalism and ability to adapt to changing conditions and (b) legal traditions that adapt efficiently to minimize the gap between the economy's contracting needs and the legal system's capabilities will more effectively foster financial development than more rigid systems (Merryman, 1985).

Countervailing theories and evidence challenge both parts of the law and finance theory. Many researchers accept that effective investor protection facilitates efficient corporate financing and growth-enhancing financial development but reject the law and finance's view that legal origin is a central determinant of investor protection laws and financial development (Roe, 1994; Pagano & Volpin, 2001; Rajan & Zingales, 2003). Furthermore, while some scholars accept the importance of legal tradition in shaping the efficiency of financial contracting, there are sharp disagreements about which legal systems work best to promote the efficient evolution of the law (Rubin, 1982). Alternatively, some studies directly question the importance of investor protection laws by arguing that changes in investor protection laws did not drive the evolution of corporate ownership and financial development in the United Kingdom and Italy (Franks et al., 2008; Aganin & Volpin, 2003).

Given debates about the role of legal institutions in shaping financial development, the remainder of the chapter is organized as follows. Section 16.2 describes the law and finance theory along with skeptical and competing views.¹ Section 16.3 reviews empirical evidence on both parts of the law and finance view. That is, we assess (i) whether legal origins account for cross-country variations in property rights protection, support of private contractual arrangements, investor protection laws, and financial development and (ii) the degree to which cross-country differences in investor protection laws explain differences in corporate finance and financial development. Besides examining supportive and conflicting evidence on these two parts of the law and finance theory, we also summarize recent findings on the mechanisms—the political and adaptability mechanisms—through which law and finance may be related.

This chapter updates and expands our 2005 examination of the connection between legal institutions and financial development given continuing advances in this field. We also note that La Porta et al. (2008, 2013) offer extensive reviews of the law and finance literature and Beck (2012) evaluates the connections between legal institutions and economic development more generally.

¹To qualify our approach, however, we recognize that many participants in the law and finance debate may disagree that the law and finance view necessarily comprises the two parts mentioned above. This is not crucial for our review. We simply note that many contributors to the debate on the links between legal institutions and financial development examine (i) the impact of legal origin on property rights protection, support for private contractual arrangements, and investor protection laws (ii) the impact of investor protection laws and their enforcement on financial development or (iii) both. This review examines these different components.

16.2 Legal Theories of Financial Development

This section describes the law and finance theory. We devote a considerable space to tracing the historical evolution of legal institutions because the law and finance theory stresses that historically determined differences in legal heritage continue to shape private property rights protection, investor protection laws, and financial development today. Furthermore, this section describes two mechanisms through which legal origin may influence the contracting environment: the political and adaptability mechanisms. Finally, we review countervailing views that question the law and finance theory.

16.2.1 Law, Enforcement, and Financial Development

The first part of the law and finance theory stresses that legal institutions influence how firms raise funds, the governance of firms, and the overall functioning of financial markets and institutions (LLSV, 1998). This emphasis on the impact of legal institutions on finance follows naturally from the evolution of corporate finance theory during the past half-century. Modigliani and Miller (1958) view debt and equity—the two central mechanisms for raising funds—as legal claims on firms' cash flows. Jensen et al. (1976) stress that statutory laws and the degree to which courts enforce those laws shape the types of contracts that are used to address governance problems within firms, including the problem that inside managers and controlling shareholders are often in positions to expropriate minority shareholders and creditors (Shleifer & Vishny, 1997). Furthermore, as summarized by Hart (1995), financial economists have increasingly focused on how legal institutions shape (i) the control rights that financial securities bring to their owners and (ii) the impact of different legal rules on corporate control.

Thus, the law and finance theory emphasizes that cross-country differences in (i) contract, company, bankruptcy, and securities laws, (ii) the legal systems' emphasis on private property rights, and (iii) the efficiency of enforcement influence (a) how firms raise funds and the costs of those funds, (b) how investors govern firms and the effectiveness of those mechanisms, and (c) the confidence and extent to which people purchase securities and participate in financial markets and institutions.

16.2.2 The Historical Development of Europe's Legal Systems

The second part of the law and finance theory stresses that a country's legal heritage shapes its approach to property rights, private contracting, investor protection, and, hence, financial development. Comparative legal scholars note that the world's major legal families were formed in Europe over centuries and then spread internationally. Thus, we begin our discussion with Roman law.

Hayek (1960) notes that when Emperor Justinian had the Roman law compiled in the sixth century, he attempted to eliminate jurisprudence, i.e., the development of law on a case-by-case basis. Further, while Roman law placed the law above all individuals, the Justinian texts placed the emperor above the law. This “Justinian deviation” did not take root; jurisprudence continued to shape the law.

From the fifteenth century, France’s legal system evolved as a regionally diverse if not fragmented *mélange* of customary law, law based on the Justinian texts, and case law (Dawson, 1968, p. 349). Although courts must have debated the appropriate application of conflicting Roman and customary law as new circumstances emerged, these deliberations generally occurred in private and without the same public, scholarly debates seen in Germany or England (Dawson, 1968, pp. 286–302). Further, by the eighteenth century, there was a notable deterioration in the integrity and prestige of the judiciary. The Crown sold judgeships to wealthy families, and the judges unabashedly promoted the interests of the elite and impeded progressive reforms.^{2,3}

Unsurprisingly, the French Revolution turned its fury on the judiciary and quickly strove to (a) place the State above the courts and (b) eliminate jurisprudence.⁴ Codification under Napoleon supported the unification and strengthening of the State and relegated judges to a minor, bureaucratic role. According to the theory underlying the French civil code, the legislature drafts laws without gaps, so judges do not make law by interpreting the existing laws, by choosing between laws, or by giving meaning to ambiguous laws. Like Justinian, Napoleon sought a code that was so clear, complete, and coherent that there would be no need for judges to deliberate publicly about which laws, customs, and past experiences apply to new, evolving situations.⁵ Furthermore, this approach required a high degree of procedural formalism to reduce judges’ discretion in regulating the presentation of evidence, witnesses, arguments, and appeals (Schlesinger et al., 1988). Thus, France adopted both greater procedural formalism and more limited judicial discretion.

There are conflicting views on the success of the Napoleonic code’s goal of eliminating jurisprudence. Merryman (1985, 1996) argues that the Napoleonic doctrine was a temporary, largely theoretical “deviation” from 2000 years of a legal tradition built on jurisprudence. In contrast to theory, the French courts eventually built an entire body of tort law based on Article 1382 of the Napoleon code, which states that one whose act injures another must compensate that person.

²See, Dawson (1968, p. 373). Also, while the Crown at times issued progressive reforms, the courts “...refused to apply the new laws, interpreted them contrary to their intent, or hindered the attempts of officials to administer them” (Merryman, 1985, p. 16).

³Glaeser and Shleifer (2002) stress that differences in the political experiences of France and England before the fifteenth century also shaped the evolution of their comparative legal institutions (Zywicki, 2003).

⁴Robespierre even argued that “the word jurisprudence...must be effaced from our language” (Quoted from Dawson, 1968, p. 426).

⁵When the first commentary on the code was published in 1805, Napoleon is said to have exclaimed, “My Code is lost!” (Quoted from Dawson, 1968, p. 387).

In contrast to theory, French courts have used case law to recast the law of unjust enrichment, alter the law on obligations, re-work the law of contracts regarding gifts, and change the system of administrative law (Dawson, 1968, pp. 400–415). From this perspective, practicalities in conjunction with a legal tradition grounded in jurisprudence produced in France a legal system that has increasingly employed judicial discretion over the last two centuries and thereby circumvented inefficient qualities of the code.

Others disagree and argue that antagonism toward jurisprudence and the exaltation of the state's role produced a comparatively static, rigid legal tradition.⁶ The French situation encouraged the development of easily verifiable “bright-line-rules” that do not rely on the discretion of judges (Glaeser & Shleifer, 2002). While simple and clear, Johnson et al. (2000) argue that bright-line-rules and excessive judicial formalism may not allow judges sufficient discretion to apply laws fairly to changing conditions and therefore not support evolving commercial needs.

Turning to Germany, Bismarck—like Napoleon—unified the country (in 1871) and prioritized unifying the courts through codification. Bismarck's decision in 1873 to codify and unify the whole of private law in Germany led to the adoption of the German civil law in 1900.

However, unlike in France, German courts have published (since at least the sixteenth century) comprehensive deliberations that illustrated how courts weighted conflicting statutes, resolved ambiguities, and addressed changing situations (Dawson, 1968). German university law faculties worked directly with courts and tried to reconcile emerging situations with the logic of the Justinian texts. Through active debate between scholars and practitioners, Germany developed a dynamic, common fund of legal principles that formed the basis for codification in the nineteenth century. Unlike in France, legal scholars thus took a prominent role in developing the German system.

Moreover, the German code “was not intended to abolish prior law and substitute a new legal system; on the contrary, the idea was to codify those principles of German law that would emerge from careful historical study of the German legal system” (Merryman, 1985, p. 31).⁷

Whereas the Napoleonic code was designed to be immutable, the *Bürgerliches Gesetzbuch* was designed to evolve. For instance, France technically denies judicial review of legislative actions, while Germany formally recognizes this power and German courts actively exercise it (Glendon et al., 1982, p. 57). While France's administrative courts are within the executive branch itself, in Germany, the judiciary handles these disputes. Thus, while codification had a similar role in Germany and France in unifying the country and reasserting the power of the central state, Germany had a very different approach toward jurisprudence.

⁶see Posner, 1973; Rubin, 1977; Priest, 1977

⁷The German legal scholar Karl von Savigny argued that the law of a people was a product of the history and culture of that people's development (Merryman, 1985, p. 30).

The Scandinavian civil law developed relatively independently from the other traditions in the seventeenth and eighteenth centuries and is less closely linked with Roman civil law than the French or German traditions (Zweigert & Kötz, 1998, henceforth ZK). Moreover, neither the construction nor the subsequent evolution of the Scandinavian civil law has been used to eliminate jurisprudence and boost the role of the State relative to private investors to the same extent as in the French civil law (LLSV, 1998).⁸ Comparatively less effort has been devoted to understanding the functioning of the Scandinavian civil law tradition and its influence on the development of financial systems in Scandinavia.

The British common law's historical development is unique in terms of (a) the relationship between the State and the courts and (b) jurisprudence. From 1066, English law evolved based on the resolution of specific disputes and increasingly stressed the rights of private property. While landholding rights in England were initially based on William I's feudal system, the courts developed legal rules that treated large estate holders as private property owners and not as tenants of the king. Indeed, the common law at the dawn of the seventeenth century was principally a law of private property (e.g., Littleton, 1481; Coke, 1628).

The English common law asserted its independence from the State during the great conflict between Parliament and the English kings in the tumultuous sixteenth and seventeenth centuries. The Crown attempted to reassert feudal prerogatives and sell monopoly rights to cope with budgetary shortfalls. Parliament (composed mostly of landowners and wealthy merchants) and the courts took the side of the property owners against the Crown. While King James I argued that royal prerogative superseded the common law, the courts asserted that the law is king, *Lex, Rex*. This political struggle culminated in 1688 when the Stuarts were thrown out. This allowed the courts to place the law above the Crown and limit the Crown's power to alter property rights and grant monopoly rights.⁹

Besides the power of the law vis-a-vis the State, the common law's history is also notably different from France's in terms of jurisprudence and legal formalism. Unlike in Pre-Revolutionary France, the courts in England were frequently viewed more favorably and sometimes as supporters of progressive reforms, so judges were afforded greater discretion. Regarding legal formalism, English law typically imposes less rigid and formalistic requirements on the presentation of evidence, witnesses, etc. and instead offers judges greater latitude (Schlesinger et al., 1988). Regarding jurisprudence, the English common law tradition is almost synonymous

⁸Coffee (2001) points to the superior performance of the Scandinavian countries relative to other civil law countries and even to common law countries and explains this with the high level of social cohesion in these countries.

⁹There are two additional related issues. First, England was unified during the formative period of the common law. This reduced political incentives for codification. Second, English courts were a liberalizing force that helped dismantle the feudal system and protected the rights of landowners against the Crown (Hayek, 1960). Whereas the French Revolution sought individual rights through strict prohibitions on the discretion of judges, England found liberty through an independent and influential judiciary.

with judges having broad interpretation powers and courts molding and creating law as circumstances change. The common law is obsessed with facts and deciding concrete cases rather than adhering to the logical principles of codified law. Thus, the famous dictum: “The life of the law has not been logic: it has been experience” (ZK, 1998, p. 181). Unlike the Napoleonic doctrine, judges continually—and as a matter of general practice—shape the law through their decisions.

16.2.3 The Spread of Europe’s Legal Systems

The English, French, and German legal traditions spread throughout the world through conquest, colonization, and imitation. Napoleon secured the adoption of the code in all conquered territories, including Italy, Poland, the Low Countries, and the Habsburg Empire. Also, France extended her legal influence to parts of the near east, northern, and sub-Saharan Africa, Indochina, Oceania, French Guyana, and the French Caribbean islands during the colonial era. Furthermore, the French code heavily influenced the Portuguese and Spanish legal systems, which helped spread the French legal tradition to Central and South America. The English common law spread through colonization and conquest to all corners of the world. The Austrian and Swiss civil codes were developed at the same time as the German civil code, and the three influenced each other heavily. In turn, Czechoslovakia, Hungary, Yugoslavia, and Greece relied on German civil law in formulating and modernizing their legal systems in the early part of the twentieth century. The German civil code was not imposed but greatly influenced Japan. At the end of the nineteenth century, Japan looked toward Europe as it sought to draft a commercial code. While Japan considered the French civil code, Japanese legal scholars were attracted to the systematic theorizing of the German code and its emphasis on fitting the evolution of the law into a country’s historical context (ZK, 1998, pp. 296–302). The Japanese commercial code of 1899 is squarely based on the German counterpart. Although Japan came under the influence of the common law during the post-World War II occupation period (especially in the area of public law), it is not uncommon to classify Japan as a German civil law country, particularly when focusing on commercial and company law. Similarly, the German code influenced the development of commercial law in Korea, especially through the Japanese occupation. During the early decades of the twentieth century, China (and hence Taiwan) examined European law in seeking to improve the operation of their commercial law. China introduced civil codes in 1925 and 1935 that, except for family and inheritance law, were shaped by German civil law. Of course, China has its own ancient legal tradition and also experienced Mao and the Cultural Revolution. The Scandinavian legal system was not spread to any country outside Northern Europe.

While the subject of active debate, Merryman (1996) advances four interrelated reasons for why the exportation of the Napoleonic code had more pernicious effects in French, Belgian, Dutch, Spanish, and Portuguese colonies than in France itself. According to this view, the adoption of the French civil code has crippled the judicial

systems of many French legal origin colonies and hindered their ability to develop efficiently adaptive legal systems.

First, the French rigidly imposed the Civil Code in its colonies even though there were—and remain—serious conflicts between the code and local laws (ZK, 1998, pp. 109–113). Tensions between local law and the transferred doctrine may impede the efficient development and application of the law with negative implications for financial development (Berkowitz et al., 2002). In comparison, it is valuable to note that England did not try to replace Islamic, Hindu, or unwritten African law, and the flexibility of the common law eased its transfer. For instance, the English courts in India were instructed to apply Islamic or Hindu law depending on the faith of the parties in cases of inheritance, marriage, caste, etc. In Africa, judges were to apply the English law only to the extent that local circumstances permitted and matters were to be decided by equity and good conscience as rendered necessary by local circumstances (ZK, 1998, pp. 225–229). While somewhat chaotic, this arguably set the stage for the evolution of an independent, dynamic common law in the post-colonial era.

Second, when the French instilled the code, they brought the theory of the Napoleonic doctrine with its antagonism toward jurisprudence and its reliance on judicial formalism to minimize the role of judges. The French did not also bring the practical knowledge of how to circumvent some of the negative attributes of the code and create an efficient role for judges (Merryman, 1996).

Third, given the Napoleonic doctrine, judges in French legal origin countries—though not necessarily France itself—frequently “. . . are at the bottom of the scale of prestige among the legal professions in France and in many nations that adopted the French Revolutionary reforms, and the best people in those nations accordingly seek other legal careers” (Merryman, 1996, p. 116). Consequently, it is more difficult to develop efficiently responsive legal systems if the courts do not attract the best minds. Also, the static theory of the Napoleonic doctrine may become self-fulfilling: the best minds choose other professions, which hinders efficient legal flexibility. As a consequence, the legislature will tend to write “bright line laws” to limit the role of the courts. As argued by Pistor et al. (2002), once a country adopts the “bright line” approach to law-making, it is very difficult to change. Courts will not be challenged to develop legal procedures and methods to deal with emerging conditions. Thus, according to some scholars, these characteristics of French law have worked to retard the development of efficiently adaptive legal systems that support financial development.

Fourth, France has a long history of avoiding open disputes about legal interpretation (Dawson, 1968). Moreover, Napoleonic doctrine formally inhibits open disputations by judges on how they weigh competing statutes, ambiguous laws, and past court decisions in deciding new cases. The exportation of this characteristic to French legal origin colonies, i.e., the absence of a legal culture of openly discussing the application of the law to evolving conditions, hindered the development of efficient legal systems around the world accordingly. From this perspective, French legal origin colonies imported a restrictive, formalistic legal doctrine under particular conditions that enhanced the probability that their legal systems would be less efficiently adaptable than common and German civil law countries and even than the legal system in France itself.

16.2.4 From Legal Origin to Finance: Political and Adaptability Mechanisms

We now describe two mechanisms through which legal origin may influence financial development. The political mechanism is based on two premises. First, legal traditions differ in the emphasis they place on protecting private investors' rights relative to the State's rights. Second, private property rights protection forms the foundation for financial development. Thus, historically determined differences in legal origin can help explain the existing differences in financial development according to this component of the law and finance view (LLSV, 1998).

Some scholars argue that the civil law has tended to support the rights of the State, relative to private property rights, to a greater degree than the common law with adverse implications for financial development. Indeed, La Porta et al. (2004) find that in civil law countries, the State is less likely to grant judges tenure, give court jurisdiction over cases involving the government, or permit judicial review of the constitutionality of laws. LLSV (1999, pp. 231–232) state that a civil legal tradition, then, can be taken as a proxy for the intent to build institutions to further the power of the State. A powerful State with a responsive civil law at its disposal will tend to divert the flow of society's resources toward favored ends, which is antithetical to competitive financial markets. Furthermore, a powerful State will have difficulty credibly committing not to use its coercive powers to interfere in financial markets (Greif, 2005), which will also hinder financial development. Thus, the law and finance theory holds that civil law countries will have weaker property rights protection and lower levels of financial development than countries with other legal traditions.

In contrast, the common law has historically tended to side with private property owners against the State according to this view. Rather than becoming a tool of the State, the common law has acted as a powerful counterbalance that promotes private property rights. Rajan and Zingales (2003) note that governments in civil law countries were more effective than governments in common law countries in expanding the role of government at the cost of financial market development during the interwar period between 1919 and 1939. They attribute this to the stronger role of the judiciary vis-a-vis the legislature in common law countries. Thus, the law and finance theory holds that the British common law supports financial development to a greater degree than the civil law systems.

The second mechanism linking legal origin with financial development is the adaptability mechanism, which is built on two premises. First, legal systems differ in their ability to adjust to changing circumstances. Second, if a country's legal system adapts only slowly to changing circumstances, large gaps will open between an economy's financial needs and the legal system's ability to support those needs.

An influential, though by no means unanimous, line of inquiry holds that legal systems that embrace case law and judicial discretion tend to adapt more efficiently to changing conditions than legal systems that adhere rigidly to formalistic procedures and that rely more strictly on judgments based narrowly on statutory law (Coase, 1960). Posner (1973) argues that while legislators consider the impact on

particular individuals and interest groups when writing statutes, judges are forbidden from considering the deservedness of specific litigants and, therefore, more likely to render decisions based on objective efficiency criteria (Rubin, 1982, p. 205). Rubin (1977) and Priest (1977) hold that common law systems are more efficient than statutory-based systems because inefficient laws are routinely litigated and re-litigated pushing the law toward more efficient outcome. In contrast, Posner (1973) and Bailey and Rubin (1994) argue that statutory law evolves slowly and is subject to a greater degree of inefficient political pressures than the common law.¹⁰ If statutes are constantly playing “catch-up” and are continually pushed in inefficient directions by the legislative process, this will hinder efficient corporate finance and financial development.

Thus, while subject to countervailing views presented below, the adaptability channel predicts that French legal origin countries, albeit not necessarily France (as analyzed by Merryman (1996) and explained above), have a lower probability of developing efficiently flexible financial systems than German civil law and especially common law countries. The adaptability channel holds that the common law is inherently dynamic as it responds case by case to the changing needs of society. This limits the opportunities for large gaps to grow between the demands of society and the law. Indeed, La Porta et al. (2004) show that common law countries are more likely to admit judicial decisions as a source of law. In addition, Djankov et al. (2003a) stress that differences in legal formalism also influence the adaptability of the law. They find that common law countries tend to have less legal formalism in regulating the collection and presentation of evidence, while civil law countries require elaborate and extensive procedures throughout judicial processes, insisting on written documentation at every stage of the process and setting rigid procedural requirements on communication between parties.

In contrast, the Napoleonic doctrine’s distrust of judges induces a reliance on judicial formalism. This hinders the flexibility of the legal system in many French law countries, with adverse implications on financial development. Furthermore, as noted, many legal scholars argue that German law falls close to the common law in terms of adaptability since it rejected the Napoleonic doctrine and maintained its historical roots in jurisprudence.

While the political and adaptability mechanisms are interrelated parts of the law and finance theory and both predict that legal origin shapes financial development, they make conflicting predictions regarding French versus German civil law

¹⁰For example, in the United States, corporate officers and directors have a legal responsibility to maximize a firm value for shareholders. Macey and Miller (1993) argue that the efficiency justification for these broad fiduciary responsibilities is to fill in gaps because it is impossible to pre-contract for all contingencies. This gap-filling role of fiduciary duties can lower transaction costs and improve corporate governance by requiring directors to promote the interests of shareholders above their own interests. For example, in a legal system where judges do not go beyond the statutes, “. . . a corporate insider who finds a way not explicitly forbidden by the statutes to expropriate outside investors can proceed without fear of an adverse judicial ruling” (LLSV, 2000a, p. 9).

countries. The political channel holds that the civil law tradition—both French and German—tends to centralize and intensify state power and, therefore, takes a more wary stance toward the development of free financial systems than the common law. In contrast, the adaptability channel stresses that common law and German civil law countries have notably more adaptable legal traditions than French ones.

The two mechanisms also make different predictions concerning the channels through which legal systems influence the development of financial markets. The political mechanism contends that State control of the judiciary produces a system that focuses more on the power of the State and less on the private contracting rights of individual investors than a legal system characterized by an independent judiciary. Thus, the political channel stresses that cross-country differences in the independence of the judiciary are critical for explaining cross-country differences in financial development. In contrast, the adaptability mechanism stresses that cross-country differences in the law's flexibility are critical for explaining cross-country differences in financial development.¹¹

One can overemphasize the differences between the political and adaptability channels, however. The political channel focuses on the power of the State, while the adaptability channel highlights differences in the ability of legal systems to evolve with changing conditions. Jurisprudence, however, may be much less likely in a system where the State controls the judiciary than in a system where the judiciary enjoys greater independence (Damaska, 1986; Glaeser & Shleifer, 2002).

16.2.5 Skeptical Views Regarding the Law and Finance Theory

Many influential legal scholars and economists question each premise underlying the law and finance theory. There are disagreements about the comparative flexibility of the common and civil law traditions, doubts about the view that the common law places greater emphasis on private property rights protection than the civil law, skepticism about classifying countries by legal origin and thus ignoring other influences including indigenous legal systems (Siems, 2022), questions about whether legal origin is a fundamental determinant of financial development, and doubts about the central role of investor protection laws in promoting financial development.

Specifically, Backhaus (1998) and Blume and Rubinfeld (1982) argue that precedent can hamper the efficient evolution of the law. Indeed, Epstein (1975) and Rubin (1982) provide a rich set of examples, including the evolution of the law of property

¹¹ Proponents of the political channel argue that historically Germany had much more efficient institutions than France did. Citing Ertman (1997) and Finer (1997), LLSV (1998, 1999) note that Germany built a professional bureaucracy based on the military and professional civil servants, while France developed a patrimonial bureaucracy with strong links to political elites. Arguably, these differences have also worked to create German courts that are more independent from the State, more efficient at protecting private contracting rights, and less focused on the rights of the State than in France.

during the nineteenth century and the design of private clauses in contracts, when statutory law changes were necessary to produce more efficient outcomes in the United States. As another example, English law has clung with remarkable tenacity to the principle that “only a person who is a party to a contract can sue on it” (ZK, 1998, p. 468). In contrast, the continental countries granted greater rights to third parties through statutory changes. Furthermore, Lamoreaux and Rosenthal (2002) compare the laws of incorporation and partnerships in the United States and France. They argue that the French civil law system responded more effectively to evolving economic conditions than the US common law system. Finally, Bentham (1789) noted that the common law’s lack of coherence hinders its ability to evolve efficiently.

Another line of criticism questions Posner’s (1973) argument that the courts have better incentives to select socially efficient outcomes than the legislature. Galanter (1974) and Tullock (1980) argue that rich disputants and well-endowed special interest groups can litigate and re-litigate cases, which blur Posner’s (1973) delineation between the processes of legislation and litigation. Furthermore, the choice of litigation and legislation may be primarily a strategic decision regarding which avenue offers the highest probability of success. From this perspective, there is no reason to presume that common or civil law systems will produce more efficient outcomes.

Research also questions whether common law systems emphasize property rights relative to the rights of the State to a greater degree than civil law systems. For instance, Ekelund and Tollison (1980) and Rubin (1982) argue that while the courts in England sided with Parliament against the Crown’s efforts to grant monopolies in the sixteenth and seventeenth centuries, this should not be viewed as a general characteristic that common law legal systems favor private property rights and competition more than civil law systems. Along similar lines, Coffee (2000) argues that civil law systems are not inherently against minority shareholder rights, but instead the law has evolved sufficiently in civil law countries to protect minority shareholders effectively, given the patterns of corporate ownership in those countries.

Furthermore, many question whether categorizing countries as having British, French, German, or Scandinavian legal origins is appropriate and analytically useful. For example, when the French legal system was exported to colonies around the world, it operated less effectively than in France itself (Dawson, 1960, 1968; Merryman, 1985, 1996) as stressed above, and legal scholars highlight differences in the operation of the French civil law across countries in Latin America. Turning to the common law, Franks and Sussman (2005) describe material differences in the adaptability of two common law countries, the United Kingdom and the United States, highlighting potential weaknesses in placing common law countries into a single category. Along the same lines, Berkowitz, Pistor, and Richard (2002) stress that the way national legal systems were initially transplanted and received, e.g., through conquest, colonization, or imitation—not just whether countries are classified as having British, French, German, or Scandinavian legal origins—shapes the

functioning legal systems. Building on this view, Oto-Peralis Romero-Ávila (2014a, b) show that the common law operated very differently after colonization depending on whether there was a high proportion of European settlers in the colony. At the same time, France imposed its civil law more rigidly and uniformly across its empire.

Some researchers question whether legal heritage is a crucial determinant of legal and financial institutions and instead stress that politics determines the degree of investor protection laws, the energy devoted to private contract enforcement, the extent to which legal systems emphasize the rights of property owners relative to the rights of the State, and hence the development of competitive financial markets (Pound, 1991; Roe, 1994; Pagano & Volpin, 2001; Rajan & Zingales, 2003; Haber et al., 2003). From this perspective, those in power shape policies and institutions—including legal and financial institutions—to stay in power and enrich themselves. The elite may or may not favor financial development, which ultimately influences the operation of legal and financial institutions. This view does not reject the importance of legal institutions in shaping financial systems. Instead, it stresses the political roots of differences in legal and financial institutions.¹²

16.2.6 Alternative Views Regarding the Development of Financial Systems

Scholars propose several alternative explanations of the heterogeneous development of financial systems worldwide. Most of these theories relate to the persistence of historical shocks and the distribution of economic and political power.

First, the endowment view stresses that differences in geography and disease environment have critically shaped patterns of political, institutional, and economic development (McNeill, 1963; Jones, 1981; Crosby, 1989; Diamond, 1997; Engerman & Sokoloff, 1997; Sokoloff & Engerman, 2000; Acemoglu et al., 2001, 2002; Haber et al., 2021).

Acemoglu, Johnson, and Robinson (2001, henceforth AJR) base their theory on how endowments influence enduring institutions on three premises. First, AJR note that Europeans adopted different types of colonization strategies. At one end of the spectrum, the Europeans settled and created institutions to support private property and check the power of the State, while at the other end of the spectrum, Europeans sought to extract as much from the colony as possible. Second, the feasibility of European settlement heavily influenced the type of colonization strategy. In inhospitable environments, Europeans tended to create extractive states (AJR, 2001). In areas where endowments favored settlement, Europeans tended to form settler

¹² Glaeser and Shleifer (2002) model the evolution of legal institutions, while Glaeser and Shleifer (2003) show that legal and regulatory institutions may evolve together and sometimes substitute for each other depending on specific conditions. For broad discussions of the co-evolution of legal, regulatory, and political institutions, see Olson (1993), North (1981, 1990), Djankov, Glaeser, LaPorta, Lopez-de-Silanes, and Shleifer (2003b), and Barth, Caprio, and Levine (2004). Easterly and Levine (1997) show that ethnic division may shape the wide range of institutions and policies.

colonies. Third, institutions created during the colonial times were persistent, including after independence, thus influencing today's level of financial development. AJR (2001, 2002), Beck, Demirgüç-Kunt, and Levine (2003a, henceforth BDL), and Easterly and Levine (2003) provide empirical support for the view that endowments influence institutions, including financial institutions.

A related strand of literature stresses that the dominance of natural resources in an economy influences institutional development (Sachs & Warner, 2001). According to this view, extracting natural resources, such as oil, does not require well-developed institutions, efficient markets, and human capital underlying successful manufacturing activities (Beck & Laeven, 2006; Besley & Persson, 2010). Focusing on financial institutions, Beck and Poelhekke (2023) show that countries whose GDP increases due to an unexpected positive shock to the world price of its basket of natural resources experience slower increases in bank deposits and loans than countries experiencing similarly sized economic booms due to non-resource-related shocks.

Second, some argue that religious traditions shape the financial sector. Stulz and Williamson (2003) note that different religions have different attitudes toward creditors' rights. In particular, the Catholic Church has historically taken a hostile stance toward charging interest and creditor rights. Similarly, the Qur'an prohibits charging interest. In contrast, according to this culture-religion view, the Reformation advanced a different religious attitude toward finance, whereby the payment of interest was considered a normal part of commerce and the rights of creditors were more naturally supported in countries dominated by Protestant religions. From this perspective, countries with a predominantly Protestant religious heritage tend to have comparatively well-developed credit markets and loan-issuing financial institutions.

Exploring data across and within six countries in South Eastern Europe that were formerly only partly occupied by the Ottoman Empire, Grosjean (2011) finds that the former Islamic rule (which prohibited interest lending into the nineteenth century) is consistently associated with lower contemporaneous formal financial development, both across and within countries, and localities with Armenian, Jewish, or Greek minorities, who were allowed to practice interest lending under the Ottoman rule, have higher levels of bank penetration.

Religious prejudices, such as antisemitism, can also shape financial development. When Pope Leo IX prohibited Christians from charging interest in the eleventh century and Jews were banned from owning farms and other lands in much of Europe, this pushed Jews into providing finance. These developments reinforced anti-Jewish and anti-finance sentiments and strengthened the public's association between the two. D'Acunto, Prokopczuk, and Weber (2019) test whether enduring antisemitism is associated with enduring anti-finance sentiments. Consistent with this view, they find that present-day households in German counties with higher anti-Jewish sentiments during the Nazi period participate in the stock market less than other households and are less likely to have a home mortgage.

Third, important cultural differences across countries, including in individualism and risk-taking, might shape legal and financial development. Licht, Goldschmidt, and Schwartz (2005) show that the variation in the quality of legal institutions across countries can be partly explained by variations in societal attitudes toward (a) assertiveness and individualism and (b) risk avoidance and collectivism. Guiso et al. (2004) hold that “social capital,” the informal rules that govern social interactions, played a critical role in determining financial development in Italy, while Guiso et al. (2008) show that less trusting individuals are less likely to invest in stocks. Levine et al. (2020) show that cross-regional differences in the intensity of the historic slave trade in Africa had enduring influences on social trust that continue to shape households’ willingness to use financial services.

Finally, specific historic events might turn into a traumatic experience for nations, with long-ranging implications for institutions. Murphy (2005) sees the 1720’s Mississippi bubble, with its subsequent banking crisis and hyperinflation, as critical for the negative French attitude toward the financial sector. Similarly, the hyperinflationary experience in Germany resulted in a hawkish approach toward monetary policy that was deeply entrenched for the following 80 years. Malmendier and Nagel (2010) show that “Depression babies,” individuals growing up during the Depression era in the United States, are less likely to invest in equity and have overall more risk-averse investment strategies.

16.3 Empirical Evidence on Law and Finance

This section reviews the empirical evidence on the law and finance view. The first subsection discusses evidence on the links between legal origin and financial development, investor protection laws, and private property rights protection. Next, we assess whether investor protection laws influence corporate valuations, corporate governance, and the operation of financial markets. The third subsection reviews emerging evidence on the political and adaptability mechanisms.

16.3.1 Legal Origin and Financial Development

To measure legal origin, many researchers follow LLSV (1998) in classifying a country as having either a British common law, French civil law, German civil law, or Scandinavian civil law based on the source of each country’s company or commercial code. David and Brierley (1985) argue that commercial legal systems of most countries derive from these four major legal families. Reynolds and Flores (1989) provide information on the origins of national laws for over 100 countries. Using these legal origin dummy variables, researchers have initiated an energetic examination of the relationship between legal and financial institutions.

LLSV (1997, 1998) find that French civil law countries have the lowest levels of financial development even after controlling for the overall level of economic development. French civil law countries have smaller stock markets (as measured by market capitalization divided by GDP), less active initial public offering markets, and lower levels of bank credit as a share of GDP. These results are broadly consistent with the theories of law and finance discussed above.¹³

Empirical work also examines the connection between legal origin and specific legal institutions governing the rights of external investors in firms. For example, LLSV's (1998) shareholder rights index measures the degree to which majority shareholders have less discretion in exploiting minority shareholders. LLSV (1998) show that French civil law countries have lower levels of shareholder rights. LLSV (1997) and Levine (2003b) find that low levels of shareholder rights are associated with poorly developed equity markets. In contrast, common law countries have high levels of shareholder rights with correspondingly high levels of equity market development. Following criticism of the original indicator (e.g., Pagano & Volpin, 2005; Spamann, 2010), Djankov et al. (2008b) develop a new measure of the legal protection of minority shareholders against expropriation by corporate insiders, e.g., large shareholders and executives. Djankov et al. (2008b) find that common law countries have significantly stronger laws protecting minority shareholders, and laws are positively associated with stock market development.

Furthermore, LLSV (2006) examine laws and regulations that (1) force information disclosure, (2) facilitate private enforcement of financial contracts through strict liability rules, and (3) give a large role to a public enforcer of securities laws, such as the US Securities and Exchange Commission. They show that French legal origin countries tend to have relatively weak information disclosure and liability rules. Moreover, they show that countries with stronger information disclosure and liability rules tend to have more developed stock markets, while there is not a strong link between public enforcement rules and regulations and the operation of stock markets.

Similarly, the extent to which the legal system protects the rights of creditors positively influences credit market development. LLSV (1997) develop *Creditor Rights*, an index measuring creditors' rights when firms become insolvent, and show that this index is higher in countries with a common law tradition than in French civil law countries. Furthermore, LLSV (1997), Levine (1998, 1999), and Djankov et al. (2007) show that greater creditor rights are positively associated with financial intermediary development. Djankov et al. (2008a) develop and analyze an alternative measure of creditor protection. They collect data across 88 countries on debt enforcement procedures, using the case study of an insolvent firm. After showing that common law countries have comparatively more efficient debt enforcement procedures, they show that their debt enforcement efficiency measure is positively associated with debt market development.

¹³ Additional work further shows that common law countries have significantly greater market capitalization than the combined group of civil law countries (BDL, 2001).

Nevertheless, legal origin is certainly not the whole story. Rajan and Zingales (2003) argue that financial development does not always evolve monotonically and that cross-country differences in financial development also change materially over time. Thus, time-invariant factors such as legal origin cannot fully explain time variation in the relative levels of financial development across countries. Rajan and Zingales (2003) stress the critical role of political forces in shaping policies toward financial markets and intermediaries and, hence, the development of financial systems. Pistor et al. (2002) disagree with Rajan and Zingales (2003) about corporate law and argue that even acute political changes in Germany, France, and England during the twentieth century did not substantively alter the evolution of corporate law.

While recognizing the limitations of the law and finance theory's ability to explain intertemporal changes in relative levels of financial development across countries, recent research has conducted several robustness checks regarding the linkages between legal origin and financial development. Levine (1998, 1999, 2003a), Levine, Loayza, and Beck (2000), and BDL (2003a) use different measures of financial development and also expand the set of countries to over 100. This research confirms that legal origin helps explain cross-country differences in financial development. In particular, French civil law countries, though not France itself, tend to have particularly low levels of equity market development. To the extent that competitive securities markets rely more on legal institutions than banks, these results are consistent with theories suggesting a strong link between legal institutions and financial development.

Furthermore, BDL (2003a) show that French civil law countries tend to have lower levels of private property rights protection. Again, this is consistent with the view that French legal origin countries place comparatively less emphasis on the rights of private property holders than countries with a common or German civil law tradition.

Research has also run statistical horse races between theories that stress the role of legal institutions and alternative theories. As noted earlier, an influential body of works stresses the dominating role of political forces in shaping financial development. While it is challenging to measure cross-country differences in political institutions, BDL (2003a) make an initial attempt to control for differences in political systems in assessing the law and finance relationship. They include measures of the degree of competitive and executive elections, the number of influential veto players in the legislative process, and an overall index of national openness based on trade openness. BDL (2003a) continue to find that legal origin explains differences in equity market development, banking sector development, and the level of private property rights protection even when controlling for these proxies for characteristics of the political environment.

BDL (2003a) also control for natural resource endowments and religion in examining the robustness of the connection between legal heritage and financial development. They find that endowments help explain cross-country differences in financial institutions, confirming the AJR and Engerman and Sokoloff (1997) theories of institutional development. Nevertheless, legal origin continues to explain

property rights differences and stock market development even when controlling for endowments.

Cross-country analyses of the relationship between legal origin and financial sector development have been complemented by within-country analysis. Specifically, France, Spain, and Mexico, all civil law countries, colonized territories within lands that eventually became the United States. Within these territories, they implemented their civil law traditions until those lands were acquired by the British Crown or the US government, which established the common law. Berkowitz and Clay (2005, 2011) show that states initially settled by civil law countries had less independent judiciaries and lower-quality courts a couple of centuries later than states settled by England.

As this discussion suggests, a primary challenge to assessing the legal origins' view of comparative financial development is identifying exogenous changes in legal systems. To address this challenge, Levine et al. (2023) explore differences between British and French concessions over adjacent and similar plots of land in Shanghai following the First Opium War in the nineteenth century. They compare three legal regimes: the initial period without the British and French legal systems operating, when the respective legal codes were applied, and when jurisdiction reverted to Chinese courts. By examining the changing application of different legal traditions to adjacent neighborhoods within the same city and controlling for military, economic, and political characteristics, they evaluate the impact of legal systems on financial development. Consistent with the law and finance view, Levine et al. find that bank density grew comparatively faster in the British concession after the establishment of European courts, and this financial development gap narrowed appreciably when the legal jurisdiction over the concessions reverted to the Chinese courts.

16.3.2 Investor Protection Laws, Corporate Finance, and Financial Development

We now examine the empirical evidence concerning the relationship between investor protection laws and (a) corporate financing decisions and (b) financial market operations and thus the second part of the law and finance hypothesis.

First, recent work suggests that legal institutions influence the valuation of firms and banks and, hence, the cost of capital. Claessens et al. (2002), LLSV (2002), and Caprio et al. (2003) find that stronger shareholder rights tend to enhance stock market valuations of corporations. Furthermore, LLSV (2000b) show that countries with strong shareholder rights **can** force firms to disgorge cash and pay higher dividends. This evidence is consistent with the view that investor protection laws influence corporate governance with measurable implications on stock prices and dividend policies. Atanasov et al. (2010) shows how legal reforms in Bulgaria that strengthened minority shareholder rights increased firms' value. Berkowitz, Lin, and Ma (2015) analyze China's 2007 Property Law, which strengthened property rights for private firms and their creditors (primarily banks), and show that the enactment of

the law had a positive effect on firms' value, especially those with a relatively large share of tangible assets (to be used as collateral) and no political connections.

Second, empirical analyses also find a strong connection between investor protection laws and both ownership concentration and the private benefits of corporate control. The data are consistent with the view that stronger legal protection of investor rights makes minority investors more confident about their investments, which reduces the need for firms (Claessens et al., 2000; LLSV, 1999) and banks (Caprio et al., 2003) to use concentrated ownership as a mechanism for alleviating corporate governance problems. For example, Nenova (2023) shows that the legal environment, law enforcement, investor protection, takeover regulations, and power-concentrating corporate charter provisions explain more than two-thirds of the cross-country variation in the value of control-block votes. Furthermore, Zingales (1994) and Dyck and Zingales (2004) show that greater statutory protection of minority shareholder rights and more effective legal enforcement of those rights lower the private benefits of controlling a corporation.

Third, legal institutions influence the ability of firms to raise capital. Thus, laws may affect the degree to which firms operate at financially constrained levels. Kumar, Rajan, and Zingales (2001) and Beck, Demircug-Kunt, and Maksimovic (2006) find that countries with legal institutions that more effectively protect property rights tend to have larger firms, while Laeven and Woodruff (2007) find that firms in Mexican states with stronger legal institutions are larger, with the effect of being stronger in industries dominated by proprietorships, suggesting that the legal system reduces idiosyncratic firm risk. This is consistent with the law and finance theory that firms are less constrained by retained earnings and operate at more efficient scales in countries with better legal institutions.

Fourth, recent work has also drawn a connection between legal institutions and the efficiency of equity markets. More effective contractual institutions are associated with larger and more liquid stock markets and local bond markets (Djankov et al., 2007; Burger & Warnock, 2006). Musacchio (2008) studies the case of Brazil and shows that the weakening of creditor rights after 1945 contributed to a decline in bond market development. Morck, Yeung, and Yu (2000) find that in countries where legal institutions do not protect shareholders effectively, fewer individuals and institutions participate in domestic stock markets, reducing the efficiency of those markets. Eleswarapu and Venkataraman (2006) show that trading costs are significantly higher for stocks from French civil law than from common law countries and effective spreads and the price impact of trades are significantly lower for stocks from countries with higher judicial efficiency.

Fifth, a growing body of work explores how collateral and bankruptcy laws—critical features of the legal environment—shape the operation of credit markets and structure of credit contracts. For example, Haselmann et al. (2010) gauge the effect of legal reforms in 12 transition economies. They find that banks increase the supply of credit following such reforms but react more to collateral law reforms that increase the likelihood that creditors realize their claims against debtors than bankruptcy reforms that ensure a more orderly resolution process after debtors become insolvent. Assunção et al. (2013) exploit a reform in Brazil that simplified the sale of

repossessed cars used as collateral and find that strengthening the usefulness of this form of collateral expanded credit to riskier, self-employed borrowers and led to larger loans with lower spreads and longer maturities. However, this reform also led to a higher incidence of delinquency and default. Love, Martinez Peria, and Singh (2016) show that collateral registries for movable assets facilitate firm financing.¹⁴ Davydenko and Franks (2008) and Liberti and Mian (2010) show that better creditor protection leads lenders to ease collateral requirements, reducing the impact of wealth on access to credit. Bae and Goyal (2009) show banks reduce loan maturities in countries with less efficient contract enforcement, while Giannetti (2003) documents that firms with intangible assets have easier access to loans in countries with legal systems than more effectively protect creditors.

Qian and Strahan (2007) explore how creditor rights shape the structure of syndicated loans and the terms of those loans. In countries with creditor protection laws that more effectively allow creditors to seize collateral when borrowers default on loans, loans have longer maturities and lower interest rates. Furthermore, stronger creditor rights increase concentration within loan syndicates, which is consistent with stronger collateral laws lowering the benefits of diversifying loan ownership within syndicates. On the other hand, Esty and Megginson (2003) show that syndicates for project finance in countries with weak creditor rights and poor legal enforcement are larger and more diffuse. Fan et al. (2012) find a bias toward more short-term debt in countries with weaker legal institutions. Johnson, Boone, Breach, and Friedman (2000a) show that weak legal institutions—legal institutions that do not effectively support the claims of outside investors—help account for cross-country differences in stock market declines and exchange rate depreciations during the Asian crisis.

Finally, cross-country studies have explored the relationship between legal institutions and the efficiency of capital allocation. Wurgler (2000) and Beck and Levine (2002) show that countries with legal institutions that define and enforce strong rights for small, outside investors more effectively reallocate the flow of finance toward growing firms and away from declining firms. Demirgüç-Kunt and Maksimovic (1998) show that countries with legal institutions that protect outside investors tend to create better-functioning financial systems that fund faster-growing firms. Beck et al. (2005) find that better legal institutions reduce the impact of firms' financing obstacles on firm growth, while Beck et al. (2008) show that firms in countries with poor institutions use less external finance, especially bank finance and that protection of property rights increases external financing of small firms significantly more than of large firms, mainly due to its effect on bank finance. Claessens and Laeven (2003) show that legal rules regarding investor protection influence the types of firms getting external finance. Specifically, in countries with strong investor protection laws, firms with less collateral have an easier time getting external finance than similar firms in countries with poorly functioning legal institutions. Brown et al. (2013) find that strong shareholder protections and better access to stock market

¹⁴For more on the impact of insolvency regimes on the economy, see Claessens and Klapper (2005), Acharya and Subramanian (2009), and Gine and Love (2010).

financing lead to substantially higher long-run rates of R&D investment, particularly in small firms.

Some careful case studies, however, challenge the importance of investor protection laws. For example, Franks et al. (2008) trace the history of investor protection laws and corporate ownership in the United Kingdom. In a landmark court case, *Foss v. Harbottle* (1843), the judge found that no individual shareholder could sustain an action against the company, thereby rejecting the notion of minority investor protection. Not until 1948 did Parliament begin to enact limited legislation to protect minority shareholders, and Franks et al. (2008) stress that it was not until 1980 that Parliament enacted strong minority shareholder rights statutes. The United Kingdom should have had comparatively inactive equity markets and concentrated ownership in the nineteenth and early twentieth centuries and more dispersed ownership and greater equity market activity after 1948, especially after 1980. The evidence, however, is at best mixed. Ownership concentration was similar in 1900 and 1960, which is not consistent with the law and finance prediction, but market liquidity did jump substantially with the enactment of stronger shareholder rights legislation.

Similarly, Aganin and Volpin (2003) argue that the history of investor protection laws and corporate ownership in Italy during the twentieth century does not support the law and finance view. They hold that investor protection laws were weak at the beginning of the century and did not change much after World War II, but were strengthened after 1974 and especially after 1990. They note that the law and finance theory predicts a fall in corporate ownership concentration after 1974 as stronger investor protection laws make small shareholders more confident about their investments. However, corporate ownership concentration did not fall after 1974; it rose, and ownership concentration was more diffuse at the beginning of the twentieth century than at the start of the twenty-first century. Therefore, Aganin and Volpin (2003) question the applicability of the law and finance view in Italy and stress the importance of considering politics in explaining corporate ownership and the evolution of investor protection laws.

16.3.3 Law and Finance Theory's Political and Adaptability Mechanisms

While an exploding body of research examines (a) the links between legal origin and investor protection and financial development and (b) the links between investor protection laws and corporate financing efficiency, researchers are only beginning to examine the mechanisms through which legal origin operates. The political channel postulates that legal traditions differ in the priority they give to private property rights relative to the state's rights. The adaptability channel stresses that legal traditions differ in their responsiveness to changing socioeconomic conditions.

BDL (2003b, 2005) study whether legal origin influences financial development primarily through the political or adaptability mechanism by exploiting the data assembled by Djankov et al. (2003a) and La Porta, Lopez-de-Silanes, Pop-Eleches, and Shleifer (2004). BDL (2003b) find that French and German civil law countries

have significantly less powerful supreme courts (as measured by lifelong tenure and power over administrative cases), consistent with the view that the State grants less independence in a civil law tradition than in a common law system. They also find that French civil law countries rely significantly less on case law—i.e., they grant a considerably smaller role for judicial decisions as a source of law—than German civil law or British common law. This is consistent with the view that German civil and British common legal traditions rely more on jurisprudence than French civil law systems.

BDL (2003b) also find support for the adaptability channel but not the political channel in explaining cross-country variation in financial sector development: the power of supreme courts enters either insignificantly or negatively, while case law is positively associated with stock market development, bank development, and private property rights protection.

Research also focuses on judicial formalism, which is related to the adaptability mechanism. Excessive formalism may slow legal processes, increase legal costs, and hinder the ability of courts to arrive at fair judgments due to the rigid adherence to bright-line-rules (Johnson et al., 2002; Djankov et al., 2003a; Glaeser & Shleifer, 2002). Indeed, Djankov et al. (2003a) construct an index of legal formalism that measures the need for legal professionals, written documents, statutory justification, the statutory codification of evidence, and the formal procedural steps associated with legal processes. They find that legal formalism is lower in common law countries and less legal formalism is associated with shorter proceedings and less corruption.

Other studies further explore features of the political and adaptability channels. Acemoglu and Johnson (2005) find that legal formalism is negatively associated with stock market development, consistent with the adaptability channel's prediction that greater legal formalism impedes efficient contracting and lowers stock market development. In contrast to BDL (2003b), however, Acemoglu and Johnson (2005) find evidence consistent with the political channel, showing that coercion-constraining institutions, including legal constraints on executive action and strong protection of private property from expropriation by the state, are positively associated with financial development. Focusing on firm-level data from China, Cull and Xu (2005) find evidence that coercion-constraining and contract enforcement institutions shape firms' reinvestment decisions.

Looking back to ancient Rome, Malmendier (2009) also highlights the importance of the political channel. She documents that the Roman form of the shareholding company developed in the early—legally less developed—days of the Roman Republic when the political environment supported it. However, this shareholding ownership form—and the accompanying trading of shares—disappeared during the Roman Empire, when the coercion-constraining climate was not favorable toward such ownership firms, despite the Empire's more advanced legal system relative to the Republic's early period.

Beyond the discussion on the channels through which legal origin affects financial sector development, several papers focus specifically on the efficiency of the judiciary and its importance for corporate finance in India. Chemin (2009) shows

that (1) Indian High Court amendments that complicate judicial procedures and (2) conflicting judicial decisions that arise temporarily due to ambiguities in the Code of Civil Procedure increase expected trial duration, which negatively impacts credit markets. Chemin (2012), in turn, documents that the staggered reform of inefficient reforming court procedures across India led to fewer breaches of contract and facilitated small firms' access to finance. Using the staggered introduction of fast-track debt recovery courts in India as an exogenous shock to legal system efficiency, Visaria (2009) shows that loan delinquency and interest rates fall, and Gopalan, Mukherjee, and Singh (2016) show that the new, more efficient debt recover courts increase the ratio of long-term debt to assets, consistent with the legal institutions' view of financial development.

Brown et al. (2017) exploit externally imposed differences in legal institutions across Native American reservations to assess the law's impact on financial markets. The US Congress assigned state courts to adjudicate contract disputes on some reservations, leaving other reservations in the hands of tribal courts. After showing that state courts resolve contract disputes more efficiently and predictably than tribal courts, Brown et al. (2017) discovered that credit market development is significantly higher in reservations assigned to state courts.

16.4 Conclusions

A rapidly growing body of research examines the role of legal institutions in explaining financial development. While this question has been explored by legal scholars, historians, and economists for several decades, the debate on the importance of legal adaptability and the relationship between the judiciary and the State continues. New technologies (digital finance, big data, and artificial intelligence) yield new financial products and new providers of financial services, challenging the existing legal institutions. Which legal institutions will best adapt to these challenges to foster investment, innovation, and improvements in living standards? How will political systems shape the adaptation of legal institutions and financial development to these technological innovations? These are some of the questions that researchers and policymakers must now address.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., & Johnson, S. (2005). Unbundling institutions: Law versus politics. *Journal of Political Economy*, 113, 949–995.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91, 1369–1401.

- Acemoglu, D., Johnson, S., & Robinson, J. A. (2002). Reversal of fortunes: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics*, 117, 1133–1192.
- Acharya, V., & Subramanian, K. V. (2009). Bankruptcy codes and innovation. *Review of Financial Studies*, 22, 4949–4988.
- Aganin, A., & Volpin, P. (2003). *History of corporate ownership in Italy*. London Business School mimeo.
- Assunção, J., Benmelech, E., Silva, F. (2013). Repossession and the democratization of credit. *Review of Financial Studies*, 27, 2661–2689.
- Atanasov, V., Black, B., Ciccotello, C., & Gyoshev, S. (2010). How does law affect finance? An examination of equity tunneling in Bulgaria. *Journal of Financial Economics*, 96, 155–173.
- Backhaus, J. G. (1998). Efficient statute law. In P. Newman (Ed.), *New Palgrave dictionary of economics and law* (pp. 24–28). Macmillan.
- Bae, K.-H., & Goyal, V. K. (2009). Creditor Rights, Enforcement, and Bank Loans. *Journal of Finance*, 64, 823–60.
- Bailey, M. J., & Rubin, P. H. (1994). A positive theory of legal change. *International Review of Law and Economics*, 14, 467–477.
- Barth, J., Caprio, G., & Levine, R. (2004). Bank supervision and regulation: What works best? *Journal of Financial Intermediaries*, 13, 205–248.
- Beck, T. (2012). The Role of Finance in Economic Development – Benefits, Risks, and Politics. In D. Müller (Ed.), *Oxford Handbook of Capitalism* (pp. 161–203).
- Beck, T., & Levine, R. (2002). Industry growth and capital allocation: Does having a market- or bank-based system matter? *Journal of Financial Economics*, 64, 147–180.
- Beck, T., & Laeven, L. (2006). Institution building and growth in transition economies. *Journal of Economic Growth*, 11, 157–186.
- Beck, T. & Poelhekke, S. (2023). Follow the money: Does the financial sector intermediate natural resource windfalls? *Journal of International Money and Finance*, 130, Article 102769.
- Beck, T., Levine, R., & Loayza, N. (2000). Finance and the sources of growth. *Journal of Financial Economics*, 58, 261–300.
- Beck, T., Demirgüç-Kunt, A., & Levine, R. (2001). Legal theories of financial development. *Oxford Review of Economic Policy*, 17, 483–501.
- Beck, T., Demirgüç-Kunt, A., & Maksimovic, V. (2006). The influence of financial and legal institutions on firm size. *Journal of Banking and Finance*, 30, 2995–3015.
- Beck, T., Demirgüç-Kunt, A., & Levine, R. (2003a). Law, endowments, and finance. *Journal of Financial Economics*, 70, 137–181.
- Beck, T., Demirgüç-Kunt, A., & Levine, R. (2003b). Law and finance. Why does legal origin matter? *Journal of Comparative Economics*, 31, 653–675.
- Beck, T., Demirgüç-Kunt, A., & Levine, R. (2005). Law and firms'. *Access to Finance American Law and Economic Review*, 7, 211–252.
- Beck, T., Demirgü.-Kunt, A., & Maksimovic, (2008). Financing patterns asround the world: are small firms different? *Journal of Financial Economics*, 89, 467–487.
- Bentham, J. (1789). *An introduction to the principles of morals and legislation*. T. Payne & Sons.
- Berkowitz, D., & Clay, K. (2005). American civil law origins: Implications for state constitutions and state courts. *American Law and Economics Review*, 7, 62–84.
- Berkowitz, D., & Clay, K. (2011). *The evolution of a nation: How geography and law shaped the American states*. Princeton University Press.
- Berkowitz, D., Pistor, K., & Richard, J.-F. (2002). Economic development, legality, and the transplant effect. *European Economic Review*, 47, 165–195.
- Berkowitz, D., Lin, C., & Ma, Y. (2015). Do property rights matter? Evidence from a Property Law Enactment. *Journal of Financial Economics*, 116, 583–593.
- Besley, T., & Persson, T. (2010). State Capacity, Conflict, and Development. *Econometrica*, 78, 1–34.

- Blume, L. E., & Rubinfeld, D. L. (1982). The dynamics of the legal process. *Journal of Legal Studies*, 11, 405–419.
- Brown, J. R., Martinsson, G., & Petersen, B. C. (2013). Law, stock markets, and innovation. *Journal of Finance*, 68, 1517–1549.
- Brown, J. R., Cookson, J. A., & Heimer, R. Z. (2017). Law and finance matter: Lessons from externally imposed courts. *Review of Financial Studies*, 30, 1019–1051.
- Burger, J. D., & Warnock, F. E. (2006). Local Currency Bond Markets. *IMF Staff Papers* 53, 1–7.
- Caprio, G., Laeven, L., & Levine, R. (2003). *Governance and bank valuation* (National Bureau of Economic Research Working Paper, 10158).
- Chemin, M. (2009). Do judiciaries matter for development? Evidence from India. *Journal of Comparative Economics*, 37, 230–250.
- Chemin, M. (2012). Does the quality of the judiciary shape economic activity? Evidence from a judicial reform in India. *Journal Law, Economic and Organization*, 28, 460–485.
- Claessens, S., & Laeven, L. (2003). Financial development, property rights, and growth. *Journal of Finance*, 58, 2401–2436.
- Claessens, S., & Klapper, L. (2005). Bankruptcy around the world: explanations of its relative use. *American Law and Economics Review*, 7, 253–283.
- Claessens, S., Djankov, S., & Lang, L. H. P. (2000). The separation of ownership and control in East Asian Corporations. *Journal of Financial Economics*, 58, 81–112.
- Claessens, S., Djankov, S., Fan, J. P. H., & Lang, L. H. P. (2002). Expropriation of minority shareholders in East Asia. *Journal of Finance*, 57, 2741–2771.
- Coase, R. (1960). The problem of social cost. *Journal of Law and Economics*, 3, 1–44.
- Coffee, J. C. (2000). *Privatization and corporate governance: The lessons from securities market failure* (Unpublished working paper 158). Columbia Law School.
- Coffee, J. C. (2001). *Do norms matter? A cross-country examination of the private benefits of control* (Unpublished working paper 183). Columbia Law School.
- Coke, E. (1628) [1979 version]. *The first part of the institutes of the laws of England*. Garland Publishing.
- Crosby, A. W. (1989). *Ecological imperialism: The biological expansion of Europe, 900–1900*. Cambridge University Press.
- Cull, R., & Xu, L. C. (2005). Institutions, ownership, and finance: the determinants of profit reinvestment among Chinese firms. *Journal of Financial Economics*, 77, 117–146.
- D’Acunto, F., Prokopczuk, M., & Weber, M. (2019). Historical antisemitism, ethnic specialization, and financial development. *Review of Economic Studies*, 86, 1170–1206.
- Damaska, M. R. (1986). *The faces of justice and state authority: A comparative approach to the legal process*. Yale University Press.
- David, R., & Brierley, J. E. C. (1985). *Major Legal systems in the world today*. Stevens and Sons.
- Davydenko, S., & Franks, J. (2008). Do Bankruptcy Codes Matter? A Study of Defaults in France, Germany, and the U.K. *Journal of Finance*, 63, 556–608.
- Dawson, J. P. (1960). *A history of lay judges*. Harvard University Press.
- Dawson, J. P. (1968). *The oracles of the law*. University of Michigan Law School. (Reprinted in 1986 by William S. Hein & Co.).
- Demirgüç-Kunt, A., & Maksimovic, V. (1998). Law, finance, and firm growth. *Journal of Finance*, 53, 2107–2137.
- Demirgüç-Kunt, A., & Maksimovic, V. (2002). Funding growth in bank-based and market-based financial systems: Evidence from firm level data. *Journal of Financial Economics*, 65, 337–363.
- Diamond, J. (1997). *Guns, germs, and steel: The fates of human societies*. W.W. Norton.
- Djankov, S., McLiesh, C., & Shleifer, A. (2007). Private Credit in 129 Countries. *Journal of Financial Economics*, 84, 299–329.
- Djankov, S., Hart, O., McLiesh, C., & Shleifer, A. (2008a). Debt enforcement around the world. *Journal of Political Economy*, 116, 1105–1149.
- Djankov, S., LaPorta, R., Lopez-de-Silanes, F., & Shleifer, A. (2008b). The Law and Economics of Self-Dealing. *Journal of Financial Economics*, 88, 430–446.

- Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2003a). Courts. *Quarterly Journal of Economics*, 118, 453–518.
- Djankov, S., Glaeser, E., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2003b). The new comparative economics. *Journal of Comparative Economics*, 31, 595–619.
- Dyck, A., & Zingales, L. (2004). Private benefits of control: An international comparison. *Journal of Finance*, 59, 537–600.
- Easterly, W., & Levine, R. (1997). Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics*, 112, 1203–1250.
- Easterly, W., & Levine, R. (2003). Tropics, germs, and crops. *Journal of Monetary Economics*, 50, 3–39.
- Ekelund, R. B., & Tollison, R. D. (1980). Economic regulation in mercantile England: Heckscher revisited. *Economic Inquiry*, 18, 567–599.
- Eleswarapu, V., & Venkataraman, K. (2006). The impact of legal and political institutions on equity trading costs: A Cross-Country Analysis. *Review of Financial Studies*, 19, 1081–1111.
- Engerman, S. L., & Sokoloff, K. L. (1997). Factor endowments, institutions, and differential paths of growth among new world economies. In S. Haber (Ed.), *How Latin America fell behind* (pp. 260–304). Stanford University Press.
- Epstein, R. A. (1975). Unconscionability: A critical reappraisal. *Journal of Law and Economics*, 18, 293–315.
- Ertman, T. (1997). *Birth of the Leviathan*. Cambridge University Press.
- Esty, B., & Megginson, W. (2003). Creditor rights, enforcement, and debt ownership structure: evidence from the global syndicated loan market. *Journal of Financial and Quantitative Analysis*, 38, 37–59.
- Fan, J. P. H., Titman, S. & Twite, G. (2012). An international comparison of capital structure and debt maturity choices. *Journal of Financial and Quantitative Analysis*, 47, 23–56.
- Finer, S. (1997). *The history of government* (Vols. I–III). Cambridge University Press.
- Franks, J., & Sussman, O. (2005). Financial innovations and corporate bankruptcy. *Journal of Financial Intermediation*, 14(3), 283–317.
- Franks, J., Mayer, C., & Rossi, S. (2008). Ownership: Evolution and regulation. *The Review of Financial Studies*, 22(10), 4009–4056.
- Galanter, M. (1974). Why the 'haves' come out ahead: Speculation on the limits of legal change. *Law and Society Review*, 9, 95–160.
- Gianetti, M. (2003). Do Better Institutions Mitigate Agency Problems? Evidence from Corporate Finance Choices. *Journal of Financial and Quantitative Analysis*, 38, 185–212.
- Gine, X., & Love, I. (2010). Do reorganization costs matter for efficiency? Evidence from a bankruptcy reform in Colombia. *Journal of Law and Economics*, 53, 833–864.
- Glaeser, E., & Shleifer, A. (2002). Legal origins. *Quarterly Journal of Economics*, 117, 1193–1230.
- Glaeser, E., & Shleifer, A. (2003). The rise of the regulatory state. *Journal of Economic Literature*, 41, 401–425.
- Glendon, M. A., Gordon, M. W., & Osakwe, C. (1982). *Comparative legal tradition in a nutshell*. West Publishing Co.
- Gopalan, R., Mukherjee, A., & Singh, M. (2016). Do debt contract enforcement costs affect financing and asset structure? *Review of Financial Studies*, 29, 2774–2813.
- Greif, A. (2005). Commitment, Coercion and Markets: The Nature and Dynamics of Institutions Supporting Exchange. In C. Ménard, & M. M. Shirley (Eds.), *Handbook of New Institutional Economics* (pp. 727–788). Dordrecht: Springer.
- Grosjean, P. (2011). The institutional legacy of the Ottoman Empire: Islamic rule and financial development in South Eastern Europe. *Journal of Comparative Economics*, 39, 1–16.
- Guiso, L., Sapienza, P., & Zingales, L. (2008). Trusting the Stock Market. *Journal of Finance*, 63, 2557–2600.
- Guiso, L., Sapienza, P., & Zingales, L. (2004). The role of social capital in financial development. *American Economic Review*, 94, 525–556.

- Haber, S., Razo, A., & Maurer, N. (2003). *The politics of property rights: Political instability, credible commitments, and economic growth in Mexico*. Cambridge University Press.
- Haber, S., Elis, R., & Horrillo, J. (2021). The ecological origins of economic and political systems. Available at SSRN 3958073.
- Hart, O. (1995). *Firms, contracts, and financial structure*. Oxford University Press.
- Haselmann, R., Pistor, K., & Vig, V. (2010). How law affects lending. *Review of Financial Studies*, 23, 549–580.
- Hayek, F. A. (1960). *The constitution of liberty*. The University of Chicago Press.
- Jensen, M. C., William, H., & Meckling, W. (1976). Theory of the firm: Managerial behavior, agency costs, and ownership structure. *Journal of Financial Economics*, 3, 305–360.
- Johnson, S., Boone, P., Breach, A., & Friedman, E. (2000a). Corporate governance in the Asian financial crisis. *Journal of Financial Economics*, 58, 141–186.
- Johnson, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2000b). Tunneling. *American Economic Review* (Papers and Proceedings), 90, 22–27.
- Johnson, S., McMillan, J., & Woodruff, C. (2002). Courts and relational contracts. *Journal of Law, Economics, and Organization*, 18, 221–277.
- Jones, E. L. (1981). *The European miracle: environments, economies, and geopolitics in the history of Europe and Asia*. Cambridge University Press.
- King, R. G., & Levine, R. (1993a). Finance and growth: Schumpeter might be right. *Quarterly Journal of Economics*, 108, 717–738.
- King, R. G., & Levine, R. (1993b). Finance, entrepreneurship, and growth: Theory and evidence. *Journal of Monetary Economics*, 32, 513–542.
- Kumar, K. B., Rajan, R. G., & Zingales, L. (2001). *What determines firm size?* (University of Chicago unpublished working paper).
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1997). Legal determinants of external finance. *Journal of Finance*, 52, 1131–1150.
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1998). Law and finance. *Journal of Political Economy*, 106, 1113–1155.
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (1999). The quality of government. *Journal of Law, Economics, and Organization*, 15, 222–279.
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (2000a). Investor protection and corporate governance. *Journal of Financial Economics*, 58, 3–27.
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (2000b). Agency problems and dividend policies around the world. *Journal of Finance*, 55, 1–33.
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. W. (2002). Investor protection and corporate valuation. *Journal of Finance*, 57, 1147–1170.
- La Porta, R., Lopez-de-Silanes, F., Pop-Eleches, C., & Shleifer, A. (2004). Judicial checks and balances. *Journal of Political Economy*, 112, 445–470.
- La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2006). What works in securities laws? *Journal of Finance*, 61, 1–33.
- La Porta, R., Lopez-de-Silanes, F. & Shleifer, A. (2008). The economic consequences of legal origins. *Journal of Economic Literature*, 46, 285–332.
- La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2013). Law and Finance After a Decade of Research. In: G. M. Constantinides, M. Harris, & R. Stulz, (Eds.), *Handbook of the Economics of Finance* (Vol. 2) (pp. 425–491).
- Laeven, L., & Woodruff, C. (2007). The quality of the legal system, firm ownership, and firm size. *Review of Economics and Statistics*, 89, 601–614.
- Lamoreaux, N., & Rosenthal, J.-L. (2002). *Organizational choice and economic development: A comparison of France and the United States during the mid-19th century*. University of California at Los Angeles mimeo.
- Levine, R. (1997). Financial development and economic growth: Views and agenda. *Journal of Economic Literature*, 35, 688–726.

- Levine, R. (1998). The legal environment, banks, and long-run economic growth. *Journal of Money, Credits, and Banking*, 30, 596–620.
- Levine, R. (1999). Law, finance, and economic growth. *Journal of Financial Intermediation*, 8, 36–67.
- Levine, R. (2003a). Bank-based or market-based financial systems: Which is better? *Journal of Financial Intermediation*, 11, 398–428.
- Levine, R. (2003b). Napoleon, bourses, and growth: With a focus on Latin America. In A. Omar & C. A. Cadwell (Eds.), *Market augmenting government: Essays in honor of Mancur Olson*. University of Michigan Press.
- Levine, R. (2004). Finance and growth: Theory and evidence. In P. Aghion & S. Durlauf (Eds.), *Handbook of economic growth*. North-Holland Elsevier. forthcoming.
- Levine, R. (2005). Law, endowments, and property rights. *Journal of Economic Perspectives*, 19(3), 61–88.
- Levine, R. (2024). Banks and long-run economic prosperity. In A. N. Berger, P. Molyneux, & J. O. S. Wilson (Eds.), *The Oxford handbook of banking* (4th ed.). Oxford University Press. forthcoming.
- Levine, R., & Zervos, S. (1998). Stock markets banks and economic growth. *American Economic Review*, 88, 537–558.
- Levine, R., Loayza, N., & Beck, T. (2000). Financial intermediation and growth: Causality and causes. *Journal of Monetary Economics*, 46, 31–77.
- Levine, R., Lin, C., & Xie, W. (2020). The African slave trade and modern household finance. *The Economic Journal*, 130, 1817–1841.
- Levine, R., Lin, C., Ma, C., & Xu, Y. (2023). The legal origins of financial development: Evidence from the Shanghai concessions. *Journal of Finance*, 78, 3423–3464.
- Liberti, J., & Mian, A. (2010). Collateral Spread and Financial Development. *Journal of Finance*, 65, 147–177.
- Licht, A. N., Goldschmidt, C., & Schwartz, S. H. (2005). Culture, law and corporate governance. *International Review of Law and Economics*, 25, 229–255.
- Littleton, T. (1481) [1903 version]. *Littleton's tenures in English* (E. Wambaugh, ed.). John Bryne & Co.
- Love, I., Martínez Pería, M. S., & Singh, S. (2016). Collateral registries for movable assets: Does their introduction spur firms' access to bank finance? *Journal of Financial Services Research*, 49, 1–37.
- Macey, J. R., & Miller, G. P. (1993). Corporate stakeholders: A contractual perspective. *University of Toronto Law Journal*, 43, 401–427.
- Mahoney, P. (2001). The common law and economic growth: Hayek might be right. *The Journal of Legal Studies*, 30, 503–525.
- Malmendier, U. (2009). Law and Finance 'at the Origin.' *Journal of Economic Literature*, 47, 1076–1108.
- Malmendier, U., & Nagel, S. (2010). Depression Babies: Do Macroeconomic Experiences Affect Risk-Taking? *Quarterly Journal of Economics*, 126, 373–416.
- McNeill, W. H. (1963). *The rise of the west: A history of the human community*. University of Chicago Press.
- Merryman, J. H. (1985). *The civil law tradition: An introduction to the legal systems of Western Europe and Latin America*. Stanford University Press.
- Merryman, J. H. (1996). The French deviation. *The American Journal of Comparative Law*, 44, 109–119.
- Modigliani, F., & Miller, M. (1958). The cost of capital, corporation finance, and the theory of investment. *American Economic Review*, 48, 261–297.
- Morck, R., Yeung, B., & Yu, W. (2000). The information content of stock markets: Why do emerging markets have synchronous stock price movements? *Journal of Financial Economics*, 58, 215–260.

- Murphy, A. (2005). Corporate Ownership in France: The Importance of History. In: R. Morck (Ed.), *A History of Corporate Governance around the World* (pp. 185–219). Chicago: University of Chicago Press.
- Musacchio, A. (2008). Can civil law countries get good institutions? Lessons from the history of creditor rights and bond markets in Brazil. *Journal of Economic History*, 68, 80–108.
- Nenova, T. (2023). The Value of Corporate Voting Rights and Control: A Cross-Country Analysis. *Journal of Financial Economics*, 68, 325–351.
- North, D. (1981). *Structure and change in economic history*. W.W. Norton and Company.
- North, D. (1990). *Institutions, institutional change, and economic performance*. Cambridge University Press.
- Olson, M. (1993). Dictatorship, democracy, and development. *American Political Science Review*, 87, 567–576.
- Oto-Peralis, D., & Romero-Ávila, D. (2014a). The distribution of legal traditions around the world: A contribution to the legal origins theory. *Journal of Law and Economics*, 57, 561–628.
- Oto-Peralis, D., & Romero-Ávila, D. (2014b). Legal traditions and initial endowments in shaping the path of financial development. *Journal of Money, Credit and Banking*, 46, 43–77.
- Pagano, M., & Volpin, P. (2001). The political economy of finance. *Oxford Review of Economic Policy*, 17, 502–519.
- Pagano, M., & Volpin, P. (2005). The political economy of corporate governance. *CSEF Working Paper*, 29.
- Pistor, K., Keinan, Y., Kleinheisterkamp, J., & West, M. D. (2002). The evolution of corporate law. *University of Pennsylvania Journal of International Economics and Law*, 23, 791–871.
- Popov, A. (2018). Evidence on finance and economic growth. In: T. Beck, & R. Levine (Eds.), *Handbook of Finance and Development*. Edward Elgar Publishing.
- Posner, R. A. (1973). *Economic analysis of the law*. Little-Brown.
- Pound, J. (1991). Proxy voting and the SEC. *Journal of Financial Economics*, 29, 241–285.
- Priest, G. L. (1977). The common law process and the selection of efficient rules. *Journal of Legal Studies*, 6, 65–82.
- Qian, J., & Strahan, P. (2007). How law and institutions shape financial contracts: The case of bank loans. *Journal of Finance*, 62, 2803–2834.
- Rajan, R. G., & Zingales, L. (1998). Financial dependence and growth. *American Economic Review*, 88, 559–586.
- Rajan, R. G., & Zingales, L. (2003). The great reversals: The politics of financial development in the 20th century. *Journal of Financial Economics*, 69, 5–50.
- Reynolds, T. H., & Flores, A. A. (1989). *Foreign law: Current sources of basic legislation in jurisdictions of the world*. Rothman and Co.
- Roe, M. J. (1994). *Strong managers weak owners: The political roots of American corporate finance*. Princeton University Press.
- Rubin, P. H. (1977). Why is the common law efficient? *Journal of Legal Studies*, 6, 51–64.
- Rubin, P. H. (1982). Common law and statute law. *Journal of Legal Studies*, 11, 205–233.
- Sachs, J., & Warner, A. (2001). The curse of natural resources. *European Economic Review*, 45, 827–838.
- Schlesinger, R. B., Baade, H. W., Damaska, M. R., & Herzog, P. E. (1988). *Comparative law. Case-text-materials*. The Foundation Press.
- Shleifer, A., & Vishny, R. (1997). A survey of corporate governance. *Journal of Finance*, 52, 737–783.
- Siems, M. (2022). *Comparative Law*. Cambridge University Press.
- Sokoloff, K., & Engerman, S. L. (2000). Institutions, factor endowments, and paths of development in the new world. *Journal of Economic Perspectives*, 14, 217–232.
- Spamann, H. (2010). The “antidirector rights index” revisited. *The Review of Financial Studies*, 23, 467–486.
- Stulz, R., & Williamson, R. (2003). Culture, openness, and finance. *Journal of Financial Economy*, 70, 313–349.

- Tullock, G. (1980). *Trials on trial*. Columbia University Press.
- Visaria, S. (2009). Legal reform and Loan repayment: the microeconomic impact of debt recovery tribunals in India. *American Economic Journal: Applied Economics*, 1, 59–81.
- Wurgler, J. (2000). Financial markets and the allocation of capital. *Journal of Financial Economy*, 58, 187–214.
- Zingales, L. (1994). The value of the voting right: A study of the Milan stock exchange. *The Review of Financial Studies*, 7, 125–148.
- Zywicki, T. J. (2003). The rise and fall of efficiency in the common law: A supply-side analysis. *Northwestern University Law Review*, 97, 1551–1633.
- Zweigert, K., & Kötz, H. (1998). *An introduction to comparative law*. Oxford University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Licit and Illicit Responses to Regulation: Revisited

17

Lee Benham

Contents

17.1	Introduction	392
17.2	Licit Responses to Regulation	394
17.2.1	Substitution of Other Goods	394
17.2.2	Substitution of Other Attributes of Goods	396
17.2.3	Substitution of Amenities	397
17.2.4	Substitution of Barter and Other Arrangements	398
17.2.5	Substitution of Vertical Integration	399
17.2.6	Substitution of Household Production	400
17.2.7	Substitution of Personalized Exchange	400
17.2.8	Changes in Governance and Contractual Relationships	401
17.2.9	Changes in Organization of the Market	401
17.3	Illicit Responses to Regulation	402
17.3.1	Underground Economic Activity	402
17.3.2	Collusive Activities and Norms Within Firms	403
17.3.3	Discrimination	404
17.3.4	Private Coercion and Extralegal Organizations	404
17.3.5	Corruption	405
17.4	History, Path Dependence, and Interest Groups	405
17.4.1	Broadcasting in Britain	407
17.4.2	US Sugar Production	408
17.4.3	The Old Believers in Russia	409
17.4.4	New Technology	409
17.5	Conclusion	410
	References	411

L. Benham (✉)

Ronald Coase Institute and Washington University, St. Louis, MO, USA

e-mail: lee.k.benham@gmail.com

© The Author(s) 2025

C. Ménard, M. M. Shirley (eds.), *Handbook of New Institutional Economics*,

https://doi.org/10.1007/978-3-031-50810-3_17

391

Abstract

Regulation generates both legal and illegal responses. Standard economic analysis emphasizes the impact of regulation on a small number of variables, typically price and quantity. However, the regulatory process plays out with many actors responding along many margins. The responses to a regulation—and the interest groups involved—are affected by what is being measured, which metrics are used, the transparency of the regulatory process and the regulation's consequences, how technology evolves, and which mechanisms are used for enforcement. Several examples illustrate the diverse ways in which regulatory processes have played out historically.

Keywords

Regulation · Interest groups · Corruption · Compliance · Legal · Illegal · Institutional analysis · Governance · Enforcement · Measurement · Path dependence · Coase

17.1 Introduction

This chapter examines various licit and illicit behaviors associated with responses to regulation. Here regulation is taken as the imposition of rules, price controls, and laws to influence the way markets operate. In the study of regulation, the standard neoclassical economics model has historically emphasized a small number of variables, principally prices and quantities, but the regulatory process plays out across many actors along multiple dimensions.

The standard neoclassical model assumes all transaction costs are zero. As Ronald Coase has established, in such a world, if non-equilibrium money prices are set by regulation, the parties to an exchange could simply contract around the regulation (Coase, 1960). In such a market, participants have vast options to adapt, including vertical integration, use of other media of exchange, barter, long-term contracts, and alteration of product attributes. Even with money prices assigned randomly or set to zero, the system could work efficiently. The parties could price discriminate perfectly, and no mutually advantageous trades would go unexercised (Barzel, 1997). In that world, monopolies would not lead to deadweight losses, and there would be no externalities. As noted by Stigler (1951), a zero-transaction-cost world is a strange world.

Coase views the world of zero-transaction costs as a useful first step in analysis, but he encourages us to limit the time we spend studying that world. He encourages us instead to study the world in which we actually live—the world of positive transaction costs. This is the central building block of new institutional economics.

The standard literature on regulation usually modifies the pure neoclassical model slightly by assuming that positive (sometimes even infinite) transaction costs exist along one or two specified margins, for example, “imperfect capital” or “imperfect information,” while transaction costs remain zero along all other margins. These regulatory models pay lip service to utility maximization and opportunity costs—as opposed to purely pecuniary maximization and money prices—but usually ignore

them in empirical work. The desire for parsimony in the standard model is understandable, but there exist many margins of adaptation, and there are no *a priori* grounds for believing that standard price-and-quantity responses are the most significant.

New institutional economics modifies the standard neoclassical model significantly by taking the view that transaction costs are positive and finite along all margins. This perspective draws on the fundamental tenet of economics: the relevant prices facing an individual are the full opportunity costs associated with the various choices (Benham & Benham, 2000). For a given good, these full opportunity costs include the money price of the good itself plus the transaction costs of obtaining the good.¹

This perspective also emphasizes the possibility of substitution along many margins. Markets typically adjust through a mixture of changes in money prices and changes in other characteristics of the goods traded in that market, including location, waiting time, durability, warranty, freshness, and associated services. Indeed, in many markets, waiting time and quality are highly variable, while money prices are relatively sticky.² Inside households, inside firms, and even inside some markets, money prices may not be used at all.

This has many implications for measuring the full range of responses to regulation. Attributes of the goods and also attributes of associated property rights, contractual forms, organizational structures, the medium of exchange, and informal networks can vary in response to new regulations. These responses can reduce or increase the deadweight losses traditionally associated with formal regulations.

Regulatory intervention is not usually driven by abstract scientific curiosity. Interest groups—both outside and within governments—influence the introduction of a regulation and its implementation. Many elements are in play: what is measured, what metrics are used in that measurement, how transparent is the regulatory process, how transparent are the regulation's consequences, how does the technology evolve, and what are the enforcement mechanisms. Moreover, new regulations will generate both new technologies and new interest groups.³

We need to discover which consequences are big and which are small, to learn how various effects play out over time, and to understand the conditions under which a particular type of response will occur. In some contexts, a single small regulation may cascade into a large regulatory system. In other contexts, a Draconian regulation may have little real impact because of the adaptations made.

¹ The opportunity costs an individual faces are likely to be influenced by the specific characteristics of the individual, the type of exchange, and the institutional setting within which the individual is operating. See Benham and Benham (2001).

² Facing short- and long-term variations in market conditions, many restaurants deliberately alter waiting time and quality of service as well as money price.

³ An overview of the central role of politics in the regulatory process is provided by Pablo Spiller (2010). See also Mancur Olson (1982) and Roger Noll and Bruce Owen (1983). Mary Shirley (2002) has examined interest group issues concerning regulation of public water supplies. Most studies of collective action have emphasized the opposite direction of analysis: interest groups' efforts to create regulations serving their interest, rather than new interest groups generated by or emboldened by new regulations.

Table 17.1 shows a broad range of possible licit and illicit responses to regulation. While far from exhaustive, this list extends the set usually examined and illustrates various approaches taken by new institutional economics. Some examples of these responses follow.

17.2 Licit Responses to Regulation

Let us consider some specific types of licit responses.

17.2.1 Substitution of Other Goods

If a regulation raises the price of a given good, consumers will tend to consume less of that good and to substitute other goods. The higher price will also elicit a supply response. The standard neoclassical model focuses principally on these price and quantity substitution effects, but substitution exists along many other margins.

At a given time and place, the set of goods available in the marketplace is merely a subset of all goods that are potentially available. The set existing in the marketplace is established as the outcome of an equilibrium process that depends on competition, production possibilities, income, and the formal and informal rules of the game. Changes in regulation can affect these factors and hence change the available set of

Table 17.1 Categories of responses to regulation

Licit responses	
Substitute	For
Other goods	Regulated good
Other attributes of good	Regulated attribute
Amenities	Profits in excess of regulated maximums
Barter and other arrangements	Money
Vertical integration	Market exchange
Household production	Market production
Personalized exchange	Impersonal exchange
Alter	
Governance and contractual relationships	
Organization of the market	
Illicit responses	
Development of	
Underground economic activity	
Collusive activities and norms within firms	
Discrimination	
Private coercion and extralegal organizations	
Corruption	

goods. This alters not only the range of substitutes available to consumers but also other features of the economy and governance.

Historically governments have often banned products. There are many illustrations of how the substitution effect plays out.

In 1689 the English banned the import of French wines. King Louis XIV of France supposedly commented, “I have substitutes for their manufactures, but they have no substitute for Champagne.” However, the ban resulted in such high prices for French wines that the English substituted beer consumption for wine consumption. The English government then increased the tax rate on beer. This beer tax became a major source of revenue for the British government, helping to support the growth and maintenance of the British navy and the British empire (Nye, 2007).

In the United States, Prohibition was in effect from 1920 to 1933. During this period the transportation and sale of alcoholic beverages were illegal. There were, however, exceptions. These included “medicinal whiskey” which could be legally sold to individuals who presented a medical prescription for diagnoses such as anxiety, depression, or cancer. The medical prescription itself cost about \$3 (several hours of an average wage then). The patient could then pay another \$3 or \$4 to have this prescription filled at a pharmacy for a pint of whiskey. This could be repeated every 10 days (Gambino, 2013). Even Winston Churchill found this system helpful. In 1932 as he was recuperating from an automobile accident in New York City, Dr. Otto Pickhardt provided him with a universal prescription stating “This is to certify that the post-accident convalescence of the Hon. Winston S. Churchill necessitates the use of alcoholic spirits especially at meal times. The quantity is naturally indefinite, but the minimum requirements would be 250 cubic centimeters” (Sharp, 2023). This system generated a substantial income for physicians and pharmacies, encouraged new interest groups and also the Mafia, and altered features of the healthcare industry. For example, at the beginning of Prohibition, Walgreens was a small pharmacy chain with 20 stores. It marketed medicinal whiskey actively, and by 1929 it had expanded to 525 stores.

In 1867, the Russian government sought to raise revenue by monopolizing the production of matches. The resulting matches were expensive. One box could cost two-thirds of a worker’s daily wage. And these matches were physically large. Impoverished entrepreneurs then started splitting each large match into four or five smaller matches and selling them separately at a lower price. The Russian government then increased the severity of punishment for match-splitting and also redesigned the matches so they would ignite only when struck against a government-produced matchbox.

Governments commonly attempt to raise the price of certain products, sometimes by becoming a new buyer in the market and commonly underestimating the elasticity of supply. The classic example is agricultural price supports. These subsequently lead to increased output which soon requires production controls which in turn cascade into other outcomes which require additional regulation.

An extreme example of such government intervention occurred during World War II (WWII) when the Allies sought to prevent Germany from obtaining tungsten, essential for weapons production. At that time, Spain was selling tungsten to

Germany. To foreclose this source for Germany, the Allies started trying to purchase the entire supply of Spanish tungsten, and the price was rapidly bid up, eventually increasing 15-fold. The supply response in Spain was large, indeed much larger than expected—from 20 tons annually before WWII to 2000 tons annually late in the war—a hundred-fold increase (Bowen, 2006).

One further example of supply response: in Canada, the government of British Columbia attempted to reduce overfishing by reducing the number of fishing licenses and by shortening the fishing season from 60 days to 6 days. Fishermen then bought larger fishing vessels. During their new 6-day season, the total volume of halibut they caught was 50% greater than before. This also meant that most consumers could only obtain frozen halibut, rather than fresh (Libecap, 2018; Ménard & Shirley, 2022).

17.2.2 Substitution of Other Attributes of Goods

Regulation also affects the equilibrium attributes of goods. Even “simple” goods have many attributes beyond their money price, such as size, color, quality, reliability, warranty, availability of credit, associated service, location, and waiting time. All of these can vary with market conditions and regulations. The standard economic model assumes that money price and quantity are the two dimensions along which adjustments take place between buyers and sellers and that these adjustments continue until net gains are zero for the marginal unit exchanged. Through these adjustments, sellers and buyers minimize the deadweight loss. As Yoram Barzel (1997) has discussed, the same logic applies to the non-money attributes of goods: buyers and sellers jointly adjust these other attributes until the net gains are zero for marginal changes there. When a regulation alters one attribute (or a few—it can never control all of them), it is in the mutual interest of buyers and sellers to minimize the deadweight loss by varying other attributes.⁴

Depending on the substitution possibilities, therefore, regulation of money prices can lead not only to shortages or surpluses and welfare losses but also to changes in other attributes. Consider the price ceilings on gasoline in the United States in effect 1971–1974. Faced with regulated prices set below the market-clearing level, service stations reduced their hours of business to the minimum needed to sell their allocation, lowered their quality of service, and offered lower-octane fuel. Customers

⁴Price controls typically involve attributes that are relatively straightforward for the regulators to measure and monitor, like number of units, weight, and money price, rather than aspects that are more costly to measure, like quality, reliability, and associated service. Similarly, empirical work tends to focus on attributes that can be measured at low cost to the researcher. Studies of gasoline regulation typically focus on money price, volume, and octane rating (for which data are readily available) and not on the quality of service and waiting time. However, real-world responses to regulation may also involve attributes that are very costly for regulators and researchers to measure. But these can involve significant transaction costs, e.g., long waits in line, or significant externalities, e.g., people waiting in their cars with their engines running.

waited longer in line to obtain gasoline. They purchased new vehicles designed with larger gas tanks. Buyers and sellers jointly sought the lowest cost adaptation to the regulation.

Hong Kong rent controls provide another illustration. The first regulations, imposed in 1921, prevented landlords from raising rents except when they demolished an existing building and replaced it with a new one. In such cases, the existing tenants received no compensation. A building craze ensued. By 1923, some landlords were even replacing buildings that were only 2 years old, and many former tenants were sleeping in the streets (Cheung, 1975).

The building owners captured some of the returns from attribute substitution (a building just constructed vs. a building already some time in use), but many potential gains were dissipated by excessive construction. Even when rent controls carried over into new buildings, the new tenants would give the owners “key money” for access to the apartments. These side payments could not easily be regulated because they could be disguised as the sale of other items such as old chairs. In principle, key money could be paid on a continuing basis to keep a landlord from tearing down a building, but this was illegal, and it also involved high costs of collective action, since most buildings had multiple tenants.

Eventually, in 1955, under another rent control regulatory system, the government introduced an enforced-compensation scheme for displaced tenants. This new scheme clarified ownership rights and greatly reduced the transaction costs of negotiations between landlords and tenants. It better aligned incentives for recognizing costs borne by displaced tenants and costs for new construction. Steven Cheung argues that the Hong Kong rent control system at that time was highly efficient compared to rent control systems elsewhere in the world. He argues further that Hong Kong’s continuing growth through the twentieth century would not have been possible without relatively low transaction costs between tenants and landlords.⁵

17.2.3 Substitution of Amenities

Certain regulations, such as ceilings on money profits in firms, lower the cost to the firms’ decision-makers of consuming amenities. Whenever the money profits of a regulated firm are potentially greater than the maximum permitted, this reduces the opportunity cost to the firm of having elegant company dining rooms, chauffeured cars, congenial colleagues, good relations with the unions, nepotism, and the quiet life (Alchian & Kessel, 1962). During the years when US banking regulations severely constrained competition among banks, imposing fixed interest rates,

⁵Cheung (1979). Over time the tenant-landlord problem also diminished because builders increasingly sold apartments as condominiums rather than renting them out. This movement toward owner-occupants rather than tenants is a form of vertical integration which addresses some problems of incentive alignment. It depends on the existence of capital markets with low transaction costs for prospective condominium owners.

restricted entry, and limited branch banking, the top executives in that industry should have led less stressful lives than their counterparts in more competitive industries. A study of executives' life expectancy during that time period found that bankers indeed had a longer average life expectancy than did executives in other industries. This is consistent with the view that these regulations reduced the costs to the bank executives of choosing a desirable lifestyle (Yen & Benham, 1986). Of course the bankers working in this environment benefitted, but the reduced competition implied higher costs for the consumers of bank services.

Regulations and taxes are often less onerous for non-profit organizations than for for-profit firms. This lowers the cost of non-profits surviving, and the "profits" can be taken out in alternative forms like salaries and amenities, in retaining less productive colleagues or undertaking costly and unproductive missions.

17.2.4 Substitution of Barter and Other Arrangements

When regulation lowers the transaction costs of using barter instead of money, barter will obviously become more prevalent. Trade restrictions and foreign exchange controls can have a major impact here. In recent years, approximately 10% of world trade has been conducted through barter.⁶ Historically, barter on a wide scale has been a common occurrence. Consider the case of Germany in the period around World War II. In 1936, before the war, the German government froze prices. At the end of the war, the occupying military governments retained these price controls. In 1947 the money in circulation was ten times as great as in 1936, and real income had fallen by half. Official prices did not reflect the existing scarcity, and the black market was confined to a narrow range of consumer goods. Barter was illegal, but in order to obtain raw materials, firms developed a complicated chain of barter arrangements. Even under the direct control of the British military government, Volkswagen utilized 5% of its production for barter arrangements. Legal wage levels were set too low to secure a regular supply of employees, so employers compensated employees with goods in kind, which could then be used to barter for food (Carlin, 1989). Against the devastation of the war and the harsh regulations of the period, barter and other adaptations kept real income and output from falling further. Subsequently in Eastern Europe during the Soviet central planning period, money prices were suppressed, and barter was widespread as an adaptation. For example, a champagne producer in Hungary regularly sent out truckloads of champagne to barter for machinery repair parts.

In addition to barter, many other substitutes for money arise where they are cost-effective. Radford gives an instructive example in his study of the economics of a prisoner-of-war camp (Radford, 1945). Money was very limited in the camp, and

⁶Even in open market conditions, barter is still sometimes the lowest-cost way to trade. Barter exchange among US corporations is growing, partly because computers allow low-cost measurement and tracking of barter arrangements. See Akbar Marvasti and David Smyth (1998).

barter began. This was then superseded by the use of a money substitute as the medium of exchange, cigarettes, which were reasonably homogeneous, divisible, and durable.

17.2.5 Substitution of Vertical Integration

Regulations such as price controls and cartel pricing enforcement can alter the relative advantages of vertical integration versus market exchange. Price controls decrease the relative cost of internalizing production within the firm and can fundamentally affect transaction costs within and across firms. This was strikingly illustrated in the United States during and immediately after World War II, when the imposition of price controls led to a substantial increase in vertical mergers.

Another example comes from Germany in the late nineteenth century, when the government encouraged the formation of cartels. Thereupon the Rhenish-Westphalian Coal Cartel increased its coal prices substantially. After that, many electricity and gas utilities, railroads, and even municipalities acquired their own coal mines to produce their own coal and avoid the high cartel prices (Stockder, 1932).

Regulations that affect the existence or costs of particular markets such as future markets affect the incentives for vertical integration. Among its many roles, a future market serves as a synthetic storage mechanism and thus an alternative to vertical integration into storage. When the future market in oil began in the United States in 1983, this lowered the cost to firms of using the market to obtain oil in the future as compared with storing the oil themselves. This was followed by a reduction in vertical integration in the oil industry and by changes in the terms of contracting (Sykuta, 1994, 1996).⁷

Similar arguments apply to regulation of the insurance market. For example, firms seeking to keep the costs of theft down will provide fewer or more guards depending on the cost of insurance. Insurance is a partial substitute for guards. Firms, insurance companies, and regulators jointly influence preventative actions. Insurance also affects criminal incentives and criminal behavior, and this criminal behavior coevolves with the behavior of other participants. For example, insured property owners are less likely to use guards and violence to self-protect. Criminals have an interest in knowing that. The regulation of insurance thus involves complex tradeoffs. As an illustration, insurers of stolen artwork may have an interest in providing a reward for information from someone who “just happens to know” where that painting is hidden. This helps recover that painting at lower cost than fully compensating the legal owner, but this raises the returns to art theft overall. It is not so easy to sell stolen art—a registry of stolen art is available to potential buyers, and in addition an “innocent” buyer of stolen art runs the risk of having to return the stolen art without compensation to the original owner when the location of the art is

⁷The literature in economics and finance traditionally viewed future markets as hedging mechanisms, not as substitutes for vertical integration. See Marshall (1920) and Keynes (1930).

discovered. There are thus limits to how far various participants want these rewards to go leading to regulatory limits on such rewards.⁸

17.2.6 Substitution of Household Production

Households produce goods and services both for their own consumption and for trade with the external market (Yoram Ben-Porath, 1980). As Robert Pollak describes it (1985, pp. 605–606):

The transaction cost approach views marriage as a ‘governance structure,’ emphasizes the role of ‘bargaining’ within families and draws attention to the advantages and disadvantages of family and to the special roles of ‘altruism’ and ‘family loyalty.’ It also recognizes the disadvantages of family governance: conflict spillover, the toleration of inefficient personnel, inappropriate ability match, and inability to realize economies of scale. If activities are assigned to institutions in an efficient or cost-minimizing fashion, the balance of these advantages and disadvantages plays a major role in determining which activities are carried out within families and which are performed by firms, nonprofit institutions, or the state.

Levels of transaction costs within the family and within the external market will affect the extent and type of household production. Regulations that increase the costs of impersonal exchange in the market will increase the importance of household production and other forms of personalized exchange. In the Soviet Union, when the market system was severely repressed, families vertically integrated into food production for themselves by working on their small garden plots in the countryside. In many settings, land regulations limit individuals’ ability to obtain property rights that are clear and protected. This leads to difficulties in finance and contracting. Houses built in such settings are often constructed by the hands of the household members, with bricks purchased when cash is available (de Soto, 1989).

17.2.7 Substitution of Personalized Exchange

One way to characterize regulatory regimes is the extent to which there is impersonal exchange, personalized exchange, or little exchange at all. Indeed, Douglass North (1991) has described the process of economic development as a movement from personal to impersonal exchange.

Impersonal exchange requires low costs of measurement and enforcement. Adam Smith’s invisible hand depends upon a regulatory regime that encourages impersonal exchange by keeping transaction costs low. By altering such costs, regulation affects the extent of impersonal exchange. Some regulations like

⁸“Insurance as crime governance’ stimulates demand for security, shapes criminal incentives, engages with the state to combat crime, and tolerates some crime in the interest of profitability” (Baker & Shortland, 2023).

standardization can lower measurement and enforcement costs and thus promote impersonal exchange. Many other kinds of regulations raise transaction costs and thereby promote more personalized exchange. When regulatory regimes are arbitrary, impose excessive entry costs, establish price controls, or generally cause high transaction costs, then personalized exchange is likely to dominate.⁹ For example, in late twentieth-century Egypt, the markets for many goods were heavily regulated. Shortages arose, and it was often difficult to obtain goods in the impersonal market. Egyptians adapted in part by developing their own reciprocity networks of personalized exchange. These informal networks increased the likelihood of locating rationed goods or finding jobs (Singerman, 1995).

17.2.8 Changes in Governance and Contractual Relationships

Firms and households shape their governance structure—the forms of contracting, ownership, and decision-making—in part by efforts to economize on transaction costs. Since regulations affect transaction costs, changes in regulations can affect governance structures. This argument holds a central place in the work of Oliver Williamson (Williamson, 1985; Williamson & Masten, 1995; Williamson, 1999) and in the associated literature on transaction cost economics.

To illustrate, consider the impact of unit banking regulations in the United States. Historically some US states prohibited bank takeovers by other banks, while other states did not. In states that permitted takeovers, the banks were more efficient and had higher profits. Why? Permitting takeovers increased competition. Where takeovers were not possible, managers were subject to less outside scrutiny. This was somewhat mitigated by changes in bank ownership structure. Where bank ownership is more concentrated, some individual shareholders are more likely to have sufficient stakes to make it worthwhile for them to monitor the managers. In the US states that prohibited takeovers, ownership concentration was indeed higher. This reduced the inefficiencies associated with the restrictions on takeovers, but it was not a perfect substitute for a takeover market (Schranz, 1993).

17.2.9 Changes in Organization of the Market

In many markets, regulatory standards exist for advertising, disclosure, and measurement. These regulations affect the types of information and goods produced and can give rise to dramatically different configurations of economic organization—sometimes improving efficiency and sometimes not. In the case of the market for eyeglasses, regulations in many US states historically restricted severely the types of advertising that could be provided to consumers. This greatly limited the ability of the large firms, which were the low-cost providers, to compete. These regulations

⁹Trade that involves political favors usually involves personalized exchange.

were also associated with higher prices of eyeglasses to consumers, lower frequency of purchase, and more adverse effects for those with less education (Benham, 1972; Benham & Benham, 1975).

Regulations that lower the costs of measurement, such as standardization of weights and measures, can increase enforceability and credibility, thereby enhancing the efficiency of markets. Regulations of quality standards for food can increase the credibility of brand names. Together these appear to have assisted the development of the market for manufactured foods. Food-manufacturing firms in the United States at the beginning of the twentieth century recognized this, and they themselves supported the introduction of quality-control regulations (Law, 2003a, b).

Although seemingly innocuous, even regulations concerning how land is measured can have a substantial impact. Libecap and Lueck show this in their study of the two land delineation systems used in Ohio in the early nineteenth century: (1) the metes and bounds method of land measurement, based on natural features of the landscape and artificial markers, and (2) the grid system, based on centralized grid of uniform square plots. Two centuries later, land measured with the grid system had substantially higher land prices, lower levels of violence, and higher population density relative than nearly identical land measured in metes and bounds (Libecap & Lueck, 2011).

Note that even if a regulatory environment lowers the costs of measurement and enforcement, this does not mean that the total resources in the economy devoted to transaction costs will necessarily decline. The transaction costs per transaction may fall, but the total number of transactions may increase dramatically, so that total transaction costs rise. John Wallis and Douglass North (1986) have evidence that bear on this issue. Between 1870 and 1970 in the United States, specialization and per capita income grew enormously. Simultaneously, the transaction sector—those activities which involve measurement and enforcement, such as accounting and law—increased from 25% of GNP to 45% of GNP.

17.3 Illicit Responses to Regulation

Let us now consider various illicit responses.

17.3.1 Underground Economic Activity

The underground economy, the “informal economy,” includes (a) legal activities conducted outside the formal legal system (manufacture of legal goods without a permit, providing services like house repairs without reporting income to tax authorities), and (b) illegal activities (illegal drugs, stolen goods, prostitution, gambling). Informal economies play a significant role throughout the world; their relative

importance varies by country.¹⁰ Available evidence suggests that variation in regulatory climate is a major reason for variation in the size of the underground economy.

Of particular relevance here are barriers to entry, perhaps the most intensively studied form of all regulation. In 1776 Adam Smith gave them a central place in *The Wealth of Nations* (1776). In the 1980s Hernando de Soto (1989) documented startlingly high regulatory barriers to small-business entry in Peru.¹¹ A great proportion of business in Peru was transacted informally. In the early 2000s, Djankov et al. studied regulations governing business entry in 85 countries and found high official costs of entry in most countries. Djankov et al. (2002) and Djankov (2003) also found that where regulation of entry is more extreme, the unofficial economy is larger, corruption is higher, and the quality of public service is not better.

17.3.2 Collusive Activities and Norms Within Firms

If the norms encouraging group cooperation are sufficiently strong and the formal governance structure is sufficiently weak, then long-term violation of the formal rules can be sustained. In 2011 the Japanese Olympus Corporation was exposed as having hidden \$1.7 billion US in losses for decades. How could this persist for so long, and what was the catalyst for exposure? Michael Woodford, a British executive who had worked his way up through the ranks of the Olympus Corporation, became CEO in October 2011. This was the first time a non-Japanese had been named to the top role at Olympus. Woodford noticed unusual large transactions, and after documenting some of the specifics, he urged the Board to resign and to allow a full investigation of the company's accounts. The Board voted unanimously to dismiss him from his position as CEO, citing a management culture clash.

After his dismissal, Woodford made his concerns public. The ensuing scandal led to the resignation of the company's chairman and president and to official investigations in Japan, the United Kingdom, and the United States. They confirmed that Olympus had indeed been hiding substantial investment losses for more than a decade. The scandal led to an 80% decline in the company's share price and criminal charges for several senior Olympus executives. The extent to which there are powerful cooperative norms within an organization—and Japan is a comparatively homogeneous society with strong norms of cooperation—makes such private arrangements more sustainable.

¹⁰In year 2000, the share of the economy that was informal averaged 41% in developing countries, 38% in transition countries, and 18% in OECD countries. Differences within regions were substantial: Bolivia 67% vs. Chile 20%, Zimbabwe 59% vs. South Africa 28%, Thailand 53% vs. Japan 11%, and Greece 29% vs. Switzerland 9%. See Friedrich Schneider (2002).

¹¹In an experiment, de Soto demonstrated that a person seeking legal permission to open a very small clothing factory on the outskirts of Lima in 1983 needed to spend 289 days of full-time effort and pay an irreducible minimum of two bribes to obtain the necessary permits from the bureaucracy to operate legally.

17.3.3 Discrimination

Regulations can alter the price of discrimination and thereby influence the extent of discrimination. They can also work indirectly. For example, if the money profits of a regulated firm are restricted below the level achievable, this will lower the cost to the firm of employing workers who are less productive but who have preferred characteristics along other dimensions. In other words, regulatory limits on profits will lower the cost to the firm of engaging in discrimination (Alchian & Kessel, 1962).

The regulation of US physicians in the early 1900s provides another illustration. At that time, medical licensure came under the monopoly control of state medical associations which regulated standards. This strengthened the medical associations' position as an interest group, and they dramatically reduced the number of medical schools. As a result, physicians' earnings increased, and applications to the surviving medical schools increased. The long waiting lines for applications for admission lowered the cost of discrimination to the admissions committee. The relative numbers of women and blacks admitted to medical schools then declined sharply (Kessel, 1970).

How a change in the rules can affect the price of discrimination and hence its extent is also illustrated by a change in regulations in US baseball. Until 1947, baseball was a segregated sport with a separate Negro league. This color bar broke down in 1947, permitting all teams to hire players from the Negro league, which had excellent players available at lower salaries. The teams quickest to stop discriminating then improved both their team performance and their box office returns relative to the remaining teams. As the color bar diminished, baseball owners who persisted in discriminating paid a higher price for their preference. Eventually the color bar fell for all teams (Gwartney & Haworth, 1974).

17.3.4 Private Coercion and Extralegal Organizations

Regulations that raise the cost of engaging in voluntary exchange will affect the extent to which private coercion is used, individually and through illegal organizations. Regulations such as prohibitions on alcohol, drugs, and gambling increase the potential gains of dealing in these arenas and thereby facilitate the rise of illegal organizations like the Mafia.

Regulations that increase uncertainty concerning ownership of property rights can lead to environmental problems like land invasions, violence, and deforestation. In Brazil, the conflicted regulatory environment concerning property rights has led to all of these (Alston et al. 1999, 2000). In southern Italy in the nineteenth century, the formal rules were changing, but the institutions for enforcement were weak; hence the claims of ownership and control became less certain. Violence and threats of violence became more relevant for de facto ownership. This led to a great increase in the strength of the Mafia which had a comparative advantage in violence and contract enforcement (Gambetta, 1993).

17.3.5 Corruption

Corruption is the use of public office for private gains in carrying out a public task. Regulations can significantly affect the benefits and costs of engaging in corrupt practices. In settings where regulations are excessive or arbitrary and where civil servants have discretion over enforcement, where discretionary permits are highly valuable, and/or where the likelihood of exposure is low, in these settings corruption is likely to be extensive.

The number of possible corrupt practices is very large. The organization of corruption and the property rights to associated payoffs vary considerably by country (Shleifer & Vishny, 1993). Where the property rights to bribes are clearly held and enforced, a given level of corruption may produce less inefficiency. If the highest-ranking politicians and political parties are disciplined enough to refrain from seeking additional payoffs after initial lump sums have been paid, then corruption on the margin may be circumscribed. In South Korea, for example, political parties are highly disciplined, and bribery has been highly centralized. Many payoffs are lump-sum contributions by major business leaders to presidential campaigns. This arrangement does not tax economic activity at the margin. If the government is weak and fragmented, however, local government regulators often engage in decentralized looting. This undermines credible commitments in both public and private sectors and can be particularly damaging to economic performance (Bardhan, 1997).

Regulators are often active in this process. They may intentionally introduce regulations to create further opportunities for corrupt exchanges. If corruption becomes common in one sector (say, because of high import duties or restrictions on gambling), then the cost of spreading corruption to other sectors falls. This in turn lowers the returns to normal entrepreneurial activity, which further slows down economically productive activity, which induces more people to engage in corrupt practices, and so on (Murphy et al., 1993). Furthermore, it may be easier to introduce corrupt practices than to reduce them: long periods of time may be needed to move from a corrupt to a non-corrupt equilibrium.

Accession of Eastern European countries into the EU provides a natural experiment of the impact of an external change in regulation oversight procedures on corruption. When a new country aspires to join the European Union (EU), it must undergo a rigorous accession process that requires it to adopt, implement, and enforce a wide range of EU rules and regulations. The role of the legacy interest groups within the country is thus altered. This new external oversight process has helped reduce corruption problem in the countries admitted, but the experience varies. Corruption is more persistent where the old legacies and networks are more enduring (Ganev, 2019).

17.4 History, Path Dependence, and Interest Groups

Let us turn now to Oliver Williamson's view of the role of history and path dependence. In most of the standard literature on regulation, these have been ignored. However, historical experience is fundamental to the formation and

Table 17.2 Williamson: levels of institutional constraints

Level of embeddedness
Informal institutions, traditions, norms, religion
Frequency of change: 100–1000 years
Level of institutional environment
Formal rules of the game—especially property (polity, judiciary, bureaucracy)
Frequency of change: 10–100 years
Level of governance
Play of the game—especially contract (aligning governance structures with transactions)
Frequency of change: 1–10 years
Level of resource allocation and employment
Prices and quantities, incentive alignment
Frequency of change: continuous

evolution of formal and informal rules, as well as to the consequences of any new formal regulation. Furthermore, one regulation often leads to others, and the historical reach can be long. There is path dependence in regulation, substantially rooted in the political process and in historical experience.

Oliver Williamson's classification of the different levels of institutional constraints, shown in Table 17.2, provides a perspective on the different time frames involved. Neoclassical economics has focused almost entirely on the bottom level, resource allocation, and employment, where adaptation to regulation is restricted to changes in prices and quantities and where adaptation is continuous. However, countries have differing historical experiences, norms, and traditions. In Williamson's terminology, "embeddedness" differs. The specifics of Table 17.2 can be disputed; for example, norms and religions may change more frequently than every 100–1000 years, but this perspective is valuable in emphasizing the varying time spans over which institutional changes can take place. At any point in time, the formal rules, informal norms, and their enforcement characteristics vary substantially across countries. The impact of any new formal regulation—and responses to it—will be affected by these elements and by their historical patterns within a country.

Depending on context and time period, therefore, a particular type of regulation can give rise to very different responses in different settings. The direct allocative effects of a regulation—measured at Williamson's levels of resource allocation and governance—are often only a small part of its total impact.

Small and seemingly unimportant details of regulations can lead to large downstream effects.¹² Or apparently far-reaching regulations may turn out to be

¹²The impact of any regulatory change will depend in part on the informal rules. If a regulatory change is greatly at variance with the preferences and interests of a particular group, their informal norms can evolve into "opposition norms." These opposition norms can have highly negative consequences for performance. See Victor Nee (1998).

inconsequential because of innovative adaptations over time. While standard micro-economic analysis of regulation invokes the notion of the long run to some extent, in the perspective of new institutional economics, the long run is usually longer, more varied, and more significant. The following case studies illustrate how regulations, constrained by path dependence, can generate long-run responses by interest groups with further policy consequences.

17.4.1 Broadcasting in Britain

Small initial regulatory choices can lead to large downstream effects. Ronald Coase found this in his detailed investigation of the origins of radio broadcasting in Britain (Coase, 1947). Initially, the transmission of sound by radio was regarded as a new means for sending telephone messages. Because of this, regulation of broadcasting was assigned to the organization that had jurisdiction over telephones: the Post Office. To avoid the problem of broadcast interference and to avoid having to select from among the many firms who would seek the valuable radio spectrum, the Post Office favored having a single broadcasting company.¹³ Strategies that could have been used to allocate the radio spectrum, including assigning property rights to the spectrum and selling them off, were not on the agenda at that time.¹⁴ The main advocates for radio broadcasting at that time were the radio manufacturers, who sought their revenues from the sale of radios rather than from broadcasting. Newspapers did not want competition for their advertising revenue and therefore opposed commercial forms of radio broadcasting. There was thus little opposition to the Post Office's proposal to create only one broadcasting company and to finance its programming and transmission from taxes on radio sales. There followed a broadcast monopoly that limited British citizens to BBC programming for many decades. Non-BBC programming could not even be transmitted by wire (although that involved no problem of externalities). When television was eventually introduced, the BBC monopoly was extended to cover television.

In the years just before World War II, the BBC gave the major political parties control over access to the airwaves for political speeches. The political parties then allocated their shares of airtime to politicians who endorsed their standard party positions. When Winston Churchill sought to criticize the government's appeasement policy toward Hitler, he was unable to get permission to speak on the BBC.

¹³C.A. Lewis, deputy director of the British Broadcasting Company, stated in 1923, "The chaotic state of affairs in America, where a large number of stations are transmitting on a narrow band of wavelengths and no form of control exists, was an object lesson in what not to do, and consequently the control was put into one company's hands" Lewis (1924), quoted in Coase (1947), p. 208, fn. 2.

¹⁴"Only many decades later were regulations introduced to define some rights to some spectrum such that sales could take place. The common view at the time among experts was that the regulation came as a necessary result of reckless competition among broadcasters, which retarded the orderly development of radio and subjected the listeners to intolerable strain" Coase (1959), p. 13.

Given the rules of the game, this result is not surprising. An apparently small regulatory choice driven by specific historic circumstances had these profoundly important later consequences.

17.4.2 US Sugar Production

Sugar production in the United States offers another example of the downstream impact of initial regulatory decisions. Sugar is the only US agricultural good now excluded from global free trade. In 2021 raw sugar prices in the United States were twice as high as the world spot price. All 11,800 US sugar farmers receive subsidies. As a result of these subsidies, the Everglades region in Florida has been planted with sugar cane, creating significant ecological problems. Sugar-processing and sugar-using industries have moved out of the United States because of the high domestic sugar prices. The total costs of this program are vast. How did all this happen?

Anne Krueger's classic study (1990) shows how the complexity and the cost of current sugar regulation can be understood only by an appreciation of past policies. Through most of the nineteenth century, 90% of the sugar consumed in the United States was imported. Tariffs on sugar were originally imposed as a revenue source. Then in the late 1800s, the US Department of Agriculture took initiatives to encourage domestic sugar growing and processing. The sugar industry developed because of this government-initiated program, rather than the reverse.¹⁵

Several historical accidents appear to have had a major impact on the subsequent evolution of the program, including idiosyncratic committee assignments in Congress and a focus on foreign relations issues. Once the very complex sugar program was in place, a network of program specialists arose in government and industry and became independently influential. Various interest groups were involved over time. A long-standing—and unlikely—coalition of sugar producers, processors, and users persisted until 1978. At that time a new corn-based substitute for sugar began to take market share. Corn producers benefited from having high prices for sugar, and they therefore supported a sugar quota system, opposing plans that would have permitted the US sugar price to fall to the world level.

The high domestic price of sugar has led to many market responses: the use of sugar substitutes, the importing of products such as cake mixes with high sugar content, and the subsequent ban on imported cake mixes. Diplomatic complications have been associated with the allocation of sugar import quotas across countries, e.g., Cuba. Although this program has been very costly to consumers and to sugar-using industries, there appear to be few long-term rents earned by domestic sugar producers and processors. Most of the subsidies are dissipated in inducing production where it would not otherwise take place.

¹⁵In 1934, a quota system that in fact benefited the sugar producers was imposed over their objections. They learned from experience, and 3 years later they began to support the quota system.

Examples from many other industries could be provided. The regulation and deregulation of natural gas, standards for automobile emissions, pharmaceutical regulation, medical licensure, and airline regulation all show a history of interest group politics in which new regulations frequently create new interest groups.

17.4.3 The Old Believers in Russia

The previous examples have shown that small initial regulations can lead step by step to major downstream consequences. On the other hand, Draconian regulations are sometimes followed by surprisingly resilient responses over long time periods. Consider the case of the Old Believers in Russia.

Starting in the seventeenth century, the Old Believers were heavily persecuted for refusing to conform to the new religious doctrines of the Russian Orthodox Church. During the period of most intense persecution, the Old Believers were subjected to removal of all their legal rights, forced exile, no standing in court, and even mass executions. Although the intensity of persecution diminished over time, they remained for centuries without full legal status. However, the Old Believers survived and eventually prospered by developing strong informal norms of trust, honesty, sobriety, and cooperation within their own community. These norms lowered their internal transaction costs substantially, permitting them to trade over space and through time in ways that outsiders in the broader Russian society could not. By the mid-nineteenth century, in relatively unregulated sectors like the textile industry, they were the leading entrepreneurs in Russia. When the government's direct regulatory role increased in the late nineteenth century, their role in this sector then diminished (Raskov, 2002).

In the Russian economic system, informal norms evolving over long periods of time had fundamental consequences for who did what and how efficiently. When excluded from formal state contract enforcement, the Old Believers developed their own informal system of transacting and enforcement. The later state regulations that elevated the importance of political connections, such as state licensing for entry, put them at a disadvantage.

These three historical examples illustrate the broad range of responses to regulation that arise from interest groups rather than from individual action. New regulations create new interests and destroy old interests. The rise and demise of shared interests can change the original politics irreversibly. The overall response to a regulation depends only partially on responses at the level of resource allocation. How the various interest groups respond to the regulation and which ones seize the regulatory mechanisms over time are very significant issues. At this stage in our knowledge, these outcomes are extremely difficult to predict.

17.4.4 New Technology

Innovation creates new possibilities and is essential for long-term growth. However, when new technology is changing the landscape rapidly, it can also create new

opportunities to circumvent the existing regulatory policy. This is illustrated by the 2023 trial of Sam Bankman-Fried, CEO of FTX, one of the largest exchange platforms for cryptocurrencies. After his exchange collapsed, trial testimony showed that adding just one line—“allow_negative”—to the FTX code permitted Bankman-Fried to access 65 billion dollars of his customers’ accounts without their permission. Public trials such as this make clear the limited extent to which regulations protect customers, especially where technology is evolving rapidly. This highlights the importance of understanding alternative approaches to address externalities, risks, and uncertainties. Options include not only the fine-tuning of complex regulation—which can and often does retard innovation while simultaneously giving citizens a false sense of security—but also educating citizens concerning risks and caveat emptor.

17.5 Conclusion

Ronald Coase’s fundamental insight that the organization of economic activity is determined by transaction costs was transformative. This is the central building block for new institutional economics. This insight also applies to the consequences of regulation. As illustrated in the examples above, regulation plays out on many margins usually with surprising adaptations. How a regulation plays out depends on the alternatives available under the new rules and their costs. The adaptations to a regulation and the speed of adaptation cannot be determined by ad hoc assumption about options and transaction costs.

New institutional economics emphasizes the role of history, path dependence, and the variability of experience across individuals, firms, and countries. A long-run task is to articulate a general approach that covers both economic and political actors and both individual and collective actions. A prudent scholar would examine policies with a keen eye on the process of creation and destruction of interests.¹⁶ If we can understand the potential consequences of regulators’ actions in this broader sense, supplemented by detailed studies of specific regulations in many countries, we may do a better job of understanding the impact of regulation, responses to it, and the regulatory process itself (Coase, 2002; Krueger, 1990).

Acknowledgments I thank Alexandra Benham for many discussions of these ideas and for her generous work in commenting on and editing this manuscript several times. Kurt Annen, Krishna Ladha, Claude Ménard, Marc Law, John Nye, Mary Shirley, Gili Yen, and two anonymous referees have provided helpful comments. Any errors remaining are mine.

Acknowledgment of Funding Open access was made possible through a generous donation in honor of the Ronald Coase Institute.

¹⁶Krishna Ladha (2002) has provided edification on this topic.

References

- Alchian, A., & Kessel, R. (1962). Competition, monopoly, and the pursuit of pecuniary gain. In *National Bureau of economic research, aspects of labor economics* (pp. 157–175). Princeton University Press.
- Alston, L., Libecap, G., & Mueller, B. (1999). *Titles, conflict, and land use: The development of property rights and land reform on the Brazilian Amazon frontier*. University of Michigan Press.
- Alston, L., Libecap, G., & Mueller, B. (2000). Land reform policies, the sources of violent conflict, and the implications for deforestation in the Brazilian Amazon. *Journal of Environmental Economics and Management*, 39(2), 162–188.
- Baker, T., & Shortland, A. (2023). Insurance as crime governance: How crime shapes insurance and insurance shapes crime. *Journal of Legal Analysis*, 15, 183, U of Penn, Inst for Law & Econ Research Paper No. 23-10.
- Bardhan, P. (1997). Corruption and development: A review of issues. *Journal of Economic Literature*, 35(3), 1320–1346.
- Barzel, Y. (1997). *Economic analysis of property rights* (2nd ed.). Cambridge University Press.
- Benham, L. (1972). The effect of advertising on the price of eyeglasses. *Journal of Law and Economics*, 15(2), 337–352.
- Benham, L., & Benham, A. (1975). Regulating through the professions: A perspective on information control. *Journal of Law and Economics*, 18(2), 421–448.
- Benham, A., & Benham, L. (2000). Measuring the costs of exchange. In C. Ménard (Ed.), *Institutions, contracts and organizations: Perspectives from new institutional economics* (pp. 367–375). Edward Elgar.
- Benham, A., & Benham, L. (2001). Costs of exchange. <https://www.coase.org/workingpapers/wp-1.pdf>. Accessed 24 Oct 2023.
- Ben-Porath, Y. (1980). The F-connection: Families, friends, and firms and the organization of exchange. *Population Development Review*, 6(1), 1–30.
- Bowen, W. H. (2006). *Spain during world war II*. University of Missouri Press.
- Carlin, W. (1989). Economic reconstruction in Western Germany, 1945: The displacement of ‘vegetative control. In I. D. Turner (Ed.), *Reconstruction in post-war Germany, British occupation policy and the Western zones, 1945–55* (pp. 37–65). Berg.
- Cheung, S. N. S. (1975). Roofs or stars: The stated intents and actual effects of a rents ordinance. *Economic Inquiry*, 13(1), 1–21. Also in *Empirical Studies in Institutional Change*, Lee J. Alston, Thrainn Eggertsson, and Douglass C. North, eds. Cambridge: Cambridge University Press, pp. 224–243.
- Cheung, S. N. S. (1979). Rent control and housing reconstruction: The postwar experience of prewar premises in Hong Kong. *Journal of Law and Economics*, 22(1), 27–54.
- Coase, R. (1947). The origin of the monopoly of broadcasting in Great Britain. *Economica*, 14, 189–210.
- Coase, R. (1959). The federal communications commission. *Journal of Law and Economics*, 2, 1–40. Also in the Yearbook of Broadcasting Articles, anthology edition. 1980. Washington DC: Federal Publication.
- Coase, R. (1960). The problem of social cost. *Journal of Law and Economics*, 3, 1–44.
- Coase, R. (2002). Personal communication, 23 October 2002.
- de Soto, H. (1989). *The other path*. Harper & Row.
- Djankov, S. (2003). The informal economy: Large and growing in most developing countries. World Bank Hot Topic Discussion Board. <http://rru.worldbank.org/hottopics/HotTopicsdjankov2.asp>. Accessed 16 Oct 2003.
- Djankov, S., La Porta, R., Lopez-De-Silanes, F., & Shleifer, A. (2002). The regulation of entry. *The Quarterly Journal of Economics*, 117, 1–37.
- Gambetta, D. (1993). *The Sicilian mafia: The business of private protection*. Harvard University Press.
- Gambino, M. (2013). *Smithsonian Magazine*, 7 October.

- Ganev, G. (2019). Bulgaria's economy 1989–2019: An open-ended story of structural changes. *Munich Personal RePEc Archive*. 11 July.
- Gwartney, J., & Haworth, C. (1974). Employer costs and discrimination: The case of baseball. *Journal of Political Economy*, 82, 873–881.
- Kessel, R. (1970). The AMA and the supply of physicians. *Law and Contemporary Problems*, 35(2), 267–283.
- Keynes, J. M. (1930). *A treatise on money, volume 2, the applied theory of money*. Macmillan.
- Krueger, A. O. (1990). The political economy of controls: American sugar. In M. Scott & D. Lal (Eds.), *Public policy and development: Essays in honour of Ian Little*, Reprinted in Empirical studies in institutional change, Lee J. Alston, Thrainn Eggertsson, and Douglass C. North, eds. 1996. Cambridge University Press, pp. 169–218.
- Ladha, K. (2002). Personal communication, August 2002.
- Law, M. (2003a). *Specialization, information, and regulation in American economic history*. Ph.D. dissertation, Washington University, St. Louis.
- Law, M. (2003b). The origins of state pure food regulation. *Journal of Economic History*, 63, 1103–1130.
- Lewis, C. A. (1924). *Broadcasting from within*. George Nennes Limited.
- Libecap, G. (2018). Douglass North: Transaction costs, property rights and economic outcomes. <https://www.nber.org/papers/w24585>
- Libecap, G., & Lueck, D. (2011). The demarcation of land and the role of coordinating property institutions. *Journal of Political Economy*, 119(3), 426–467. (42 pages).
- Marshall, A. (1920). *Industry and trade* (3rd ed.). Macmillan.
- Marvasti, A., & Smyth, D. (1998). Barter in the US economy: A macroeconomic analysis. *Applied Economics*, 30, 1077–1088.
- Ménard, C., & Shirley, M. M. (2022). *Advanced introduction to new institutional economics* (pp. 50–51). Edward Elgar.
- Murphy, K., Shleifer, A., & Vishny, R. (1993). Why is rent-seeking so costly to growth? *American Economic Review*, 83(2), 409–414.
- Nee, V. (1998). Norms and networks in economic and organizational performance. *The American Economic Review*, 88(2), 85–89. Papers and Proceedings.
- Noll, R., & Owen, B. (1983). *The political economy of deregulation: Interest groups in the regulatory process*. American Enterprise Institute for Public Policy Research.
- North, D. (1991). Institutions. *Journal of Economic Perspectives*, 5(1), 97–112.
- Nye, J. (2007). *War, wine and taxes*. Princeton University Press.
- Olson, M. (1982). *The rise and decline of nations*. Yale University Press.
- Pollak, R. A. (1985). A transaction cost approach to families and households. *Journal of Economic Literature*, 23(2), 581–608.
- Radford, R. A. (1945). The economics of the P.O.W. Camp. *Economica*, 12, 189–210.
- Raskov, D. E. (2002). *Norms in the economic evolution: Old believers in the Russian nineteenth-century textile industry*. Paper presented at the 2002 conference of the International Society for New Institutional Economics, Cambridge, MA.
- Schneider, F. (2002). *Size and measurement of the informal economy in 110 countries around the world*. Paper presented at a workshop of the Australian National Tax Centre, Canberra, Australia, 17 July 2002. <http://rru.worldbank.org/documents/informaleconomy.pdf>. Accessed 16 Oct 2003.
- Schranz, M. (1993). Takeovers improve firm performance: Evidence from the banking industry. *Journal of Political Economy*, 101(2), 299–326.
- Sharp, C. (2023). Churchill's doctor prescribed him 'naturally indefinite' amount of alcohol after car crash. *Express News*. 21 May 2023.
- Shirley, M. (Ed.). (2002). *Thirsting for efficiency: Experiences in reforming urban water systems*. Elsevier.
- Shleifer, A., & Vishny, R. (1993). Corruption. *Quarterly Journal of Economics*, 108(3), 599–617.

- Singerman, D. (1995). *Avenues of participation: Family, politics, and networks in urban quarters of Cairo*. Princeton University Press.
- Smith, A. (1776). *The wealth of nations*. University of Chicago Press, 1976 edition.
- Spiller, P. (2010). Regulation: A transaction cost perspective. *California Management Review*, 52(2), 147–158.
- Stigler, G. (1951). The division of labor is limited by the extent of the market. *The Journal of Political Economy*, 59(3), 185–193.
- Stockder, A. H. (1932). *Regulating an industry: The Rhenish-Westphalian coal syndicate 1893–1929*. Columbia University Press.
- Sykuta, M. (1994). *Real effects of futures markets on firm and industry behavior: A Study of institutions and contracting in the crude oil industry*. Ph.D. dissertation, Washington University, St. Louis.
- Sykuta, M. (1996). Futures trading and supply contracting in the oil refining industry. *Journal of Corporate Finance: Contracting, Governance and Organization*, 2(4), 317–334.
- Wallis, J. J., & North, D. C. (1986). Measuring the transaction sector in the American economy. In S. L. Engerman & R. E. Gallman (Eds.), *Long term factors in American economic growth*. University of Chicago Press.
- Williamson, O. (1985). *The institutions of capitalism*. The Free Press.
- Williamson, O. (1999). The new institutional economics: Taking stock/looking ahead. *Newsletter of the International Society for New Institutional Economics*, 2, 2.
- Williamson, O., & Masten, S. (Eds.). (1995). *The economics of transaction costs*. Edward Elgar.
- Yen, G., & Benham, L. (1986). The best of all monopoly profits is a quiet life. *Journal of Health Economics*, 5(4), 347–353.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part IV

Organizations and Governance

Paul L. Joskow

Contents

18.1	Introduction	418
18.2	Comparative Governance Arrangements from a Transaction Cost Economics (TCE) Perspective: The Basic Story	419
18.3	Traditional Approaches to Explaining Vertical Integration	421
18.4	Incomplete Contracts and Asset Specificity	424
18.4.1	Asset Specificity in Detail	425
18.4.2	Effects on Ex Ante Investment and Ex Post Adaptation and Performance	427
18.4.3	Ex Post Contractual Adaptation and Performance Problems	430
18.5	Vertical Integration as an Alternative Governance Structure	432
18.6	Empirical Evidence	437
18.7	Conclusions	442
	References	443

Abstract

This essay reviews research that examines the choice between governance of vertical relationships involving suppliers of intermediate goods and services (“upstream”) and the purchasers of those goods and services (“downstream”) through some form of market-based contractual arrangement versus governance based on internalization of these transactions within a single hierarchical organization through vertical integration. The emphasis is on the transaction cost economics (TCE) framework for understanding the choice of governance arrangements, though I will briefly discuss several other theories of vertical integration as well. The general comparative governance framework that is a

P. L. Joskow (✉)

Department of Economics, Massachusetts Institute of Technology, Cambridge, MA, USA

e-mail: pjoskow@mit.edu

© The Author(s) 2025

C. Ménard, M. M. Shirley (eds.), *Handbook of New Institutional Economics*,

https://doi.org/10.1007/978-3-031-50810-3_18

417

basic feature of transaction cost economics (TCE) analysis is discussed first. The essay then proceeds to discuss traditional “neoclassical” theories of vertical integration that rely on various market imperfections including market power, double marginalization, free riding, uncertainty, and economies of scale as explanations. A discussion of theories of vertical integration from a TCE perspective, focusing on the role of incomplete contracts and relationship-specific investment, follows. The essay concludes with a brief discussion of the extensive empirical literature on vertical integration, focusing on key methodological issues.

Keywords

Vertical integration · Governance · Transaction costs · Relationship-specific investments · Hold-ups

18.1 Introduction

Understanding the factors that determine which types of transactions are mediated through markets and which within hierarchical organizations called firms has been an important subject of theoretical and empirical work in microeconomics generally and central to work in New Institutional Economics (NIE) in particular for at least the last 25 years. This essay reviews research that examines the choice between governance of vertical relationships involving suppliers of intermediate goods and services (“upstream”) and the purchasers of those goods and services (“downstream”) through some form of market-based contractual arrangement versus governance within an organization through vertical integration. My primary emphasis is on the transaction cost economics (TCE) framework for understanding the choice of governance arrangements, though I will briefly discuss several other theories of vertical integration as well.

The essay proceeds in the following way. First, I discuss the general comparative governance framework that is a basic feature of transaction cost economics (TCE) analysis and within which theoretical and empirical analysis of vertical integration associated with the NIE has proceeded. Next, I discuss traditional “neoclassical” theories of vertical integration that rely on market imperfections associated with market power, free riding, uncertainty, and economies of scale as explanations for vertical integration. I then proceed to discuss theories of vertical integration from a TCE perspective, focusing on the role of incomplete contracts and relationship-specific investment. This section includes a brief discussion of the “property rights” (PR) approach to vertical integration, emphasizing the similarities and differences between the TCE and PR frameworks. Finally, I provide a brief discussion of the extensive empirical literature on vertical integration that has been stimulated by this theoretical research, focusing on key methodological issues (Klein, 2005 reviews empirical studies of the “make or buy” decision in more detail).

18.2 Comparative Governance Arrangements from a Transaction Cost Economics (TCE) Perspective: The Basic Story

I want to emphasize at the outset that there is not and will never be one unified theory of vertical integration. While the literature on vertical integration tends to focus on a simple dichotomy between the decision to “make” internally and “buy” through the market, from a TCE or NIE perspective we must be sensitive to the fact that there are a wide array of market-based governance arrangements that represent alternatives to both simple anonymous repeated spot market transactions and vertical integration. These two governance arrangements are polar cases. Theoretical and empirical research in the NIE tradition examines not only the determinants of the boundaries between firms and markets but also the origins of various “hybrid forms” of governance structure that lie between simple anonymous spot market transactions and unified hierarchical organizations with varying expanses of vertical and horizontal control. These hybrid forms include various types of long-term contracts, joint ventures, dual sourcing (partial vertical integration), holding companies, and public enterprises. In addition, NIE research examines the attributes of internal organizations with different internal structures, incentive arrangements, and vertical, horizontal, and multi-product dimensions (Williamson, 1996, 2000). Accordingly, vertical integration is only one of many potential vertical governance arrangements that transacting parties may choose from and represents only one component of broader theories of the governance of contractual relationships and theories of the firm. Indeed, except in cases where there are significant market imperfections “...market mediation is generally to be preferred over internal supply...” (Williamson, 1971, p. 113).

Virtually all theories of vertical integration turn in one way or another on the presence of market imperfections of some type. Traditional approaches to vertical integration have tended to focus (though not exclusively as I discuss further below) on vertical integration as a response to pre-existing market power problems or as a strategic move to create or enhance market power in upstream or downstream markets. While not excluding these rationales for vertical integration, the NIE approach to the analysis of alternative market and internal organizational governance arrangements is much broader. It focuses on a well-defined array of attributes of individual transactions between buyers and sellers of goods or services and how they affect the performance (total cost) of alternative governance arrangements. It recognizes that there is a wide array of governance structures through which transactions can be mediated—from anonymous spot markets to internal administrative procedures within hierarchical organizations. It recognizes further that the task of consummating transactions must confront a variety of potential transaction costs, contractual, and organizational hazards, which are related to the attributes of the transactions at issue and their interplay with the attributes of alternative governance arrangements. These transactions costs involve the direct costs of writing, monitoring, and enforcing contingent contracts as well as the costs associated with the ex ante investment and ex post performance inefficiencies that arise as a consequence of

contractual hazards of various types and various bureaucratic costs associated with internal organization. The inefficiencies of particular interest are those that arise as a consequence of *ex post* bargaining, haggling, pricing, and production decisions, especially those that arise as the relationship must adapt to changes in supply and demand conditions over time, though inefficiencies in *ex ante* investments are also relevant (Williamson, 1975, 2000). The governance structures that are chosen, whether market or hierarchical, are those that are best adapted to the attributes of the transactions of interest in the sense that they economize on the total costs of the trading relationship. That is, governance mechanisms are chosen in an effort to reduce inefficiencies associated with both *ex ante* investment and *ex post* performance of a trading relationship.

Contractual incompleteness, and its interaction with the attributes of different types of transactional attributes including asset specificity, complexity, and uncertainty, plays a central role in the evaluation of the relative costs of governance through market-based bilateral contracts versus governance through internal organization. When transactions are mediated through market-based contracts, circumstances may arise where the buyer and seller have conflicting interests. Consider the situation where transacting parties are locked into a bilateral trading relationship, in the sense that the potential aggregate value of continuing the bilateral relationship is higher than terminating it and turning to alternative buyers or sellers. In this situation one or both parties may have the incentive and ability to behave “opportunistically” to serve their own interests. The resulting bargaining over the terms and conditions of trade will affect both the distribution of the rents associated with this particular bilateral trading relationship and potentially the efficiency (quantities and production cost distortions) of the trades that are consummated *ex post* as well (reducing the rents about which the parties can argue as well as the aggregate value of the trading relationship *ex ante* and *ex post*). The potential advantage of internal organization in this case is that internal organizations are likely to better harmonize these conflicting interests and provide for a smoother and less costly adaptation process under these circumstances, facilitating more efficient *ex ante* investment in the relationship and more efficient adaptation to changing supply and demand conditions over time. As Williamson (1971, pp. 116–117) observed many years ago:

...The contractual dilemma is this: On the one hand, it may be prohibitively costly, if not infeasible, to specify contractually the full range of contingencies and stipulate appropriate responses between stages. On the other hand, if the contract is seriously incomplete in these respects but, once the original negotiations are settled, the contracting parties are locked into a bilateral exchange, the divergent interests between the parties will predictably lead to individually opportunistic behavior and joint losses. The advantages of integration thus are not that technological (flow process) economies are unavailable to non-integrated firms, but that integration harmonizes interests (or reconciles differences, often by fiat) and permits an efficient (adaptive, sequential) decision process to be utilized. ...

Accordingly, there do not exist, except perhaps at very high cost, complete contingent contracts that can specify at the time a contractual relationship is being contemplated how each of the parties will perform under all possible contingencies

that could arise as the trading relationship proceeds over time. Contracts may be incomplete because of the direct costs of specifying and writing contracts that anticipate all contingencies, because of “bounded rationality” that makes it unlikely that the transacting parties can foresee all possible contingencies and/or because of high monitoring, verification, and enforcement costs.

Incomplete contracts *per se* do not necessarily lead to market inefficiencies. It is the interaction between contractual incompleteness and certain attributes of transactions that can lead the parties to a trading relationship to become “locked in” to the relationship once the relationship is consummated. This in turn can lead to adaptation problems that can adversely affect *ex ante* investment incentives and the *ex post* efficiency of the trading relationship. These potential problems are especially acute when supply and demand conditions that are uncertain *ex ante* change over time and the bargaining threat points of the parties to the relationship move outside of a “self-enforcing range” (Klein & Leffler, 1981) or “off the contract curve” (Williamson, 2000) anticipated in the design of the initial market governance arrangements or contracts.

In this regard, as this literature has developed, relationship-specific investments of various kinds, when they are required to support an efficient trading relationship, have come to play a central, though not exclusive, role in creating bilateral trading relationships that are susceptible to *ex post* bargaining and contractual performance problems. As I will discuss in more detail below, relationship-specific investments are investments which, once made, have a value in alternative uses that is less than the value in the use originally intended to support a specific trading relationship. Once specific investments have been made, a potential “hold-up” or “opportunism” situation is created if the parties can bargain over the appropriable *ex post* quasi-rents (the difference in asset values between the intended and next best use—Klein et al., 1978; Williamson, 1979, 1996) created by specific investments or must bargain or “haggle” to adapt to changing circumstances as the relationship proceeds over time. If contractual arrangements cannot be fashioned *ex ante* to mitigate these *ex post* incentives to bargain opportunistically without a full regard for the total surplus produced by the relationship, *ex ante* incentives to make specific investments in the first place will be adversely affected, and *ex post* performance and adaptation may be inefficient as well. To protect against these potential problems, transacting parties will explore the availability of alternative governance arrangements that reduce the costs of these contractual hazards, stimulating more efficient investment incentives *ex ante*, more efficient contract execution *ex post*, and, more generally, that reduce the overall cost of the relationship. Vertical integration is favored when the benefits of mitigating opportunism problems by moving the transactions inside the firm, by reducing *ex ante* investment and *ex post* performance inefficiencies, are greater than other sources of static and dynamic inefficiency associated with resource allocation within bureaucratic organizations.

18.3 Traditional Approaches to Explaining Vertical Integration

The explanations of the causes and consequences of vertical integration that emerged in the field of industrial organization during the post-World War II period were heavily influenced by the sharp distinction drawn by neoclassical economics

between resource allocation mediated through markets and resource allocation taking place within private firms and related types of hierarchical organizations (e.g., public enterprises). Microeconomics in general and applied price theory in particular were concerned with the way anonymous spot markets worked to allocate resources. The factors that determined the boundaries between firms and markets were largely ignored and issues associated with the internal organization of firms and the way firms allocated resources internally were, with a few exceptions (Simon, 1947; Cyert & March, 1963; Arrow, 1974), viewed as outside of the domain of economics. Firms were conceptualized as production sets that defined the technologically most efficient opportunities to transform inputs into outputs. They relied on anonymous spot markets to buy and sell inputs and outputs. That is, what firms did and what markets did were *complementary* activities. Coase's (1937, 1972) view that firms and markets were *substitute* governance mechanisms was not an accepted part of received wisdom until relatively recently. Precisely what was in a firm's production set and what was not was, at best, rather vague, and there existed no meaningful economic theory to explain where to draw the line between firms and market transactions or to explain the diverse types of "non-standard" contractual arrangements observed in the real world.

Industrial organization theorists like Bain (1956, 1959) viewed the relevant firm production set rather narrowly as encompassing activities that were clearly physically related to one another. Both multi-plant economies and vertical integration downstream and upstream were generally viewed as being unnecessary for a firm to produce at minimum cost in the absence of technological relationships that physically joined production between plants. Instead, the presumption was that vertical integration, and non-standard vertical contractual arrangements, reflected the presence of market power somewhere in the system and/or efforts to create or exploit market power. Vertical integration could be a profitable response to costs of successive monopolies (e.g., double marginalization and related "vertical externalities"; Tirole, 1988, Chap. 4), or it could facilitate price discrimination in a variety of different ways (Perry, 1978), or vertical integration (and long-term contracts) could be used strategically to soften competition in the short run by raising rivals' costs or in the long run by increasing the costs of entry to foreclose rivals that might otherwise enter the market (Aghion & Bolton, 1987; Ordover et al., 1990; Hart & Tirole, 1990).

Moreover, vertical integration itself was viewed theoretically as being "costless." That is, no internal organization costs were recognized, but only any costs realized through distortions in market prices, quantities, or the factor proportions used to produce output from a neoclassical production function. "Costless" vertical integration was also used as a benchmark against which alternative "costless" contractual arrangements could be compared. So, for example, distortions arising in one way or another as a consequence of double marginalization can be "solved" through vertical integration as well as with alternative (costless) contractual arrangements—two-part tariffs, maximum retail price maintenance, quantity forcing contracts, requirements contracts, service obligations, etc. (Tirole, 1988, Chap. 4). Absent transaction costs of some type, the alternative instruments are all equally attractive mechanisms for

responding to double marginalization. Price discrimination could be accomplished with vertical integration or with contracts that prohibited resale and eliminated the associated arbitrage that could otherwise undermine price discrimination. The potential advantages and disadvantages of alternative contractual responses were discussed primarily in the context of their effects on market prices and quantities and their ability to allocate risk more or less efficiently between parties with different degrees of risk aversion. There was nothing embedded directly in the neoclassical theoretical approach to vertical relationships that allowed for attributes of the contractual and organization arrangements themselves to provide a basis to choose among alternative governance arrangements because there were no transactions costs associated with any of the institutional alternatives identified.

Another potential source of incentives for vertical integration is the free rider problem associated with the provision of pre-sale information and post-sale service by competing downstream retailers (Telser, 1960; Mathewson & Winter, 1986). If retailers cannot fully appropriate for themselves the benefits of retail service expenditures but instead see some of the benefits accrue to their downstream competitors, this “horizontal externality” (Tirole, 1988, Chap. 4) will lead downstream retailers to underinvest in retail service (at least from the perspective of the manufacturer). Vertical integration is one potential solution to this problem. So too are various combinations of exclusive territorial agreements, resale price maintenance, profit pass-over contracts, and other mechanisms. The unanswered question is how to choose among the alternative institutional arrangements in a systematic way.

Dennis Carlton (1979) has shown how the combination of uncertain demand for inputs and the failure of markets to be cleared by spot prices under some contingencies can create a private incentive for downstream firms to integrate backward partially or fully for “supply security” reasons. (See also Malmgren, 1961; Arrow, 1975; Green, 1986; and Bolton & Whinston, 1993, for related theoretical work.) And there is abundant support in the business history literature for such a motivation for vertical integration (Chandler, 1964, p. 84). “The strong incentives for vertical integration arise because the vertically integrated firm is able to satisfy high probability demand by itself, and pass on the low probability demand to some other firm” (Carlton, 1979; p. 207). Williamson (1971, p. 117) points out that “...arguments favorable to vertical integration that turn on ‘supply reliability’ considerations commonly reduce to the contractual incompleteness issue (footnote omitted).” Moreover, it is not clear that the market imperfections that create the incentive to vertically integrate here could not be equally well (or even better) mitigated by downstream firms by arranging a portfolio of fixed price and spot market contracts.

Finally, George Stigler (1951) proposed a theory of vertical integration based upon Adam Smith’s famous theorem that “the division of labor is limited by the extent of the market.” Stigler advanced a dynamic or life-cycle theory of vertical integration. He argued that in an infant industry producing a new downstream product vertical integration would be more likely to occur because the demand for specialized inputs would be too small to support independent firms supplying intermediate goods. As the demand for the new product grows, intermediate good suppliers whose production is characterized by increasing returns would be spun off

as independent firms supplying inputs to multiple competing downstream suppliers. The empirical prediction is that as industries grow the extent of vertical integration should decline and as industries contract vertical integration should increase. Elberfeld (2002) argues that Stigler's theory is correct as long as there are no barriers to entry. Stigler's theory turns primarily on economies and diseconomies of scale and the implicit assumption that suppliers of new products require specialized inputs. It ignores transaction costs associated with both internal organization and market contracting. The theory has found little empirical support.

There is clearly no shortage of theories identifying potential incentives for vertical integration. This should not be surprising. As long as it is assumed that there are no additional costs associated with internal organization, almost any market imperfection necessarily becomes a candidate for creating private incentives for vertical integration. However, this approach ignores both the costs of internal organization and other costs of more complex contractual alternatives to either simple linear spot market contracts or vertical integration. In principle, the TCE framework should be able to encompass these traditional sources of market failure that have been identified as factors that increase incentives for vertical integration as well and to do so in a richer and more systematic fashion. However, the emphasis of TCE to date has been on looking for other than traditional vertical and horizontal externality, foreclosure, uncertainty, and risk allocation explanations for vertical integration and non-standard vertical restraints rather than trying to incorporate these considerations into the analysis. This is a deficiency that can be and should be remedied (Joskow, 1991, 2002).

18.4 Incomplete Contracts and Asset Specificity

As noted earlier, one of the foundations of TCE is the recognition that contracts are incomplete and potentially lead to contractual hazards that adversely affect ex ante investment incentives and the efficiency of ex post performance, especially in response to adaptations required to respond to changing supply and demand conditions. These problems arise when the parties to a contractual relationship find themselves in a bilateral bargaining position ex post as a consequence of lock-in as discussed earlier. While there are several potential sources of lock-in that lead to potential bilateral bargaining and adaptation problems, much of the theoretical and empirical work in the TCE tradition has focused on relationship-specific investments (asset specificity) and/or the interaction between asset specificity and other transactional attributes such as uncertainty, product complexity, and information asymmetries. Accordingly, I will focus the discussion here on the role of asset specificity.

Before proceeding, it is useful to outline an idealized set of steps that leads to a well-functioning vertical relationship to provide a framework for further analysis. The actual structural attributes of each of these steps and their implications for the costs of the feasible set of alternative governance arrangements then become the focus of comparative governance analysis. How would an idealized contractual

relationship be structured? First, the responsibilities and authorities of the parties to the transaction (including the appropriate assignment of property rights) would be defined *ex ante*. Second, contractual provisions would need to be agreed to *ex ante* that align the parties' incentives so that they have a mutual interest in performing in conformity with the intent of the contractual agreement. These contractual provisions include contractual formulas for adjusting prices and quantities as conditions change, cost, and profit sharing provisions, assignment of investment responsibilities, financial guarantees, and collateral requirements (Joskow, 1988b). Under normal conditions these provisions are incentive compatible and self-enforcing. Finally, the parties may agree to a process through which the terms of the contract can be adjusted (a renegotiation process) to facilitate smooth adaptation to changing circumstances that were not fully anticipated in the original terms of the contract. By facilitating realignment, the costs of haggling over the consequences of changed circumstances can be reduced. These governance design challenges are confronted in a broader context that includes, among other things, the effects of parties' behavior on the value of reputational capital, and the external financial constraints on their behavior that reputational considerations imply.

What departures from these idealized contractual attributes might emerge in practice and increase the cost of consummating the transactions covered by the contract? The assignment of responsibilities, authorities, or property rights may be incorrect or fail to cover all contingencies. The performance incentives may not work as intended under all contingencies and create *ex post* rent extraction opportunities and bargaining inefficiencies. The terms of the contract may be difficult to adapt to significant changes in economic circumstances leading to further rent extraction opportunities and performance inefficiencies. Imperfections in each step of the institutional design process can lead to distortions in *ex ante* investment incentives and *ex post* performance and adaptation inefficiencies. These imperfections in turn will vary based on the attributes of the transactions covered and the relative strengths and weaknesses of alternative governance arrangements in supporting a smooth trading relationship over time. As I will discuss in more detail below, the TCE approach takes into account all aspects of the trading relationship and compares their performance attributes under alternative governance arrangements, including internal organization. It not only emphasizes *ex post* adaptation problems but also recognizes potential inefficiencies in *ex ante* investment.

18.4.1 Asset Specificity in Detail

As noted earlier, a relationship-specific investment is an investment which once made (sunk) by one or both parties to an ongoing trading relationship has a lower value in alternative uses than it has in the intended use supporting this specific bilateral trading relationship. In the extreme, an investment made by a supplier in anticipation of supplying a product to a particular customer could be worthless if used to serve any other customer. More generally, we can think of specific investments as having significantly lower values or producing lower gains from trade,

when employed other than in supporting the intended relationship with a particular customer or supplier. Such investments create a bilateral dependency after they have been sunk defined by the difference between the value of the investment in its intended use and its next best alternative use. The parties to the transaction may then have an incentive to haggle over the distribution of the ex post quasi-rents created by the specific investments. It is the problem of economically protecting the aggregate value of the specific investments from being reduced through this potential haggling that drives the choice of governance structure.

Asset specificity that is directly relevant to vertical integration is thought to arise in a number of different contexts (Williamson, 1983, 1996, Chap. 4).¹

1. *Site specificity*: The buyer and the seller are in a “cheek-by-jowl” relationship with one another, reflecting ex ante decisions to minimize inventory and transportation expenses. Once sited the assets in question are highly immobile. A mine mouth coal plant (Joskow, 1985, 1987) or a bauxite processing plant and the associated mines (Stuke, 1983) are examples of site specificity.
2. *Physical asset specificity*: When one or both parties to the transaction make investments in equipment and machinery that involve design characteristics specific to the transaction which have lower values in alternative uses. A boiler that has been designed to burn a particular type of coal (Joskow, 1985) and investments in tools and dies to produce parts that can be used in a specific downstream manufacturer’s products (Klein et al., 1978; Klein, 1988) have this characteristic.
3. *Human asset specificity*: When, as a consequence of learning by doing, workers accumulate relationship-specific human capital that makes it possible for them to produce goods and services more efficiently than can otherwise equivalent workers who do not have this firm-specific human capital. Such human capital is of particular value to the suppliers and customers that benefit from it and is of lower value to the workers (or the firms they work for) if not utilized to support the specific relationship within which it accumulated. Design engineers who have developed special skills in designing a particular type of aircraft or automotive components are examples of human asset specificity (Monteverde & Teece, 1982; Masten et al., 1989).
4. *Dedicated assets*: General investment by a supplier that would not otherwise be made but for the prospect of selling a significant amount of product to a particular customer. If the relationship is terminated prematurely, it would leave the supplier with significant excess capacity, and a lower price to support the investment would be realized ex post than had been anticipated ex ante. The development of a large natural resource deposit in a remote location to supply a large upstream user is an example of dedicated assets (Joskow, 1985).

¹Masten et al. (1991) identify “temporal asset specificity” as a sixth category. Williamson (1996, p. 106) argues that this is a form of site specificity, and I agree with his assessment.

5. *Intangible assets*: Although specific investments are most frequently conceptualized as either physical investments or relationship-specific human capital, intangible capital such as brand name loyalty can have relationship-specific attributes. For example, McDonalds has a significant brand name value which has accumulated over time through investments in product quality, advertising, and promotion. The value of these investments is tied completely to the McDonalds brand name. In order to sell its products, however, McDonalds must convey the use of its valuable brand name to its distribution outlets, some of which it owns (vertical integration) and some of which are independent franchisees.

18.4.2 Effects on Ex Ante Investment and Ex Post Adaptation and Performance

The combination of incomplete contracts and relationship-specific investments can have adverse effects on both ex ante investment and the efficiency of ex post performance by creating a bilateral trading situation in which the parties' bargaining in their own individual self-interest leads to a reduction in the overall size of the pie. The property rights literature (Grossman & Hart, 1986; Hart & Moore, 1990; Hart, 1995) discussed below focuses on ex ante investment incentives and assumes that transactions are consummated efficiently ex post. The only inefficiencies that arise from improper alignment of property rights are reflected in distortions in ex ante investments, and these investment distortions are the primary focus of the analysis. Under these and other assumptions (Maskin & Tirole, 1999), the proper assignment of property rights or decision-making authority can mitigate hold-ups over the ex post division of appropriable quasi-rents and the associated adverse effects on ex ante investments. On the other hand, TCE is concerned with both ex ante and ex post inefficiencies that arise in bilateral trading relationships, though Williamson has generally placed more emphasis on ex post haggling and associated inefficiencies (Williamson, 1971, 1975, 2000; Klein, 2002) than on ex ante investment distortions. This emphasis, in turn, apparently reflects the view that internal organization is most likely to be superior to market contracting in circumstances where ex post bargaining/haggling costs associated with market contracting are high since what internal organization is good at harmonizing the otherwise conflicting interest of the parties to the transaction and facilitating a smooth and efficient adaptation to changing supply and demand conditions.

For ease of exposition, I focus first on a simple model in which contractual incompleteness leads to an ex post expropriation of quasi-rents but not to inefficiencies in ex post trade. That is, while there is bargaining over the appropriable ex post quasi-rents, and this affects ex ante investment incentives, given a particular level of ex ante investment, the efficient quantity is produced and traded ex post. As we shall see, contractual incompleteness plus asset specificity leads to underinvestment ex ante.

Before relationship-specific investments are made to support a trading relationship, let us assume buyers and sellers may have numerous potential suppliers and customers among whom they can choose to enter into an ongoing relationship (though there is no reason why there cannot be small numbers bargaining *ex ante* and “smaller numbers” bargaining *ex post* to incorporate pre-existing market power considerations as reflected in the traditional vertical externality and foreclosure literatures discussed above). That is, the markets are reasonably competitive *ex ante*. However, once a relationship is developed and relationship-specific investments are made to support it, a competitive bargaining situation *ex ante* is transformed into a small numbers or bilateral monopoly bargaining position *ex post*. The small numbers bargaining situation results from the investment in relationship-specific assets. Once these investments are made, the fact that they have lower values in alternative uses creates a stream of potentially appropriable quasi-rents (the difference between the value in the intended uses and in alternative uses). It is these quasi-rents over which buyers and sellers may haggle and bargain in the absence of a complete and easily enforceable contract that defines clearly the rights and obligations of each party when various contingencies arise. The potential for *ex post* bargaining over these quasi-rents in turn affects the expected returns from the initial investment in specific assets and the associated incentives to invest *ex ante*. Accordingly, the level of investment in relationship-specific assets will differ from what would maximize the aggregate gains from trade that potentially could be achieved absent opportunistic behavior. This increases the total costs of providing the goods and services being traded and is a social cost of opportunistic behavior.

The nature of the problem that arises when specific investments must be supported by incomplete contracts can be articulated more precisely with the help of a set of very simple analytical examples which are drawn from Tirole (1988, pp. 25–28). Assume that we have a buyer that is interested in acquiring one unit (for simplicity) of a product with particular characteristics. In order to meet the buyer’s requirements efficiently, a supplier must make investments that are specific to the product that the buyer desires. Once these investments are sunk, their value in producing products for other buyers is lower than it is when the investments are employed to produce output for the intended buyer. The buyer and its supplier agree to enter into a supply relationship in period 0, the supplier must make its investment in period 1, and production and trade take place in period 2. (In this sense we have a “long-term” relationship.)

Let I be the amount of the supplier’s investment and $C(I)$ the *ex post* costs of producing the product to this specific buyer’s satisfaction once the investment has been made, where $C'(I) < 0$ and $C''(I) > 0$. Assume initially that the value of the investment is zero if it is employed in alternative uses. The value of the output to the buyer is $v > C(0)$. Finally, assume that the parties do not enter into a contract prior to the supplier’s investment, but instead agree to bargain over the price when the product is finally delivered.

To solve for the transaction price in period 2 and the supplier’s expected profit maximizing investment in period 1, we must specify how the transaction price is

determined ex post. Assume that the buyer and seller negotiate a price that splits the ex post gains from trade evenly between them once the investment has been made (i.e., Nash bargaining) and that the product is (efficiently) supplied *given* the level of specific investment that is forthcoming ex ante. Then the ex post price is given by:

$$\begin{aligned} P(I) - C(I) &= v - P(I) \text{ or} \\ P(I) &= [C(I) + v]/2 \end{aligned} \quad (18.1)$$

and the supplier's profit is given by:

$$\max_I [P(I) - C(I) - I] = \max [v/2 - C(I)/2 - I] \quad (18.2)$$

Equation (18.2) shows that from the supplier's perspective a \$1 ex ante investment in cost reduction yields only a \$0.5 return to the supplier. The rest of the ex post surplus is captured or "held up" by the buyer. Accordingly, the privately optimal investment in cost reduction is lower than the socially optimal investment in cost reduction. Solving Eq. (18.2) yields the result that the supplier invests up to the point where:

$$-C'(I) = 2 \quad (18.3)$$

while the optimal investment is defined by the maximization of the total surplus $[v - C(I) - I]$, not just the supplier's slice of the surplus. The socially optimal investment is then given by:

$$\max [v - C(I) - I] \text{ or} \quad (18.4)$$

$$-C'(I) = 1 \quad (18.5)$$

Since $C(I)$ is convex, the supplier chooses a level of investment in cost reduction (Eq. (18.3)) that is too low compared to the level of investment that would maximize the gains from trade (Eq. (18.5)). The difference in production costs associated with these different investment levels is the cost of contractual hazards arising from the combination of the need for specific investments to support cost minimizing trade between the buyer and the seller and the absence of a contract that defines what the terms of trade will be ex post. The example's assumption that a specific investment has no value in alternative uses is, of course, extreme. However, the problem continues to emerge as long as the value of the specific investment in alternative uses is less than its value in the intended relationship (Tirole, 1988, p. 25.)

If the investments in specific assets could somehow be protected from this type of ex post hold-up problem, the supplier would be willing to increase investment in specific assets and reduce the overall cost of the trading relationship. The challenge in this case is to find a set of institutional arrangements that mitigate the hold-up problem which leads to underinvestment ex ante without incurring other costs that

more than exceed the increased surplus resulting from additional investment in cost reducing specific assets.

Of course, in our example, we assumed that there was no contract negotiated by the parties in period 0, before the specific investments were made, which defined the terms and conditions of trade in period 2. We have assumed the ultimate in incomplete contracts—no *ex ante* contract at all. The most obvious way to solve the problem that is identified in the example above is to allow the parties to enter into a binding contract prior to the time that the supplier makes its investment decisions. In particular, the contract could specify that the buyer would pay a price $p = P^*$ in period 2 where P^* is any price that lies between v and $C(I) + I$. The critical assumption is that P^* does not depend on I , but is fixed *ex ante*. For example, P^* could be arrived in period 0 through competitive bidding by suppliers competing to gain the right to supply the goods required by the buyer in period 2. If such a contract could be written and enforced costlessly, our problem would be solved. There would be no *ex post* bargaining opportunities, and the supplier would have precisely the correct investment incentives since its payoff is given by $(P^* - C(I) - I)$ whose maximum satisfies Eq. (18.5). This is the ultimate complete contract and eliminates the contractual hazard that was previously identified with no contract.

18.4.3 Ex Post Contractual Adaptation and Performance Problems

The departure of TCE is to recognize that this type of complete contract will be impossible to write in many interesting situations. For example, even in this simple case, let's say that the seller refuses to deliver the product unless a price higher than the agreed price is paid, perhaps reflecting the quasi-rents that could be expropriated if there were no contract at all. The buyer could go to court to enforce delivery at the contract price, but court enforcement can be costly, uncertain, and time-consuming. In addition, if it turns out that the supplier is unable to deliver in a timely way (specific performance), damages must be calculated which gets the courts involved in potentially complicated valuation deliberations with an uncertain outcome. These "haggling costs" both reduce *ex post* surplus and can affect its division among the parties, reducing both *ex ante* incentive to invest efficiently and the *ex post* value of the relationship given any particular level of investment.

The latter observations suggest that the assumption that *ex post* bargaining is efficient and, as a result, that the only inefficiencies arise with regard to *ex ante* investments is not a particularly realistic assumption. More generally with uncertainty and asymmetric information (Tirole, 1988, pp. 21–24; Myerson & Satterthwaite, 1983; Williamson, 1975, 2000) in the presence of bilateral monopoly *ex post*, bargaining will generally be inefficient. These inefficiencies in turn reduce the quasi-rents available to cover investment costs which further distorts *ex ante* investment incentives. The overall costs of the relationship increase due to the combination of *ex ante* and *ex post* inefficiencies. As already noted,

Williamson (2000) argues that it is with regard to the ex post bargaining and adaptation inefficiencies where the action lies in terms of the opportunities for internal organization to reduce transactions costs and become a lower cost alternative than even the best available market contracting alternative.

Of course, real-world transactions are often more complicated than what is captured in these simple examples, but the basic problems created by ex post lock-in and incomplete contracts are similar. For example, the nature of the technology for producing the goods or services at issue may lead to a situation where cost-minimizing exchange requires complementary relationship-specific investments by *both* the buyer and the seller (there may be other effective governance-related reasons, rather than technological reasons, for a contract to require both the buyer and the seller to make specific investments to mitigate incentives to behave opportunistically—Williamson, 1983). This not only changes the attributes of the ex post bargaining space but also complicates the effects of outside opportunities on ex ante investment decisions. When we introduce uncertainty about future production and investment costs, uncertainty about the buyer's ex post valuation and the quantities of the product required, the need for bilateral investments in specific assets to be made by both parties in order to support an efficient trading relationship, and a product quality dimension, we face a much more significant contracting and enforcement problem than in the simple model presented above. Correspondingly, it becomes more and more likely that it will be extremely costly or even impossible to write credible complete contracts that specify ex ante how the buyer and seller will behave when any contingency arises; or to design an associated enforcement mechanism that will require the performance promised or assess damages for non-performance without distorting behavior and increasing the total costs of the transactions at issue. Moreover, complex long-term contracts aimed at tying the hands of the parties so that they cannot behave opportunistically when foreseeable contingencies arise may also embody costly rigidities and have poor adaptive properties when contingencies not specifically provided for in the contract arise (Joskow, 1988a, b, 1990; Williamson, 1996, Chap. 4). Accordingly, while complex long-term contracts carry potential benefits by better protecting against the opportunistic behavior associated with specific investments than would simpler but more incomplete contracts, they also incur potential adaptation costs when unanticipated contingencies arise. Inefficiencies associated with ex ante investment distortions and ex post contract performance problems will increase, and internal organization will become a relatively more attractive governance structure.

On the other hand, there may be other constraints operating to affect ex post bargaining power that may mitigate the hold-up problem and the effects of opportunism on ex ante investments. In particular, the presence of reputational capital and the potential erosion of its value by opportunistic behavior may operate to mitigate such behavior (Williamson, 1975, Chap. 9; Klein & Leffler, 1981; Klein, 2002) and facilitate efficient relational contracting in the presence of specific investments and incomplete formal contracts.

18.5 Vertical Integration as an Alternative Governance Structure

These considerations help explain why we observe a wide array of contractual arrangements in the real world that sometimes look very different from the “standard” anonymous spot market transaction that is featured in elementary and intermediate microeconomics textbooks. However, these more complex contractual arrangements are unlikely to protect completely against the opportunistic behavior associated with specific investments and other sources of ex post lock-in and necessarily incur negotiating, monitoring, enforcement, and adaptation costs when changed circumstances push the threat points of the parties outside of the “self-enforcing range.” Vertical integration represents an alternative governance structure to bilateral contracts for mediating the supply of a product that requires specific investments to support cost-minimizing exchange. Rather than fiddling with contractual protections to mitigate the inherent conflicts of interest that may arise between independent buyers and sellers in the presence of specific investments, and dealing with other distortions and rigidities that such contracts may entail, the buyer may choose instead to integrate backward (or the seller to integrate forward) into the supply of the input at issue (or sale of the downstream good). By so doing the parties to the transaction would substitute internal hierarchical governance mechanisms within a firm for governance by market-based bilateral contracts. The vertically integrated firm’s managers (arguably) will pursue a common objective to maximize the firm’s value. They are not expected to engage in the kinds of opportunistic behavior associated with specific investments and an incomplete contract between an individual buyer and seller each pursuing its individual firm’s objectives.

If we were to apply the traditional methodology for analyzing incentives for vertical integration in response to market power and other standard market imperfections as discussed above (e.g., Tirole, 1988, Chap. 4), the vertically integrated firm would simply be modeled as a costless entity having the objective function specified in Eq. (18.4). That is, the vertically integrated firm would have the incentive to maximize the aggregate gains from trade associated with the transactions at issue without any offsetting costs of internal organization. The contractual hazard associated with the specific investment is removed because there is no longer an economic agent in the picture that has the incentive or ability to haggle over the distribution of the ex post quasi-rents. This approach assumes that both pieces of the ex post quasi-rent pie are internalized into the unified objective function of the vertically integrated firm that can execute the associated transactions costlessly and efficiently.

This analytical approach is not fully satisfactory. As Coase has observed (1937), this approach implies that vertical integration will *always* either have superior efficiency properties or equivalent efficiency properties to decentralized trade between independent buyers and sellers in a market. Why then doesn’t all economic activity take place within vertically integrated hierarchies with a single objective function? The problem is that this basic approach to analyzing vertical integration

doesn't tell us what it is about internal organization that makes it a superior governance structure to imperfect market transactions in some circumstances but not in others. The attributes and associated costs of allocating resources within internal organization are missing.

An important difference between internal organization and market contracting is the nature of the delegation of authority to make decisions when contingencies arise which could not otherwise be contracted on effectively through bilateral contracts. The property rights approach focuses on ownership of physical and intangible assets (but not human capital that accrues to individual workers) where ownership carries with it the authority to determine how these assets will be used (Grossman & Hart, 1986; Hart & Moore, 1990; Hart, 1995). Ownership of specific investments (e.g., through vertical integration) gives the owner the residual authority to use the assets to further the owner's objective function. While negotiations between managers within a firm may arise, the firm establishes clear lines of authority to resolve them. Ownership and the rights of control that go along with it change the status quo bargaining point within the firm and the ultimate allocation of the rents over which the bargaining takes place. That is, when specific investments are involved, ownership of the specific assets allocates the residual rights of control to the party that makes the specific investment. The owner then has the authority over the ex post trading decision and any internal transfer prices that may be relevant. Hart (1995) shows how various combinations of specific physical and human capital can affect the allocation of resources under alternative ownership arrangements. The residual rights of control that are conveyed by ownership affect the ex post distribution of surplus which in turn affects the ex ante incentives to invest.

The property rights approach strips the firm of most of its organizational features and focuses on how ownership and the associated residual rights of control affect the bargaining power of otherwise self-interested economic agents engaged in bilateral trade. This approach does not allow for any other changes in incentives and behavior of the transacting parties when the relationship is brought from the market inside of the firm (vertical integration). Thus, it largely ignores important differences between market transactions and internal organization other than simply a change in relative bargaining power between self-interested managers (Williamson, 1996, Chap. 4). However, the objective functions possessed by managers and the incentive and payoff structure that they face are different for managers within a firm as compared to managers in separate firms. One of the key tasks of management is to develop monitoring and financial incentive arrangements within the firm that induce the managers and employees to pursue the interests of the firm rather than the interests of a hypothetically independent division of the firm producing for its own account (Williamson, 1985, Chap. 6; Holmstrom & Milgrom, 1990; Williamson et al., 1975). These incentive arrangements include compensation contracts that partially tie compensation to overall firm performance and the effects of employee behavior on promotion opportunities and continued employment. In short, other things equal, the incentive and ability of a manager within a firm to exploit specific investments to hold up another division is different from what it would be if the managers were managing two independent firms.

Monitoring behavior and the costs and distribution of information are also likely to be different within a firm than between independent firms (Williamson, 1975; Arrow, 1975; Hart, 1995, p. 72). Employees within a firm have different incentives and obligations to reveal information to senior management from those of employees of an independent firm and can be subject to swifter and different penalties for hiding information (e.g., termination). Moreover, senior management has authority to use a variety of monitoring and information gathering mechanisms that can be matched quickly to problems as they arise without adhering to formal (incomplete) auditing contracts. The internal auditing departments of large firms have substantial authority to range far and wide in identifying behavior that is inconsistent with the firm's objectives. Accordingly, internal organization is likely to be better at obtaining those types of information that are directly relevant to monitoring and controlling the opportunistic behavior by the firm's managers that would otherwise arise from the combination of asset specificity and incomplete contracts if the transactions took place between independent firms.

Internal organization can also rely on more informal and less time-consuming procedures to resolve conflicts inside the firm than would independent agents bound by formal contracts. As Williamson (1996, Chap. 4) observes, the internal "contract law" within a firm is quite different from the arbitration and litigation procedures to which independent economic agents would have to turn if they could not resolve disputes. The latter can be a costly and time-consuming process that typically involves a third-party decision-maker who must become informed about the issues *de novo*. The internal decision-maker, whether the CEO or the relevant division manager, can utilize simpler and faster internal procedures for resolving conflicts between divisions and also is likely to come to the problem with much more information of relevance than would a third-party arbitrator.

If hierarchical organizations have these attractive properties, why don't we see more economic activity taking place within very large organizations rather than through markets? The answer is that internal organization is good at some things, but not at others. Williamson (1996, Chap. 4) observes that when we look at the bigger dynamic picture, internal organization is a last resort that we turn to only in the presence of significant contracting hazards and associated transactions costs. This is because, opportunistic behavior associated with specific investments aside, decentralized market arrangements have superior adaptive properties to internal organization in many *other* important dimensions. Differences in the information structure between market and hierarchical governance structures which help mitigate opportunism problems associated with specific investments may lead to inefficiencies in other dimensions (Hart, 1995, pp. 71–72). For example, employees may be less willing to reveal information that adversely affects their promotion possibilities or continuing employment. The kinds of low-powered incentives that characterize internal compensation arrangements may also mute incentives to exert the optimal amount of worker effort (Williamson, 1985, Chap. 6; Holmstrom & Milgrom, 1990). In addition, while internal organization is likely to be better at removing certain kinds of internal information asymmetries in the short run, it may be an inferior structure for obtaining, processing, and using external information about prices,

costs, quality, and technological change in the long run compared to repeated market transactions. For example, when a firm vertically integrates (or enters into a very long-term full requirements contract), it is likely to lose some of the benefits associated with continually examining and accessing outside opportunities through repeated contracting. These opportunities include information about the “least cost” prices of the goods and services that the firm is producing internally and the availability of new technologies and production methods. While there is nothing that prohibits a vertically integrated firm continuing to look to outside opportunities to benchmark its performance, an internal division in competition with outside sources may have strong incentives to hide or misrepresent outside opportunities in order to protect itself from external competition. This type of organizational opportunism is different from the kinds of hold-up problems created by specific investments but may be even more costly in the long run.

For these reasons, even in the face of significant contractual hazards resulting from specific investments and incomplete contracts, firms may still find it advantageous to continue to rely on arms-length market transactions for all or a fraction of their input or distribution requirements (dual sourcing) involving specific investments rather than turning to complete vertical integration. This choice may be made to provide management with external information that it can use to assess the performance of its internal divisions and to counteract the costs of various types of internal organizational inefficiencies. Competitive market prices convey a tremendous amount of information that is difficult to reproduce using internal accounting cost and auditing information. Moreover, this information is updated very quickly as supply and demand conditions change if a firm is in the market repeatedly. As organizations get larger, the volume of auditing information that must be processed by management grows non-linearly with the size and scope of the firm (Williamson, 1975) and becomes more difficult to use to control costs and quality effectively and to adapt to changing market conditions. The potential shirking problems resulting from low-power internal compensation incentives are also likely to become more significant as monitoring becomes more difficult in large organizations.

There are other dynamic considerations that may make the relative attractiveness of alternative governance arrangements in a particular industry or transactional setting change over time (Langlois & Robertson, 1989). As revealed by the extensive analysis of the GM-Fisher Body relationship (Klein et al., 1978; Klein, 1988, 2000, 2002), a detailed long-term contract involving a transaction with significant relationship-specific investment may work satisfactorily for some period of time. However, when external circumstances change, the existing contractual arrangements can lead to significant adaptation problems and associated performance inefficiencies. These adaptation problems are less likely to have emerged if the production of car bodies had been governed through internal organization rather than a rigid contract that did not anticipate a big increase in demand and the conflicts that emerged over plant location decisions. Changes in technology or government regulations may also change the relative attractiveness of alternative governance arrangements. For example, with specific reference to my work on coal contracts, changes in environmental laws in the United States have independently made it

attractive to invest in fuel flexibility capabilities in coal-burning power plants. This flexibility in turn makes coal users less dependent on specific coal suppliers or coal supply locations, reducing the value of long-term contracts. Accordingly, one would anticipate seeing a shift to shorter term more flexible contracts as the *net* costs of fuel flexibility to mitigate potential hold-up problems declines as a consequence of environmental regulations that increase the value of such flexibility.

The bottom line is that there are benefits and costs of internal organization. Market transactions incur transactions costs associated with writing and enforcing contingent contracts and the inefficiencies *ex ante* and *ex post* resulting from opportunistic behavior that exploits specific investments. Internal bureaucratic allocation mechanisms can help mitigate these types of transactions costs but incur other types of transactions or organization costs. The costs of internal organization are associated with the relatively inferior adaptive properties of bureaucratic hierarchies to rapidly changing outside opportunities over the longer term and the difficulty of designing compensation mechanisms to give managers and employees appropriate incentives to control costs and product quality. No governance structure is free from at least some transaction costs. The decision whether or not to vertically integrate then becomes a tradeoff between the costs of alternative governance arrangements. Governance arrangements are selected which represent the best that can be accomplished from a set of imperfect governance alternatives. Understanding the tradeoffs between governance alternatives and how they are affected by the attributes of products, production processes, and inputs, legal, political, and regulatory institutions is what the comparative governance approach is all about.

Despite these observations, however, I think that it is fair to say that the TCE literature on vertical integration, especially the empirical literature, has focused much more on the inefficiencies of market transactions than it has on the strengths and weaknesses of internal organization. Indeed, this may be one of the reasons why Gibbons (2003) argues that there is a lot of confusion about the similarities and differences between the TCE approach attributed to Williamson, Klein, and others and the property rights or rights of control approach attributed to Grossman and Hart (1986) and Hart and Moore (1990). TCE emphasizes (verbally) *ex post* adaptation issues and the associated bargaining and performance costs, recognizing that these costs also affect *ex ante* investment incentives. The property rights literature assumes that *ex post* bargaining is efficient and emphasizes the effects of *ex post* rent expropriation on *ex ante* investment. However, both literatures have emphasized specific investments, and, as we shall see, much of the empirical literature relates variations in the costs of governance structure to variables measuring asset specificity of various kinds, rather than on direct measures of *ex post* adaptation costs (which could be argued would be off the equilibrium path anyway), good proxies for their expected magnitude, or variables measuring variations in the costs of internal organization. The full implementation of a comparative governance framework requires that the costs of alternative market governance arrangements and the costs of internal organizations with different attributes be given equal treatment. The situation has not changed all that much since 1971 when Williamson (1971, p. 113) observed that:

A complete treatment of vertical integration requires that the limits as well as the powers of internal organization be assessed. As the frictions associated with the powers of administrative coordination become progressively more severe, recourse to market exchange becomes more attractive, *ceteris paribus* ... it is simply asserted [in this essay] that, mainly on account of bounded rationality and greater confidence in the objectivity of market exchange in comparison with bureaucratic process market intermediation is generally to be preferred over internal supply in circumstances in which markets may be said to 'work well.' (footnote omitted)

The properties of the firm that commend internal organization as a market substitute... fall into three categories: incentives, controls, and what may be referred to broadly as 'inherent structural advantages.' In an incentive sense, internal organization attenuates the aggressive advocacy that epitomizes arm's length bargaining. Interests, if not perfectly harmonized, are at least free of representations of narrowly opportunistic sort. ... In circumstances... where protracted bargaining between independent parties to a transaction can be reasonably anticipated, internalization becomes attractive. (footnote omitted)

...when conflicts develop, the firm possesses a comparatively efficient conflict resolution machinery... fiat is frequently a more efficient way to settle minor conflicts (say differences in interpretation) than is haggling or litigations.

18.6 Empirical Evidence

The choice of governance structure and how this choice is affected by the kinds of transaction cost considerations that have been discussed here have attracted considerable empirical study. This empirical work has focused on decisions to vertically integrate the design of non-standard contractual arrangements and the performance of both vertical integration and non-standard contractual arrangements over time as supply and demand conditions change. This work has included both detailed case studies of particular firms or types of contractual and organizational arrangements as well as econometric analyses based on large numbers of observations on the governance arrangements chosen for transactions with different attributes. Interestingly, the TCE framework has stimulated much more empirical work than either the traditional theories of vertical integration outlined above or the more recent property rights literature. This is to the credit of the scholars who have done theoretical work in the TCE tradition since they have produced testable hypotheses and endeavored to provide guidance to empirical researchers regarding how to measure relevant attributes of transactions affecting market contracting and internal organization. It also reflects the continuing interaction between theory, empirical analysis, and public policy that has characterized developments in TCE over the last 25 years, a productive interaction that is largely absent from many other areas of industrial organization. Moreover, in the case of TCE-related research, the empirical results are much more supportive of the relevant theory than is the case with the other theories of vertical integration (see, e.g., Waterman & Weiss, 1996; Chippy, 2001; Mason & Phullips, 2000). Whinston (2003) argues that the empirical work stimulated by TCE does not do a good job distinguishing between TCE-based theories and property rights-based theories of vertical integration and provides suggestions for how the

predictions of property rights theories might be tested and its distinct predictions, including those that are contrary to the implications of TCE-based theories, tested empirically. Empirical work to date has not focused on trying to distinguish between TCE and property rights theories of vertical integration, and there has been little effort to test property rights theories directly. This is probably the case because those who have led the development of property rights theories of vertical integration have made little effort to specify clear testable hypotheses or to relate them to variables that are likely to be measurable and available for empirical analysis. That is, they have not provided adequate guidance to empirical researchers, and, as a result, the similarities and differences between TCE-based theories and property rights theories have been of much more interest to theorists than to empiricists. Perhaps the guidance provided by Whinston will lead empirical researchers to focus more attention on property rights theories.

There have been at least 500 papers published that have examined various aspects of comparative institutional choice from a TCE perspective. A significant fraction of these studies have examined the vertical integration or “make or buy” decision. There have also been several survey articles that have reviewed the empirical literature stimulated by TCE theories, including many related to vertical integration and non-standard vertical contracting arrangements (Joskow, 1988a, b; Shelanski & Klein, 1995; Crocker & Masten, 1996; Coeurderoy & Quélin, 1997; Vannoni, 2002). In addition, Peter Klein has written a chapter that reviews the empirical work on the “make or buy” decision (see Chap. 17 of the 2005 edition of this Handbook). Accordingly, my discussion here will be relatively brief and focus on methodological issues rather than providing a complete catalogue of empirical studies and their results. Masten et al. (1991) present a very useful empirical model that captures the essence of the comparative governance approach, the associated TCE predictions regarding the choice between governing vertical relationships with market contracts or through vertical integration (internal organization), and the issues raised for empirical analysis. I will use that model here since it helps organize alternative empirical approaches and identify important empirical issues. While it is most relevant to econometric studies of the choice between “market” and “organization,” it also provides useful guidance for related case study research.

Masten, Mehan, and Snyder focus on a model of choice between market contracting (*m*) and internal organization—vertical integration—(*o*). Following the comparative governance arrangement approach, they define the costs of the two alternative governance arrangements as:

C_o = cost of organizing transactions inside a firm (e.g. VI)

C_m = cost of organizing transactions through a (least cost) market contracting mechanism

Then the choice between market contracting and vertical integration depends on the relative costs of the two alternative governance arrangements. That is, the theory implies that transacting parties will choose internal organization if $C_o < C_m$ and vice versa. The total costs of transacting associated with the alternative governance arrangements then depend on those attributes of the transactions that

affect the costs of market contracting (measured by the elements of a vector Z) and the costs of vertical integration (measured by the elements of a vector X). X and Z may or may not include common elements. Accordingly, the costs of vertical integration or internal organization (o) and of market contracting (m) can be modeled as:

$$C_o = \alpha X + e \quad (18.6)$$

$$C_m = \beta Z + u \quad (18.7)$$

where α and β are the coefficients that measure the marginal governance cost associated with each relevant transactional attribute for internal organization and market governance structures, respectively, and e and u are random disturbance terms which may or may not be correlated with one another. It follows immediately that the probability of choosing vertical integration depends on the probability that the costs of internal organization are less than the costs of market transactions:

$$\begin{aligned} \text{Probability of choosing internal organization} &= \Pr(C_o < C_m) \\ &= \Pr(e - u < \beta Z - \alpha X) \end{aligned} \quad (18.8)$$

Unfortunately, it is rarely possible to measure the costs of internal organization (C_o) and the costs of market contracting (C_m) directly. As a result, we cannot estimate the parameters of the structural model (18.1) and (18.2) directly. Nor do we typically have good cardinal measures of the transactional attributes that enter X and Z . Instead, as we shall see, researchers frequently must rely on ordinal proxy variables to measure variations in the elements of X and Z . So, empirical studies often rely on observations indicating whether or not a relationship is governed by internal organization or market contract, creating a 0,1 limited dependent variable, and various proxies for variations in transaction related variables that are elements of X and/or Z , such as asset specificity, complexity, uncertainty, and frequency of transactions or repeated interaction. Hypotheses regarding the organization form can then be based on the estimated coefficients of α and β . These coefficients can be estimated using limited dependent variable techniques like probit or logit (though only up to a proportionality factor without independent information about the variance of $(e - u)$). Moreover, if X and Z share common variables (e.g., asset specificity), then one can test whether $(\beta_k - \alpha_k) > 0$, but not whether both coefficients are non-negative, which weakens the power of the associated hypothesis tests. For example, if one believed that when cost-minimizing exchange requires more resources devoted to specific investments the costs of both market contracts *and* internal organization increased, we would only be able to identify the relative costs (or net costs) of one governance arrangement compared to the other. If it were the case that asset specificity has little effect on market transactions but reduced the costs of internal organization, then α_k would be negative and $(\beta_k - \alpha_k) > 0$ even though asset specificity has no effect on the costs of transacting through bilateral contracts. These possibilities reduce the power of the reduced form approach to hypothesis

testing, though how important this is depends on whether there is good reason to believe that asset specificity affects internal organization costs in any systematic way. I have not been convinced that there is. If X and Z do not share common elements (e.g., variations in asset specificity affect the cost of market transactions but do not affect the costs of internal organization directly), we can identify the signs of the relevant coefficients independently. In many applications, the implicit assumption is that X and Z are orthogonal (e.g., variations in asset specificity affect the costs of market organization but not the costs of internal organization), the focus is on the measurement of the elements of Z (i.e., when $\beta_k > 0$, then $\alpha_k = 0$ and vice versa) and on the signs and magnitudes the estimated coefficients of the “market failure” variables included in Z . If the assumption that X and Z are orthogonal is not correct, not only can we measure only the sign and magnitude of $(\beta_k - \alpha_k)$, but if there are left-out variables in X that are correlated with the included elements of Z , it may not even be possible to get unbiased estimated of $(\beta_k - \alpha_k)$.

Masten, Meehan, and Snyder go on to show that *if* we can measure C_o or C_m directly, we can measure the individual structural coefficients (not just their difference) *and* the cost of the other organizational form using switching regression techniques (if X and Z are orthogonal *or* e and u are uncorrelated). They then apply this technique to measure the costs of governing the production of several components used in naval shipbuilding with varying transactional attributes using either vertical integration or contracting with third parties. I am familiar with only one other study that measures the costs of alternative governance arrangements directly (Kerkvliet, 1991).

With this basic empirical model in mind, let us now turn to the methods used in the empirical literature that seeks to test whether variations in transaction attributes such as asset specificity affect the choice between vertical integration and market contracting as TCE theory predicts. These studies tend to follow a similar empirical methodology. They generally focus on a particular good or service that is used as an input to produce or distribute a specific class of products: automobile components (Klein et al., 1978; Klein, 2000, 2002; Monteverde & Teece, 1982; Walker & Weber, 1984; Langlois & Robertson, 1989), coal (Joskow, 1985, 1987, 1988b, 1990; Kerkvliet, 1991), aerospace systems (Masten, 1984), aluminum (Stukey, 1983), chemicals (Lieberman, 1991), forestry (Globerman & Schwindt, 1986), carbonated beverages (Muris et al., 1992), pulp and paper (Ohanian, 1994), and property-liability insurance (Regan, 1997). Other studies focus on a set of products that can be distributed through a similar set of alternative distribution modes (Anderson & Schmittlein, 1984). The sale of these goods and services is mediated by several different governance structures (e.g., vertical integration, franchise agreements, long-term contracts, spot market sales), and the governance choices are observable.

The empirical challenge is then to determine whether the incidence of the alternative governance structures observed in practice can be explained by variations in the transactional characteristics of the goods and services whose governance structures are being investigated, in particular by the importance of asset specificity, holding other transactional attributes constant (or assuming that any associated missing variables are uncorrelated with the measures of asset specificity), that is,

to measure the coefficients β and α . In light of the empirical model discussed above, this challenge is more daunting than might first meet the eye given the limitations on measuring all of the dependent and independent variables that we would measure under ideal conditions.

To respond to this empirical challenge in the absence of direct measures of the costs of transacting under different governance arrangements, empirical work typically relies instead on information about the actual utilization of alternative governance structures to mediate specific groups of transactions within the (narrow) class of products being studied. For example, in my work on coal supply arrangements for electric utilities, I identified specific generating stations that relied on coal supplies from affiliated mines as well as generating stations that contracted for coal with unaffiliated suppliers. (My research program involved an examination of the choice between vertical integration and market contracting (1985), the duration of contracts for the bulk of the coal supply relationships that did not involve vertical integration (Joskow, 1987), the adjustment features built into the contracts to allow prices and quantities to adapt over time (Joskow, 1998), and the ex post performance of the contractual arrangements, including the use of the courts to enforce contractual commitments (Joskow, 1990). As another example, Monteverde and Teece (1982) identify the proportions of various automobile components utilized by US automobile manufacturers which are produced internally compared to the proportion procured from third parties. In both sets of studies, there is a substantial variation across the sample in the reliance on vertical integration. Many other studies take a similar approach.

The next step is to develop measures of the exogenous characteristics of the underlying transactions, with particular attention devoted to measuring the importance of specific investments and other variables that may interact with asset specificity to affect the costs of market contracting and the incidence of opportunistic behavior that could make market contracting more costly in terms of ex ante investment distortions and ex post performance inefficiencies. Measuring variations in the importance of specific investments to support cost-minimizing exchange is difficult. In a number of cases, survey data have been used to characterize the importance of specific investments in supporting different groups of transactions within the set that is being studied (Monteverde & Teece, 1982; Anderson & Schmittlein, 1984). In other cases, efforts are made to develop ordinal characterizations of the different types of asset specificity associated with specific sets of transactions within the groups (e.g., Joskow, 1985, focuses on mine-mouth plants as classical examples of site specificity).

Ideally, it would be desirable to identify attributes of the firms in the sample that are expected to affect the costs of internal organization as well. Then the comparative costs of alternative governance arrangements could be captured directly. As a practical matter, most of these studies rely on samples of firms that are reasonably assumed to have identical "internal organizational" attributes. The implicit assumption is that the variables that affect the costs of market contracting are orthogonal to the variables that affect the costs of internal organization. For example, in my work on coal contracts, the firms involved all produced electricity (a homogeneous

product), were regulated monopolies, and were subject to similar types of economic regulation. I had no reason to believe that variations in the importance of asset specificity affected the costs of internal organization significantly, while they were expected to affect the costs of market contracting significantly. Monteverde and Teece (1982) look at transactions involving a small number of automobile firms with widely varying demands for components. They implicitly assumed that some variables (e.g., firm size) affected the costs of internal organization and others affected the costs of market contracting. That is, the implicit assumption again is that X and Z are orthogonal. If this assumption is not correct, it reduces the power of the tests of the TCE theory and may lead to biased estimates of the coefficients being estimated.

With these data in hand, and recognizing the potential problems associated with the measurement issues that I have discussed, the analysis then proceeds to determine whether the variations in the governance structures observed (vertical integration or bilateral contracting, extent of vertical integration, extent of contractual pre-commitment) can be associated with variations in the importance of specific investments and other characteristics of transactions that the theory suggests will lead to opportunistic behavior and related contractual hazards. As noted earlier, in econometric application, the dependent variable is typically a limited dependent variable that indicates whether vertical integration is utilized ($y = 1$) or not ($y = 0$) or the intensity of vertical integration ($y = \% \text{ of purchases by a buyer from internal sources}$). For the reasons discussed above, and nicely illuminated by Masten et al. (1991), the results and power of the tests must be interpreted with great care.

In the end, we become convinced about the validity of a theory based on the accumulation of evidence from many studies examining a wide range of transactions and governance arrangements. The accumulation of results provides the most important power of the hypothesis tests. The empirical studies of vertical integration and how the choice of this governance structure is influenced by the importance of specific investment and other variables that could lead to ex ante and ex post contractual inefficiencies overwhelmingly show that the importance of specific investments is both a statistically and economically important causal factor influencing the decision to vertically integrate. Indeed, it is hard to find many other areas in industrial organization where there is such an abundance of empirical work supporting a theory of firm or market structure. And it is the combination of compelling theoretical analysis combined with a large body of supporting evidence that makes the TCE approach to understand vertical integration and alternative vertical governance arrangements so important.

18.7 Conclusions

Let me conclude where I began. There is no single unified theory of vertical integration that exists today or is likely to exist in the future. There are many types of market imperfections that could lead transacting parties to turn to vertical integration as an alternative governance arrangement, recognizing that vertical integration is one of many governance alternatives to relying on anonymous spot market

contracting. However, the NIE/TCE approach provides a framework that can encompass and enrich all leading theories of vertical integration. It does so by taking a comparative governance approach, recognizing that there are unique but systematic costs associated with alternative market contracting structures, with vertical integration and with various hybrid forms, and by building on a synergistic relationship between theoretical and empirical analyses.

This being said, there is still much to learn about vertical integration, alternative market contracting structures, and various hybrid forms. In my view, we have made more progress in understanding and measuring the hazards and associated costs of market contracting in the presence of alternative transactional attributes than we have about the costs of internal organization and how these costs are affected by different internal organizational and incentive structures. This observation is especially relevant to the state of empirical analysis where measurement issues remain a serious challenge. In addition, I believe that it would be very productive to focus more attention on the dynamic properties of both contractual relationships and internal organization. The “make or buy” decision is not a once-and-for-all decision. Firms may choose to vertically integrate and then decide that it is less costly to rely on market contracting. As in the case of GM-Fisher Body, firms may choose to govern a relationship by contract and then decide to take production in house. Better understanding how idiosyncratic contractual relationships adapt to changing supply and demand conditions over time, how organizations respond to changing circumstances and ageing over time, and why governance arrangements change over time will provide deeper insights into both the market imperfection and organizational imperfection considerations that affect the choice of governance arrangements.

Acknowledgments I have benefited enormously from discussions with my colleague Bob Gibbons and from comments from Claude Ménard and two anonymous referees. This chapter was previously published as “Vertical Integration” in the *Handbook of New Institutional Economics*, (Ménard and Shirley editors), 2005, Springer.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Aghion, P., & Bolton, P. (1987). Contracts as a barrier to entry. *American Economic Review*, 77, 388–401.
- Anderson, E., & Schmittlein, D. (1984). Integration of the sales force: An empirical examination. *RAND Journal of Economics*, 15, 385–395.
- Arrow, K. J. (1974). *The limits of organization*. W.W. Norton.
- Arrow, K. J. (1975). Vertical integration and communication. *The Bell Journal of Economics*, 6, 173–183.
- Bain, J. (1956). *Barriers to new competition*. Harvard University Press.
- Bain, J. (1959). *Industrial organization*. Wiley.
- Bolton, P., & Whinston, M. (1993). Incomplete contracts, vertical integration, and supply assurance. *Review of Economic Studies*, 60, 121–148.

- Carlton, D. (1979). Vertical integration in competitive markets under uncertainty. *Journal of Industrial Economics*, 27, 189–209.
- Chandler, A. (1964). *Strategy and structure: Chapters in the history of American industrial experience*. MIT Press.
- Chipty, T. (2001). Vertical integration, market foreclosure, and consumer welfare in the cable television industry. *American Economic Review*, 91, 428–453.
- Coase, R. (1937). The nature of the firm. *Economica*, 4, 386–405.
- Coase, R. (1972). Industrial organization: A proposal for research. In V. R. Fuchs (Ed.), *Policy issues and research opportunities in industrial organization* (pp. 59–73). National Bureau of Economics Research.
- Coeurderoy, R., & Quélin, B. (1997). Transaction cost theory: A survey on empirical studies on vertical integration. *Revue Economie Politique*, 107, 146–181. [in French].
- Crocker, K., & Masten, S. (1996). Regulation and administered contracts revisited: Lessons from transaction-cost economics for public utility regulation. *Journal of Regulatory Economics*, 9, 5–40.
- Cyert, R., & March, J. (1963). *A behavioral theory of the firm*. Prentice Hall.
- Elberfeld, W. (2002). Market size and vertical integration: Stigler's hypothesis reconsidered. *Journal of Industrial Economics*, 50, 23–43.
- Gibbons, R. (2003). "Four formal(izable) theories of the firm?" preliminary and incomplete draft, August 6, 2003.
- Globerman, S., & Schwindt, R. (1986). The organization of vertically integrated transactions in the Canadian forest products industry. *Journal of Economic Behavior and Organization*, 7, 199–212.
- Green, J. (1986). Vertical integration and assurance of markets. In F. G. Mathewson & J. E. Stiglitz (Eds.), *New developments in the analysis of market structure*. MIT Press.
- Grossman, S., & Hart, O. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, 94, 691–719.
- Hart, O. (1995). *Firms contracts and financial structure*. Clarendon Press.
- Hart, O., & Moore, J. (1990). Property rights and the nature of the firm. *Journal of Political Economy*, 98, 1119–1158.
- Hart, O., & Tirole, J. (1990). Vertical integration and market foreclosure. *Brookings Papers on Economic Activity*, 21(special issue), 205–276.
- Holmstrom, B., & Milgrom, P. (1990). Regulating trade among agents. *Journal of Theoretical and Institutional Economics*, 146, 85–105.
- Joskow, P. L. (1985). Vertical integration and long term contracts: The case of coal-burning electric generating stations. *Journal of Law, Economics and Organization*, 1, 33–80.
- Joskow, P. L. (1987). Contract duration and relationship specific investments. *American Economic Review*, 77, 168–175.
- Joskow, P. L. (1988a). Asset specificity and the structure of vertical relationships: Empirical evidence. *Journal of Law, Economics and Organization*, 4, 95–117.
- Joskow, P. L. (1988b). Price adjustment in long term contracts: The case of coal. *Journal of Law and Economics*, 31, 47–83.
- Joskow, P. L. (1990). Price adjustment in long term contracts: Further evidence from coal markets. *RAND Journal of Economics*, 21, 251–274.
- Joskow, P. L. (1991). The role of transactions cost economics in antitrust and public utility regulatory policies. *Journal of Law, Economics and Organization*, 7, 497–540.
- Joskow, P. L. (1998). Asset specificity and vertical integration. In P. Neuman (Ed.), *The new Palgrave dictionary of economics and law*. Macmillan.
- Joskow, P. L. (2002). Transaction cost economics, antitrust rules and remedies. *Journal of Law, Economics and Organization*, 18, 95–116.
- Kerkvliet, J. (1991). Efficiency and vertical integration: The case of mine-mouth electric generating plants. *The Journal of Industrial Economics*, 39, 467–482.
- Klein, B. (1988). Vertical integration as organized ownership: The fisher body-general motors relationship revisited. *Journal of Law, Economics and Organization*, 4, 199–213.

- Klein, B. (2000). Fisher-general motors and the nature of the firm. *Journal of Law and Economics*, 43, 105–141.
- Klein, B. (2002). Fisher body-general motors once again: How do we know when a holdup occurs? Mimeo.
- Klein, B., & Leffler, K. (1981). The role of market forces in assuring contractual performance. *Journal of Political Economy*, 89, 615–641.
- Klein, B., Crawford, R., & Alchian, A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21, 297–326.
- Langlois, R., & Robertson, P. L. (1989). Explaining vertical integration: Lessons from the American automobile industry. *The Journal of Economic History*, 69, 361–375.
- Lieberman, M. (1991). Determinants of vertical integration: An empirical test. *Journal of Industrial Economics*, 39, 451–466.
- Malmgren, H. B. (1961). Information, expectations and the theory of the firm. *The Quarterly Journal of Economics*, 75, 399–421.
- Maskin, E., & Tirole, J. (1999). Two remarks on the property-rights literature. *Review of Economic Studies*, 66, 139–149.
- Mason, C. F., & Phyllips, O. R. (2000). Vertical integration and collusive incentives: An experimental analysis. *International Journal of Industrial Organization*, 18, 471–493.
- Masten, S. (1984). The organization of production: Evidence from the aerospace industry. *Journal of Law and Economics*, 27, 403–417.
- Masten, S., Meehan, J., & Crocker, K. (1989). Vertical integration in the U.S. auto industry: A note on specific assets. *Journal of Economic Behavior and Organization*, 12, 265–273.
- Masten, S., Meehan, J., & Snyder, E. (1991). The costs of organization. *Journal of Law, Economics and Organization*, 7, 1–25.
- Mathewson, F., & Winter, R. (1986). The economics of vertical restraints in distribution. In F. Mathewson & J. Stiglitz (Eds.), *New developments in the analysis of market structures*. MIT Press.
- Monteverde, K., & Teece, D. (1982). Supplier switching costs and vertical integration in the automobile industry. *Bell Journal of Economics*, 13, 206–213.
- Muris, T., David Scheffman, J., & Spiller, P. (1992). Strategy and transactions costs: The organization of distribution in the carbonated soft drink industry. *Journal of Economics and Management Strategy*, 1, 83–128.
- Myerson, R., & Satterthwaite, M. (1983). Efficient mechanisms for bilateral trading. *Journal of Economic Theory*, 28, 265–281.
- Ohanian, N. K. (1994). Vertical integration in the U.S. pulp and paper industry, 1900–1940. *Review of Economics and Statistics*, 76, 202–207.
- Ordover, J., Salop, S., & Saloner, G. (1990). Equilibrium vertical foreclosure. *American Economic Review*, 80, 127–142.
- Perry, M. (1978). Price discrimination and vertical integration. *Bell Journal of Economics*, 9, 209–217.
- Regan, L. (1997). Vertical integration in the property-liability insurance industry: A transaction cost approach. *Journal of Risk and Insurance*, 64, 41–62.
- Shelanski, H., & Klein, P. (1995). Empirical research in transaction cost economics: A review and assessment. *Journal of Law, Economics and Organization*, 11, 335–361.
- Simon, H. A. (1947). *Administrative behavior*. Macmillan.
- Stigler, G. (1951). The division of labor is limited by the extent of the market. *Journal of Political Economy*, 59, 185–193.
- Stukey, J. (1983). *Vertical integration and joint ventures in the aluminum industry*. Harvard University Press.
- Telser, L. G. (1960). Why should manufacturers want fair trade?. *Journal of Law and Economics*, 3 (1), 86–105.
- Tirole, J. (1988). *The theory of industrial organization*. MIT Press.

- Vannoni, D. (2002). Empirical studies of vertical integration: The transaction cost orthodoxy. *International Review of Economics and Business*, 49, 113–141.
- Walker, G., & Weber, D. (1984). A transactions cost approach to make or buy decisions. *Administrative Science Quarterly*, 29, 373–391.
- Waterman, D., & Weiss, A. (1996). The effects of vertical integration between cable television systems and pay cable networks. *Journal of Econometrics*, 72, 357–395.
- Whinston, M. (2003). On the transaction cost determinants of vertical integration. *Journal of Law, Economics and Organization*, 19, 1–23.
- Williamson, O. (1971). The vertical integration of production: Market failure considerations. *American Economic Review*, 61, 112–123.
- Williamson, O. (1975). *Markets and hierarchies: Analysis and antitrust implications*. Free Press.
- Williamson, O. E. (1979). Transaction cost economics: the governance of contractual relations. *Journal of Law and Economics*, 22(2), 3–61.
- Williamson, O. (1983). Credible commitments: Using hostages to support exchange. *American Economic Review*, 73, 519–540.
- Williamson, O. (1985). *The economic institutions of capitalism*. Free Press.
- Williamson, O. (1996). *The mechanisms of governance*. Oxford University Press.
- Williamson, O. (2000). The new institutional economics: Taking stock, looking ahead. *Journal of Economic Literature*, 38, 595–613.
- Williamson, O., Wachter, M., & Harris, J. (1975). Understanding the employment relation: The analysis of idiosyncratic exchange. *Bell Journal of Economics*, 6, 250–280.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Make-or-Buy Decision Revisited

19

Peter G. Klein and João Fernando Rossi Mazzoni

Contents

19.1	Introduction	448
19.2	The Theory of Vertical Boundaries	449
19.2.1	Transaction Cost Economics	449
19.2.2	Theoretical Extensions	451
19.3	Strategies for Empirical Research	452
19.3.1	Data and Methods	453
19.3.2	Causal Inference	454
19.3.3	Assumptions About Efficiency	455
19.4	Evidence on the Make-or-Buy Decision: A Sampler	456
19.4.1	Procurement	456
19.4.2	Contracts and Contractual Design	458
19.4.3	Platforms	459
19.4.4	Informal Agreements	460
19.5	Challenges and Directions for Further Research	460
19.5.1	Measurement and Definition	461
19.5.2	The Role of Asset Specificity	462
19.5.3	Comparing Rival Explanations	463
19.5.4	The Regulatory and Legal Environment	464
19.6	Conclusion	465
	References	466

Abstract

This chapter builds on Klein (The make-or-buy decision: Lessons from empirical studies. In *Handbook of new institutional economics*, pp. 435–464, 2005) to provide an updated literature review on make-or-buy decisions incorporating the most recent theories, methods, and data sources used and considering newer phenomena related to vertical boundaries, as well as sketching out some future directions of this literature. After summarizing the most common explanations for

P. G. Klein (✉) · J. F. R. Mazzoni

Department of Entrepreneurship and Corporate Innovation, Baylor University, Waco, TX, USA

e-mail: Peter_Klein@baylor.edu; Joao_Mazzoni1@baylor.edu

make-versus-buy, we discuss methodological challenges including data, measurement, and estimation techniques and then review empirical findings on component procurement, the use of contracts and contractual design, and newer organizational forms such as platforms. We then discuss ongoing challenges and implications for future work. While the literature on make-or-buy decisions has grown and diversified over the last 20 years, the core insights of the transaction cost approach and its basic explanatory power in helping us make sense of vertical boundaries remain the same.

Keywords

Vertical integration · Outsourcing · Transaction costs · Asset specificity · Contracting · Platforms · Hybrids · Governance

19.1 Introduction

The “transaction cost” theory of the firm introduced by Coase (1937) has become a standard framework for the study of firm boundaries. The Coasian framework helps explain not only the existence of the firm but also its size and scope. Why, in Coase’s (1937, pp. 393–94) words, “does the entrepreneur not organize one less transaction or one more?” Apple, for example, writes and controls the code operating its devices, manufactures its own chips, and maintains a network of Apple retail stores. Tesla not only manufactures electric cars but also makes their battery cells and home-charging equipment and maintains a network of charging stations; it has even ventured into solar panel production and works on in-house chip design for autonomous driving technology. In contrast, companies like Samsung and Dell use Google’s Android or Microsoft’s Windows to power their mobile devices and desktop computers while outsourcing most production to contract manufacturers in countries like China, Vietnam, and Indonesia (where Apple also contracts out its production). Global spending on outsourcing was estimated at \$731 billion in 2023.¹

What explains these patterns? Why do some firms choose a vertically integrated structure, while others specialize in one stage of production and outsource the remaining stages to other firms? In other words, should a firm make its own inputs, should it buy them on the spot market, or should it maintain an ongoing relationship with a particular supplier? Traditionally, economists viewed vertical integration or vertical control as an attempt to earn monopoly rents by gaining control of input markets or distribution channels. The transaction cost approach, by contrast, emphasizes that vertical coordination can be an efficient means of protecting relationship-specific investments or mitigating other potential conflicts under incomplete contracting. As transaction cost economics was developed in the 1970s and 1980s, a stream of empirical literature emerged explaining the “make-or-buy decision” using transaction cost reasoning. Recent work extends and elaborates this research stream by

¹2023 Global Shared Services and Outsourcing Survey. (n.d.). Deloitte United States. <https://www2.deloitte.com/us/en/pages/operations/articles/shared-services-survey.html>.

integrating transaction cost economics with property rights and capabilities theories, looking more closely at legal and regulatory factors, considering more fully the role of the institutional environment, considering a broader range of “hybrid” arrangements such as platforms and ecosystems, paying closer attention to causal inference, and exploiting new, large-scale databases on firms, establishments, and transactions.

This chapter builds on Klein (2005) to provide an updated literature review on make-or-buy decisions incorporating the most recent theories, methods, and data sources used and considering newer phenomena related to vertical boundaries, as well as sketching out some future directions of this literature. After summarizing the most common explanations for make-versus-buy, we discuss methodological challenges including data, measurement, and estimation techniques and then review empirical findings on component procurement, the use of contracts and contractual design, and newer organizational forms such as platforms. We then discuss ongoing challenges and implications for future work. While the literature on make-or-buy decisions has grown and diversified over the last 20 years, the core insights of the transaction cost approach and its basic explanatory power in helping us make sense of vertical boundaries remain the same.²

19.2 The Theory of Vertical Boundaries

Coase was the first to explain that organizational boundaries depend not only on the productive technology but also on the costs of transacting business. In the Coasian framework, the decision to organize transactions within the firm instead of on the open market—the “make-or-buy decision”—depends on the relative costs of internal versus external exchange. The market mechanism entails certain costs: discovering the relevant prices, negotiating and enforcing contracts, and so on. Within the firm, the entrepreneur may be able to reduce these “transaction costs” by coordinating these activities himself. However, internal organization brings other kinds of transaction costs, namely, problems of information flow, incentives, monitoring, and performance evaluation. The boundary of the firm, then, is determined by the tradeoff, at the margin, between the relative transaction costs of external and internal exchange. In this sense, firm boundaries depend not only on technology but also on organizational considerations; that is, on the costs and benefits of various contracting alternatives.

19.2.1 Transaction Cost Economics

The transaction cost perspective (Williamson, 1975, 1985, 1996b) provides a powerful set of tools for analyzing firm boundaries. Transaction cost economics (TCE)

²Other surveys of this literature, from a variety of perspectives, include Joskow (1988a); Shelanski and Klein (1995); Rindfleisch and Heide (1997); Masten and Saussier (2000); Vannoni (2002); Macher and Richman (2008); David and Han (2004); Zhao, Luo, and Suh (2004); and Lafontaine and Shaw (2005). Masten (1996) collects many of the important earlier articles.

begins with the observation that economic organization, both internal and external, imposes costs because complex contracts are usually incomplete—they provide remedies for only some possible future contingencies. Contractual incompleteness exposes the contracting parties to certain risks. Primarily, if circumstances change unexpectedly, the original governing agreement may no longer be effective. The need to adapt to unforeseen contingencies constitutes an additional cost of contracting; failure to adapt imposes what Williamson (1991a) calls “maladaptation costs.”

The most often-discussed example of maladaptation is the “holdup” problem associated with relationship-specific investments. Investment in such assets exposes agents to a potential hazard: If circumstances change, their trading partners may try to expropriate the rents accruing to the specific assets. Rents can be safeguarded through vertical integration, where a merger eliminates any adversarial interests. Less extreme options include long-term contracts, partial ownership, or agreements for both parties to invest in offsetting relationship-specific investments. According to transaction cost theory, parties tend to choose the governance structure that best controls the underinvestment problem, given the particulars of the relationship.

In this sense, transaction cost economics may be considered the study of *alternative institutions of governance*. Its working hypothesis, as expressed by Williamson (1991b, p. 79), is that economic organization is mainly an effort to “align transactions, which differ in their attributes, with governance structures, which differ in their costs and competencies, in a discriminating (mainly, transaction cost economizing) way.” Simply put, the contractual approach tries to explain how trading partners choose, from the set of feasible institutional alternatives, the arrangement that best mitigates the relevant contractual hazards at least cost.

The theory is fleshed out by specifying which governance structures go with which transactions. Transactions differ in the degree to which relationship-specific assets are involved, the amount of uncertainty about the future and about other parties’ actions, the frequency with which the transaction occurs, and so on. Each matters for the preferred institution of governance, although the first—asset specificity—is particularly important. Williamson (1985, p. 55) defines asset specificity as “durable investments that are undertaken in support of particular transactions, the opportunity cost of which investments is much lower in best alternative uses or by alternative users should the original transaction be prematurely terminated.” This could describe a variety of relationship-specific investments, including both specialized physical and human capital, along with intangibles such as R&D and firm-specific knowledge or capabilities.

Governance structures include markets, hierarchies, and hybrids. The pure anonymous spot market suffices for simple transactions such as basic commodity sales. Market prices provide powerful incentives for exploiting profit opportunities, and market participants are quick to adapt to changing circumstances as information is revealed through prices. When relationship-specific assets are at stake, however, and when product or input markets are thin, bilateral coordination of investment decisions may be desirable, and combined ownership of these assets may be efficient. Ownership is completely combined in the fully integrated firm. The transaction cost

approach maintains that such hierarchies offer greater protection for specific investments and provide relatively efficient mechanisms for responding to change where coordinated adaptation is necessary. Compared with decentralized structures, however, hierarchies give managers weaker incentives to maximize profits and normally incur additional bureaucratic costs. Alternatively, partial alignment may be achieved within an intermediate or hybrid form, such as long-term contracts, partial ownership agreements, franchises, networks, alliances, platforms, and firms with highly decentralized assignments of decision rights.

19.2.2 Theoretical Extensions

TCE remains a core theoretical perspective for analyzing make-or-buy decisions, and the strategic management literature also builds on the resource-based view (RBV) which offers a complementary perspective, emphasizing the firm's internal resources and capabilities as a key driver of a firm's activities as well as its performance (Parmigiani, 2007; Roy & Sarkar, 2016; Davis & DeWitt, 2021). TCE and RBV are complementary in that resources and capabilities explain what the firm does (e.g., what resources will be involved in production, what industries the firm will be in), while TCE explains how these activities should be organized and governed (for instance, whether the resources will be owned or rented, whether the business units will be wholly owned subsidiaries or alliance partners, and so on).

A puzzle for the TCE explanation is the finding by Atalay, Hortaşçu, and Syverson (2014), using establishment-level data on the US manufacturing sector, that many vertically integrated firms do not transfer physical components across their own production stages. From this, these authors argue that vertical integration is often used not to govern transactions—that is, to lower the cost of procuring inputs a firm uses in production—but as a means of allocating scarce managerial talent across production units. In other words, firms add vertically related subsidiaries for the same reason they add other subsidiaries, namely, to leverage their unique intangible resources and capabilities, not to coordinate specific production processes.³

Many recent papers augment the core transaction cost explanation by emphasizing factors that moderate the relationship between asset specificity, uncertainty, frequency, and other transactional characteristics and organizational form. These moderators include managerial capabilities (Brahm and Tarzijan, 2014), the institutional environment (Bucheli et al., 2023), and the quality of corporate governance (Gartenberg & Pierce, 2017). Other legal and regulatory factors such as trade credits can also protect against opportunism (Breza & Liberman, 2017). An important setting is networks, platforms, or ecosystems or a collaborative context such as open innovation in which contracting and vertical integration decisions affect not only buyer-supplier relations but also the focal firm's relationship with network partners or rivals (Gambardella & Panico, 2014; Cassiman & Valentini, 2016). Other

³This explanation is similar to Coase's (2000) reinterpretation of the acquisition of Fisher Body by G.M., discussed in more detail below.

works (e.g., Rawley & Simcoe, 2013; Gil & Ruzzier, 2018) show how productivity, competition, and similar factors can influence vertical integration even in the absence of asset specificity. Helfat and Campo-Rembado (2016) argue that integrated and specialized firms coexist within industries not because of differences in transaction costs but because integration allows product differentiation in terms of technological advancements or unique aesthetic features which are difficult to facilitate through contract.

Vertical integration can also facilitate adaptation and learning (Sanders et al., 2007; Moschuris, 2007). For example, Forbes and Lederman (2009) examine regional contracting in the US airline industry and show that airlines are more likely to own and operate their own ground handling, maintenance, and catering services when serving routes that require rapid (weather-related) adaptation, thus avoiding costly ex post renegotiation. Zhou and Wan (2017) study the soft drink industry and find that integration of bottling by concentrate producers improves coordination by giving producers greater flexibility to adapt to consumer tastes, more direct control over investments in the bottling processes, and improved coordination along the value chain. As Ménard (2013) points out, uncertainty about the optimal organizational form, especially for transactions that could potentially be governed by market, hierarchy, or hybrid modes, can also lead to plural sourcing, i.e., firms both making and buying the same intermediate goods (see also He & Nickerson, 2006; Parmigiani & Mitchell, 2009; Sako et al., 2016).⁴

19.3 Strategies for Empirical Research

Early empirical work on the make-or-buy decision adopted the transaction cost framework and followed the same basic model. The efficient form of organization for a given economic relationship—and, therefore, the likelihood of observing a particular organizational form or governance structure—is seen as a function of certain properties of the underlying transaction or transactions: asset specificity, uncertainty, frequency, and so on. Organizational form is the dependent variable, while asset specificity, uncertainty, complexity, and frequency are independent variables. Specifically, the probability of observing a more integrated governance structure depends positively on the amount or value of the relationship-specific assets involved and, for significant levels of asset specificity, on the degree of uncertainty about the future of the relationship, on the complexity of the transaction, on the frequency of trade, and possibly on some aspects of the institutional environment.

At the transaction level, the organizational form is typically modeled as a discrete variable (“make,” “buy,” or “hybrid” or the presence or absence of a particular

⁴Some recent work in contract theory also questions the role of contractual incompleteness (Maskin & Tirole, 1999). In response, theorists have argued that contracts and vertical integration can serve as useful “reference points” for renegotiating agreements in the face of unanticipated events, even in the absence of opportunism and holdup (Hart & Moore, 2008; Hart, 2009; Halonen-Akatwijuka & Hart, 2013). This stream has not yet generated much empirical literature.

feature in a supply or distribution contract). At the firm level or the buyer-supplier dyad level, vertical integration can be measured as a continuous measure, such as the percentage or value of shipments procured internally or externally. Of the independent variables, asset specificity has received the most attention, presumably because of the central role it plays in the transaction cost approach to vertical integration.⁵

19.3.1 Data and Methods

Early empirical work on the make-or-buy decision relied on idiosyncratic, hand-collected datasets of transactions in one firm or industry with measures of relationship-specific investment such as product complexity, coded from survey data or quantitatively assigned by inspection, as a proxy for physical asset specificity (Masten, 1984; Bigelow, 2001); worker-specific knowledge, again coded from survey data, as a proxy for human asset specificity (Monteverde & Teece, 1982a, b; Masters and Miles, 2002; Shi et al., 2018; Souza Filho & Miranda, 2019; Takata & Parry, 2022); physical proximity of contracting firms, as a proxy for site specificity (Joskow, 1985, 1987, 1988b, 1990; Spiller, 1985; González-Díaz et al., 2000; Vita et al., 2010); and spatial and temporal proximity (Masten et al., 1991; Pirrong, 1993; Hubbard, 1999; Liu et al., 2008; Diez-Vial & Alvarez-Suescun, 2011).⁶ Common proxies for uncertainty include sales variance (Levy, 1985; Anderson & Schmittlein, 1984) and some measure of technological uncertainty, such as the frequency of changes in product specification and the probability of technological change (Walker & Weber, 1984; Crocker & Reynolds, 1993).

Case studies comprise the bulk of the studies on the make-or-buy decision, primarily because the main variables of interest—asset specificity, uncertainty, frequency—are difficult to measure consistently across firms and industries. In early studies, these characteristics were estimated based on surveys or interviews—for example, asking managers to rate the degree to which an investment has value in outside uses on a Likert-type scale. More recent work has taken advantage of several new, secondary datasets for measuring firm boundaries and internal organization such as the US Census Bureau's Economic Census and Commodity Flow Survey which capture shipments across vertically linked establishments of the same corporate parent. Industry-level measures of vertical integration can also be computed from the Census Bureau's Annual Wholesale Trade and Annual Retail Trade Surveys.

⁵Williamson (1991a) distinguishes among (1) site specificity, in which parties are in a “cheek-by-jowl” relationship to reduce transportation, and inventory costs and assets are highly immobile; (2) physical asset specificity as in relationship-specific equipment and machinery; (3) human asset specificity referring to transaction-specific knowledge or human capital; (4) brand-name capital reflected in intangible assets reflected in consumer perceptions; (5) “dedicated assets” referring to substantial, general-purpose investments that would not have been made outside a particular transaction, the commitment of which is necessary to serve a large customer; and (6) temporal specificity meaning assets that must be used in a particular sequence.

⁶Other proxies, such as fixed costs or “capital intensity,” have more obvious limitations and are rarely used (see discussion in Von Hirschhausen & Newman, 2008; Tan et al., 2023).

For cross-country studies, Dun and Bradstreet's World Base includes firm-level and establishment-level data for over 300 million global companies, including a vertical integration index used in several studies of firm boundaries (Alfaro et al., 2016; Acemoglu et al., 2009). At this level of aggregation, transaction costs can be measured at the country level using measures of financial development, procedural complexity, and contract enforcement from the World Bank's World Development Indicators and Doing Business index (Acemoglu et al., 2009).

Unfortunately, large-sample, cross-country measures of transaction-level attributes such as asset specificity remain elusive. A recent exception is Kermani and Ma (2023) who use bankruptcy filings to estimate the level of asset specificity among a large sample of US firms. They estimate that liquidation values of fixed assets held by distressed firms are only a third of the industry-adjusted net book value of these assets, suggesting a substantial degree of asset specificity. They also show that the physical characteristics of these assets (mobility, durability, and customization) explain much of this variation. Similar methods for measuring relationship-specific investment should prove highly valuable for future work on determinants of vertical boundaries.

On the whole, the evidence from case studies as well as large-sample econometric analyses is consistent with the argument that make-or-buy decisions mainly reflect efficiency considerations, rather than the exploitation of market power. In particular, transactions are typically organized to minimize the hazards associated with asset specificity, though many outstanding challenges, puzzles, and issues remain.

19.3.2 Causal Inference

As with empirical social science more generally, quantitative empirical research on make-or-buy decisions has paid increasing attention to causal inference. Much of the early econometric work used cross-sectional data, making it difficult to identify a treatment effect of transactional or environmental characteristics on organizational form. These studies typically tested a reduced-form model where the likelihood of observing a more hierarchical form of governance increases with the degree of relationship-specific investments. There are two sets of problems with this approach, however.

First, the reduced-form model considers only one side of the equation, namely, the impact of transactional characteristics on transaction costs between firms, while ignoring the potential impact of these characteristics of internal organizational costs. If these characteristics reduce the costs of internal organization, then asset specificity could lead to integration, independent of the holdup problem or other maladaptation costs (Masten, 1994, p. 10). Unfortunately, we have more theory and evidence on "external" transaction costs (such as those identified by Coase, 1937) than theory and evidence on internal organizational costs.⁷

⁷ A few studies attempt to measure internal organizational costs directly, including Masten, Meehan, and Snyder (1991) and Fisher and Connelly (2017).

Second, firms typically choose—or at least have influence over—both transactional attributes and organizational form, making it difficult to identify a causal impact of one on the other. Time-varying factors such as changes in technology or market conditions may also influence both the make-or-buy decision and the transactional characteristics under consideration. Addressing this problem typically involves searching for plausibly exogenous sources of variation in transactional or environmental attributes and using instrumental variables, differences-in-differences, regression discontinuity, or a similar design to estimate treatment effects.

For example, Rawley and Simcoe (2013) estimate the impact of a productivity-enhancing productivity shock (the adoption of computerized dispatching systems) on vertical integration in the US taxicab industry. They instrument for the productivity shock using a series of industry and demographic variables unrelated to the focal firm's asset ownership decision. Murcia, Panwar, and Tarzijan (2021) use a similar approach to study how the adoption of Corporate Social Responsibility (CSR) affects supply chains, finding that vertical integration allows for closer monitoring and trust-building that facilitate CSR. They instrument for the focal firm's CSR practices using the average CSR scores of same-industry peers, a technique that has been used widely in the literature on organizational practices (Klein & Saidenberg, 2010; Kolympiris et al., 2019).

Several papers exploit exogenous legal and policy changes as natural or quasi-natural experiments. Gil and Ruzzier (2018) study the impact of competition on vertical boundaries in the Spanish TV sector using regulatory shocks to instrument for competition, while Breza and Liberman (2017) exploit a change in trade-credit regulations in Chile to estimate the impact of contracting costs on vertical integration. Alfaro et al. (2016) show how trade policy impacts outsourcing prices that drive the make-or-buy decision. Other papers looking at legal or regulatory shocks include Guadalupe and Wulf's (2010) study of trade liberalization's impact on corporate hierarchies and Ferguson and Formai's (2013) finding that the impact of the quality of the legal system on contract enforcement is weaker in industries that are more vertically integrated.

A few papers (Liston et al., 2007; Uygun et al., 2023) use simulation methods to analyze make-or-buy decisions. While these models can illustrate the impact of particular mechanisms, they do not address causality directly.

19.3.3 Assumptions About Efficiency

A more general problem with the empirical literature on vertical integration—or, for that matter, any aspect of organizational form—relates to counterfactuals. Transaction cost reasoning implies firms choose transactional characteristics and governance structures as a bundle. The study of alternative institutional arrangements is thus the analysis of these bundles—which organizational or governance structure is most efficacious for a particular kind of transaction? Empirically, however, we observe only the combinations chosen—not the most efficient of all the possible combinations.

One approach is to assume that market forces cause an “efficient sort” between transactions and governance structures. Williamson (1985, pp. 119–20) writes, for instance, that “backward integration that lacks a transaction cost rationale or serves no strategic purposes will presumably be recognized and will be undone,” adding that mistakes will be corrected more quickly “if the firm is confronted with an active rivalry.” One way to examine this conjecture more closely is to see if appropriately organized firms—that is, firms that match transactional characteristics to governance structures as the theory says they should—outperform the feasible alternatives. Silverman, Nickerson, and Freeman (1997), for example, show that transaction cost efficiency is positively correlated with firm survival in the for-hire trucking industry, while Bigelow (2001) examines outsourcing arrangements in the US automobile industry and finds that transactions that are appropriately aligned tend to last longer than inappropriately organized ones.

19.4 Evidence on the Make-or-Buy Decision: A Sampler

Here we review the empirical evidence on procurement, contracting, platforms, and other contexts. The studies cited here, along with additional examples of applications to different industries, products, and technologies, are summarized in Table 19.1.

19.4.1 Procurement

The decision to make components internally or procure them on the open market was the first topic studied extensively within the transaction cost framework. Early efforts by Monteverde and Teece (1982b) and Masten (1984) use samples of components, coded as either made or bought, along with proxies for asset specificity such as worker-specific knowledge and component complexity as rated by industrial engineers. These papers use probit or logit models to test the relationship between in-house production and asset specificity, along with uncertainty and other control variables, finding that asset specificity is a statistically significant predictor of vertical integration.

Refinements to this basic approach include distinguishing among types of uncertainty (e.g., Walker & Weber, 1984) and types of asset specificity, particularly relationship-specific human capital (e.g., Masten et al., 1989). Recent work also identifies additional benefits and costs of alternative procurement modes beyond those identified in the transaction cost literature. For example, Reitzig and Wagner (2010) discuss what they call “hidden costs” of outsourcing such as loss of expertise within the firm, reduced ability to detect and respond to competition, and the potential for decreased performance in downstream activities. Vertical structure can also impact productive efficiency, which transaction cost models tend to hold constant (Zhang et al., 2018), innovation and technical change (Li & Tang, 2010; Kapoor & Adner, 2012), and the productivity of potential outsourcing partners (Bakhtiari, 2015), which in turn affects the choices available to the focal firm.

Table 19.1 Empirical evidence on make-or-buy decisions

Subject area	Industry, technology, or context	Examples
Procurement	Electricity generation	Joskow (1985); Saussier (2000)
	Aerospace	Masten (1984)
	Aluminum	Stuckey (1983); Hennart (1988)
	Forestry	Globerman and Schwindt (1986)
	Paper industry	Lafontaine and Shaw (1999)
	Chemicals	Lieberman (1991)
	Engineering	Lyons (1995)
	Semiconductor industry	Kapoor (2013)
	Trucking	Nickerson and Silverman (2003)
	Offshore oil gathering	Hallwood (1991)
	Information technology	Ulset (1996)
	Electronic components	Weiss and Kurland (1997)
	Construction	González-Díaz, Arruñada, and Fernández (2000)
	Stock exchanges	Bindseil (1997)
	Distribution shipments across several industries	Caves and Bradburd (1988)
	Vertical integration using multi-industry data	Levy (1985); MacMillan, Hambrick, and Pennings (1986); Harrigan (1986)
	Spanish construction firms	González-Díaz, Arruñada, and Fernández (2000)
	Goods-producing industries	Atalay, Hortaçsu, and Syverson (2014)
	Different firms from 93 countries	Acemoglu et al. (2009)
	Firm-level data from manufacturing	Alfaro et al. (2023)
	Firm-level data from manufacturing	Prete et al. (2017)
	Industry supply chain	Fan et al. (2022)
Contracts and contractual design	Failure of the G.M.–Fisher body contract	Klein, Crawford, and Alchian (1978); Klein (1988); Langlois and Robertson (1989); Coase (2000); Helper, MacDuffie, and Sabel (2000); Casadesus-Masanell and Spulber (2000); Freeland (2000)
	Choice between contracts and other hybrids	Pisano (1990); Oxley (1997, 1999); Breza and Liberman (2017)
	Internal use of contracts	Magelssen, Rich, and Mayer (2022)
	Specific contractual provisions	Gallick (1984); Joskow (1985, 1987, 1988b, 1990); Crocker and Masten (1988); Crocker and Reynolds (1993); DeCanio and Frech (1993); Heide, Dutta, and Bergen (1998); Saussier (2000); Gray et al. (2009)

(continued)

Table 19.1 (continued)

Subject area	Industry, technology, or context	Examples
Informal agreements	Property disputes	Ellickson (1991)
	Fresh fish	Wilson (1980)
	Lobsters	Acheson (1985)
	Whaling	Ellickson (1989); Gifford (1993)
Forward integration into marketing and distribution	Direct versus independent sales forces	Anderson and Schmittlein (1984); Anderson (1985); John and Weitz (1988); Regan (1997); Fein and Anderson (1997)
	Direct versus independent distribution	Muris, Scheffman, and Spiller (1992)
Platforms	Effect on transaction costs	Bai and Ramakrishna Velamuri (2021); Gawer (2021); Munger (2021); Akbar and Tracogna (2022); Chen et al. (2022)
	Governance, reputation mechanisms	Ter Huurne et al. (2017); Hein et al. (2020); Lahiri et al. (2023)
	Status of input providers	Uzunca et al. (2018); Hughes (2022)
Miscellaneous	Block booking	Kenney and Klein (1983)
	Multinational corporations	Hennart (1989); Yarbrough and Yarbrough (1987); Gatignon and Anderson (1988); Klein et al. (1990); Hu and Chen (1993); Henisz (2000)
	Company towns and company stores	Fishback (1986, 1992)
	Rise of medieval marketplaces and towns	Bindseil and Pfeil (1999)
	Marriage	Carbone and Brinig (1990); Hamilton (1999)

19.4.2 Contracts and Contractual Design

While the earliest literature on the make-or-buy decision—starting with Coase (1937)—treated external sourcing and in-house procurement as polar opposites, we observe firms choosing a variety of intermediate or hybrid forms of organization such as long-term contracts, partial ownership agreements, franchises, networks, alliances, and other combinations. In the transaction cost approach, a hybrid such as a long-term contract represents a blend, or compromise, between the benefits of centralized coordination and control and the incentive and informational advantages of decentralized decision-making (Williamson, 1991a; Ménard, 2004).

However, while TCE holds that contracting is appropriate at “moderate” levels of asset specificity (Williamson, 1985), vertical integration and contracting may serve fundamentally different purposes. For example, consider the controversy surrounding the purchase of Fisher Body by G.M. in 1926. Klein, Crawford, and Alchian (1978) and Klein (1988) cite the case as a classic example of vertical integration designed to mitigate holdup in the presence of asset specificity. Fisher refused to

locate its plants near G.M. assembly plants and to change its production technology in the face of an unanticipated increase in the demand for car bodies, leading G.M. to terminate its existing 10-year supply contract with Fisher and acquire full ownership. Coase (2000), revisiting the original documents, argues instead that the contract performed well and was gradually replaced with full ownership only to get Fisher's top managers (the Fisher brothers) more closely involved in G.M.'s other operations (as non-owner managers), giving G.M. better access to the Fisher brothers' managerial capabilities—which may have been greater than their capabilities as owners (Foss et al., 2021). While this interpretation is difficult to test directly, the existence of vertically integrated firms that do not transact across their subsidiaries (Atalay et al., 2014) is consistent with a knowledge-based explanation for combined ownership.

19.4.3 Platforms

Digital platforms—technologically mediated systems that enable interactions between user groups and allow those groups to carry out defined tasks using a standardized interface (Bonina et al., 2021)—can be interpreted as a form of hybrid governance that involves decentralized exchange but is distinct from conventional market mechanisms (Stallkamp & Schotter, 2021; Chen et al., 2021). Digital platforms typically focus on narrow market segments and do not own core facilities and equipment or employ their key service providers. The assets they own such as software, data analytics capability, and user data are often fungible and useful in multiple markets, reducing the level of asset specificity and allowing for greater flexibility in acquiring and utilizing digital assets.

By connecting buyers and sellers or service providers, digital platforms enable more efficient transactions and reduce transaction costs (Gawer, 2021; Munger, 2021). Like other hybrids, digital platforms rely on partners pooling strategic resources and sharing decision rights while retaining ownership of key digital assets, with partners owning their own human capital, knowledge, and other key resources. The governance mechanisms of digital platforms involve a combination of incentive alignment and the use of authority to achieve coordination and maximize the value created by complementors (Hein et al., 2020).

Sharing economy platforms blend elements of traditional business models such as payment processing, insurance, and customer support with peer-to-peer or collaborative means for producing and distributing content and facilitating interactions (Bai & Ramakrishna Velamuri, 2021; Akbar & Tracogna, 2022). The growth of sharing platforms has had significant effects on technological change (Belk, 2014; Oskam & Boswijk, 2016), workplace practices (Bouncken & Reuschl, 2018), and environmental protection (Cohen & Kietzmann, 2014). Evidence suggests that platform participants are more likely to share resources such as digital or knowledge-based assets that are modular and create value that cannot be appropriated by other platform participants or the platform owner and less likely to share highly specialized physical items or complex services associated with high transaction costs (Chen et al., 2021).

As with market-based governance, reputation can substitute for formal contracting to enforce agreements on sharing-economy platforms (Ter Huurne et al., 2017;

Lahiri et al., 2023). These systems are often supported and managed by the platform operator, bridging the gap between traditional business trust mechanisms and peer-to-peer trust. Sharing economy platforms operate in regulatory gray areas in many cases (e.g., courts continue to debate whether platform service providers should be treated as employees for tax and regulatory purposes) (Uzunca et al., 2018; Hughes, 2022).

While platforms clearly reduce transaction costs, the impact of specific governance arrangements deserves closer scrutiny. For example, in the case of the video-game industry, Lee (2013) suggests that exclusive dealing arrangements between platforms and software providers reduce user well-being by reducing demand and limiting consumer choice.

19.4.4 Informal Agreements

As mentioned above, contractual relations need not be fully formal and explicit; trading arrangements are often governed by less formal, relational norms. Palay (1984, 1985) studies the role of informal, legally unenforceable agreements between rail-freight carriers and shippers. Shipment of items like automobile parts and chemicals, for example, requires specially designed rail cars and equipment that cannot be easily redeployed for other uses. Because vertical integration was prohibited by regulation, informal agreements emerged to protect these relationship-specific investments.

Baker, Gibbons, and Murphy (2001) show, more generally, how relational contracting—both between and within firms—differs from formal contracting. Relational contracts have the advantage that outcomes need not be verifiable to third parties, such as courts, to limit parties' incentives to behave opportunistically. On the other hand, the absence of third-party participation means that such agreements must be self-enforcing. Heide and John's (1992) study of buyer–supplier relations finds that relational norms often substitute for vertical integration as a means of protecting specific assets; Anderson and Weitz (1992) and Brown, Dev, and Lee (2000) show that such norms are also important in marketing.

19.5 Challenges and Directions for Further Research

The vast empirical literature on the make-or-buy decision, including the structure of long-term contracts and hybrid forms of organization, is largely consistent with the transaction cost theory of the firm: vertical arrangements are usually best understood as attempts to protect trading partners from the hazards of exchange under incomplete contracting. As Joskow (1991, p. 47) observes, the literature on the make-or-buy decision is in many ways in “much better shape than much of the empirical work in industrial organization generally.”⁸ However, important challenges, puzzles, and opportunities remain. First, the measurement and definition of asset specificity, uncertainty, and other variables remain inconsistent. Second, many studies do not

⁸Williamson (1996a, p. 55) puts it bluntly: “Transaction cost economics is an empirical success story.”

explicitly compare rival explanations for vertical relationships. Third, correlation between transactional attributes and governance structures is often mistaken for causality. Fourth, the legal and regulatory environments do not always get sufficient attention.

19.5.1 Measurement and Definition

As mentioned above, empirical research on make-or-buy decisions is often hampered by confusion about the definitions of, and therefore the empirical proxies for, key variables such as asset specificity and uncertainty. Asset specificity is difficult to measure consistently across industries, partially explaining why there are far more single-industry studies of vertical boundaries than cross-industry studies. Uncertainty is hard to define, let alone measure.⁹ Moreover, empirical studies sometimes treat uncertainty as an independent variable, regressing the choice of organizational form on the variance of sales or another variable, but without including any measure of asset specificity in the model. Absent fixed investments, however, transaction cost economics does not predict that uncertainty would itself lead to hierarchical governance. Changes in circumstances only allow for expropriation where there are quasi-rents at risk, that is, where one side's investment is exposed. Where there are no relationship-specific investments at stake, it may be less costly for a firm to contract on the market for goods and services in an uncertain environment than to assume the risk of producing them internally. In this way, the effect of uncertainty depends on competitive conditions. If there is no asset specificity and thus there are many potential suppliers of a component for which future demand is uncertain, it may be cheaper to buy the component than to make it internally.¹⁰

Asset specificity has been more successfully treated in the empirical literature; relationship-specific physical, site, and human capital investments have all been studied, both independently and comparatively. However, further refinement and analysis need to be done here, particularly concerning measurement. Proxies such as capital intensity or fixed costs are very imperfect and may not capture whether the

⁹As discussed above, the empirical literature on vertical integration tends to use fairly crude measures of uncertainty (such as the variation of sales). Distinctions between systemic and idiosyncratic risks, between demand and supply (or technological) risk, and between risk and Knightian uncertainty have rarely been addressed.

¹⁰In some situations, uncertainty is so great that efficient governance structures cannot be crafted at all, in which case trade may fail to materialize. While there is a considerable stream of theoretical literature, following Akerlof (1970), on the possibility that markets might break down due to private information, there is relatively little theoretical or empirical work on non-market exchange under these conditions. An exception is Wiggins and Libecap's (1985) study of unitization agreements in oil production. Under such an agreement producers designate a single firm to develop a given field, with the net returns shared among all producers. This reduces recovery costs and improves oil yields by eliminating the negative externalities associated with concurrent independent development of a single field. Yet very few oil fields are unitized. Wiggins and Libecap argue that asymmetric information encourages opportunistic holdout strategies that have usually prevented the agreements from being signed.

investment has value outside the transaction for which it was initially made.¹¹ Another concern is that asset-specificity effects may be confused with market power. While specific investment may lead to bilateral monopoly, a small-number bargaining situation is not by itself evidence of relationship-specific investment.

19.5.2 The Role of Asset Specificity

While the early transaction cost literature emphasized asset specificity as the key to an efficiency explanation for vertical integration, several studies use transaction cost and incomplete-contracting theory to explain vertical integration in the absence of asset specificity. Pirrong (1993) argues that long-term contracts (and sometimes vertical integration) can be efficient in the presence of smaller contracting hazards—even when physical, human, and site asset specificities are absent. In a study of bulk shipping, he finds that more integrated governance structures can dominate spot trading in the presence of what Masten, Meehan, and Snyder (1991) call “temporal specificities.” When a processing or refinery plant contracts with a particular bulk carrier, for example, both plant and carrier capacities suddenly become specific assets. Small delays in delivery can then result in large losses of quasi-rents for the plant, just as the plant’s refusal to take full delivery can impose substantial losses on the carrier. To avoid costly strategic bargaining, then, these parties will choose a complex, long-term agreement.

Studies of the US trucking market also find long-term contracts in the absence of asset specificity. Both Williamson (1985, p. 54) and Klein, Crawford, and Alchian (1978, p. 244) cite trucks as clear examples of durable, but nonspecific, assets. Yet, as shown by Hubbard (1999) and Nickerson and Silverman (2003), trucking continues to be characterized by tight vertical coordination between hired drivers and shippers, rather than market-based coordination between independent owner-operators and shippers, even after the industry was deregulated in the 1970s. Hubbard (1999) uses this evidence to challenge the scope of holdup theories more generally. Nickerson and Silverman (2003) show that the need for temporal coordination among hauls, and the shipper’s desire to protect its brand-name capital, leads to tight vertical control, explaining why many shippers continue to use company-owned trucks. Lafontaine and Masten (2002) argue that the observed variation in contractual arrangements in trucking can best be explained by driver and truck heterogeneity, not asset specificity nor marginal incentive considerations.

None of these studies denies that physical asset specificity and site specificity are important determinants of vertical relations, only that some cases of vertical control can be explained without reference to them, or to holdup problems at all. At present, these results appear as exceptions to a more general rule. Still, an accumulation of

¹¹ The same applies to measures of human asset specificity, such as training, used to explain labor outsourcing (see, e.g., Masters and Miles, 2002).

such anomalies could challenge the key underlying structure of the transaction cost approach.

19.5.3 Comparing Rival Explanations

Besides these difficulties of measurement and definition, and the role of key independent variables, empirical research on vertical boundaries is also subject to the problem found in empirical work generally: alternate hypotheses that could also fit the data are rarely stated and compared. Usually, the data are found only consistent or inconsistent with the hypothesis at hand. Undoubtedly, studies that explicitly compare competing, observationally distinct hypotheses about contractual relationships are needed, because rival theories commonly posit mutually exclusive outcomes.

One example is Spiller's (1985) comparison of asset-specificity and market-power explanations for vertical mergers, explanations that have rival predictions about the size of the gains from mergers under various competitive conditions. He finds the total gain from merger to be smaller where the distance between the merging firms is greater (i.e., where site specificity is lower). He also finds no significant relationship between industry concentration and the distance between merging firms. These findings appear to support the asset-specificity explanation over the market-power explanation.

Poppo and Zenger (1995) ask whether transaction cost or resource-based reasoning best explains outsourcing arrangements in the information-technology (IT) sector. They find that asset specificity reduces the performance of market transactions, as predicted by TCE. However, they find no evidence for the claim that asset specificity also increases the performance of in-house transactions, an implication of the resource-based view that is central to Ghoshal and Moran's (1996) critique of TCE. Other studies that assess both transaction cost and rival theories include Poppo and Zenger (1998), Silverman (1999), and Nickerson, Hamilton, and Wada (2001). All provide support for the role of asset specificity in explaining vertical integration, sometimes in conjunction with other mechanisms.

Teece (2010) argues that make-or-buy decisions depend not only on asset specificity but also on other asset characteristics emphasized in a capabilities perspective such as non-tradability and non-availability. In his approach, while TCE highlights the need to control opportunism, a capabilities perspective is needed to explain a firm's systems for sensing and seizing opportunities. He also notes that firms may integrate forward to profit from future innovations, even in the absence of transaction cost factors. Parmigiani (2007) likewise contrasted outsourcing explanations offered by TCE, neoclassical economics, and firm capabilities, finding that each offers some independent explanatory value.

Moreover, while the evidence presented in this chapter is often interpreted as supportive of both (a) the transaction cost theory as explicated by Williamson and Klein, Crawford, and Alchian and (b) the more formal version (the "property-rights approach") associated with Grossman, Hart, and Moore, there are important

differences between these two sets of explanations for firm boundaries (see Williamson, 2000, pp. 605–07 for a detailed discussion of these differences). For example, property-rights models focus exclusively on ex ante underinvestment in relationship-specific human capital brought about by inefficient ownership arrangements, while transaction cost theories look more at the ex post contract execution stage. Partly because ex post contractual problems are easier to observe than ex ante human capital underinvestment—How, for instance, is optimal investment to be measured?—there have been relatively few empirical studies explicitly in the property-rights tradition.¹² Whinston (2000) argues that, in property-rights models, the *level* of appropriable quasi-rents is not relevant for the integration decision; only *marginal* quasi-rents matter. Few empirical studies make this distinction. Furthermore, property-rights models offer specific predictions on the direction of integration (whether buyer acquires seller or seller acquires buyer), a distinction that is also generally ignored in the empirical literature.

19.5.4 The Regulatory and Legal Environment

Contracting takes place “in the shadow of the law” (Cooter et al., 1982), and the empirical work on vertical boundaries could be improved by greater attention to the effects of the regulatory and legal environments. With notable exceptions (like the work on contracting among public utilities and their suppliers), the literature has generally focused on relatively unregulated industries operating under a relatively stable legal regime. However, differences in regulatory structures, or how judges interpret contractual clauses, can have substantial effects on the performance of alternative vertical arrangements, like in Zanarone (2013). Neglecting such differences leads to biased estimates of the effects of other factors, such as asset specificity and uncertainty, on the decision to vertically integrate or to use long-term contracts.

Henisz (2000), Delios and Henisz (2000), and Henisz and Zelner (2001) look more closely at the relationship between contractual hazards and political hazards. Henisz’s (2000) study of foreign investment finds that firms tend to prefer joint ventures with foreign partners rather than majority owned plants where political hazards are high, even though majority ownership may better mitigate the contractual hazards associated with asset specificity. The available contracting options may also be limited by regulation. Moreover, as discussed above, Palay (1984, 1985) shows that informal agreements can substitute for regulation when vertical integration is prohibited. Loredó and Suárez (2000) study the coal-burning electricity plants in Spain and find that the opportunism was mitigated by the regulatory compact between firms and the state, instead of the long-term contracts used by US plants. Ménard and Klein (2004) compare vertical relations in US and European agriculture and suggest that the variation is partly explained by differences in the institutional

¹²Hart (1995, p. 49) remarks that there has been “no formal testing of the property rights approach.” (By “formal testing” he presumably excludes case studies).

environment. Gayle and Weisman (2007), analyzing the (heavily regulated) telecommunications industry, show how efficiency is reduced when regulators determine input prices. Similar industry-level studies on the impact of local regulations explore the TV (Gil & Ruzzier, 2018), telecommunications (Beard et al., 2005), and manufacturing sectors (Bartel, Lach, and Sicherman (2014). Mandy (2009) points out that regulation, particularly input-pricing policies, influences make-or-buy decisions in retail markets. And Bucheli et al.'s (2023) study of the banana industry shows how the broader institutional environment affects vertical integration decisions. They argue that multinational corporations will integrate production when they have a strong bargaining position, relative to the host country government. When that bargaining position weakens, they tend to sell off assets to host country specialized firms.

The evolution of contractual relations in rapidly changing environments, such as transition economies, is another important area (see, e.g., Boger et al., 2001). These settings not only allow for comparative analysis but also provide insights into the ability of various contractual arrangements to adapt to changing circumstances.¹³

19.6 Conclusion

Despite the ongoing challenges described above, the transaction cost theory of the firm has had remarkable success in explaining the vertical structure of the enterprise. Indeed, the empirical literature on the make-or-buy decision is generally considered one of the best-developed parts of the new institutional economics. As Williamson (2000, p. 607) remarks, “Those who have done this modest, slow, molecular, definitive work deserve enormous credit.”

What lessons have we learned from this literature? First, asset specificity is an important determinant of vertical contractual relations. It is not the sole determinant, however; even in the face of uncertainty, arms-length contracting may effectively protect parties' relationship-specific investments. Tight vertical coordination or control may also generate efficiencies unrelated to the protection of specific assets such as improvements in knowledge flows, learning, the ability to adapt rapidly to technological and environmental shocks, and the exploitation of managerial or organizational capabilities. Still, the transaction cost approach remains central to such discussions and debates. No rival theory has produced a body of evidence rivaling the transaction cost explanation for vertical integration—a conclusion that has largely survived the “credibility revolution” in applied microeconomics that places central evidence on causal methods in empirical work.

At the same time, much new theoretical work in economics on firm boundaries builds on the incomplete-contracting framework of Grossman, Hart, and Moore or on knowledge-based and capabilities ideas, not the transaction cost framework of

¹³ For details regarding contracting in university-industry research agreements, see Bercovitz and Tyler (2014).

Williamson and Klein, Crawford, and Alchian. However, the ease of testing transaction cost models, relative to alternatives, makes them an attracting starting point for research on make-or-buy decisions.

The second lesson is that vertical relations are often subtle and complex. While early empirical work on transaction cost determinations of vertical integration tended to focus on black-and-white distinctions between “make” and “buy,” researchers increasingly recognize that a wide variety of contractual and organizational options are available; there are many shades of gray. The literature on hybrids has grown dramatically in the last 10 years, while there are fewer studies of mundane issues such as outsourcing versus in-house production per se. Moreover, as discussed above, the emergence of detailed and more comprehensive firm- and establishment-level data has revealed patterns that challenge our understanding of what vertical integration is and does. Starting with Coase (1937), transaction cost and related theories have assumed that the purpose of vertical integration is to facilitate transactions across production and distribution stages. Yet we know now that many vertically integrated firms share resources (such as management) without actually exchanging intermediate goods and services across their vertically related subunits. In this sense, there are two kinds of vertical integration: one that facilitates transactions (and for which TCE and related theories have substantial explanatory power) and another that results from firms expanding across vertical, horizontal, and conglomerate boundaries by adding subsidiaries that do not interact with each other (for which capabilities or knowledge-based explanations are needed).

Third, while we know much about the transaction cost determinants of vertical relations and increasing attention is being paid to the institutional environment, particularly in cross-country studies, more work is needed to understand the interactions across Williamson’s (2000) four levels of institutional analysis (social embeddedness, institutional environment, institutions of governance, and governance of contractual relations).

Acknowledgment of Funding Open access is made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., Johnson, S., & Mitton, T. (2009). Determinants of vertical integration: financial development and contracting costs. *The Journal of Finance*, 64(3), 1251–1290.
- Acheson, J. M. (1985). The Maine lobster market: Between market and hierarchy. *The Journal of Law, Economics, and Organization*, 1(2), 385–398.
- Akbar, Y. H., & Tracogna, A. (2022). The digital economy and the growth dynamics of sharing platforms: A transaction cost economics assessment. *Journal of Digital Economy*, 1(3), 209–226.
- Akerlof, G. A. (1970). The market for ‘lemons’: Qualitative uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84, 488–500.
- Alfaro, L., Conconi, P., Fadinger, H., & Newman, A. F. (2016). Do prices determine vertical integration? *The Review of Economic Studies*, 83(3), 855–888.
- Alfaro, L., Bloom, N., Conconi, P., Fadinger, H., Legros, P., Newman, A. F., Sadun, R., & Van Reenen, J. (2023). Come together: Firm boundaries and delegation. *Journal of the European Economic Association*, 22, 34–72.

- Anderson, E. (1985). The salesperson as outside agent or employee: A transaction cost analysis. *Marketing Science*, 4, 234–254.
- Anderson, E., & Schmittlein, D. C. (1984). Integration of the sales force: An empirical examination. *RAND Journal of Economics*, 15, 385–395.
- Anderson, E., & Weitz, B. (1992). The use of pledges to build and sustain commitment in distribution channels. *Journal of Marketing Research*, 29(1), 18–34.
- Atalay, E., Hortaçsu, A., & Syverson, C. (2014). Vertical integration and input flows. *American Economic Review*, 104(4), 1120–1148.
- Bai, G., & Ramakrishna Velamuri, S. (2021). Contextualizing the sharing economy. *Journal of Management Studies*, 58(4), 977–1001.
- Baker, G., Gibbons, R., & Murphy, K. J. (2001). Bringing the market inside the firm? *American Economic Review*, 91(2), 212–218.
- Bakhtiari, S. (2015). Productivity, outsourcing and exit: the case of Australian manufacturing. *Small Business Economics*, 44, 425–447.
- Bartel, A. P., Lach, S., & Sicherman, N. (2014). Technological change and the make-or-buy decision. *The Journal of Law, Economics, & Organization*, 30(1), 165–192.
- Beard, T. R., Ford, G. S., & Koutsy, T. M. (2005). Mandated access and the make-or-buy decision: The case of local telecommunications competition. *The Quarterly Review of Economics and Finance*, 45(1), 28–47.
- Belk, R. (2014). You are what you can access: Sharing and collaborative consumption online. *Journal of Business Research*, 67(8), 1595–1600.
- Bercovitz, J. E. L., & Tyler, B. B. (2014). Who I am and how I contract: The effect of contractors' roles on the evolution of contract structure in university–industry research agreements. *Organization Science*, 25(6), 1840–1859.
- Bigelow, L. (2001). *Efficient alignment and survival in the U.S. automobile industry* (Working paper). Olin School of Business, Washington University, St. Louis.
- Bindseil, U. (1997). Vertical integration in the long run: The provision of physical assets to the London and New York stock exchanges. *Journal of Institutional and Theoretical Economics*, 153, 641–656.
- Bindseil, U., & Pfeil, C. (1999). Specialization as a specific investment into the market: A transaction cost approach to the rise of markets and towns in medieval Germany, 800–1200. *Journal of Institutional and Theoretical Economics*, 155, 738–754.
- Boger, S., Hobbs, J. E., & Kerr, W. A. (2001). Supply chain relationships in the polish pork sector. *Supply Chain Management*, 6, 74–82.
- Bonina, C., Koskinen, K., Eaton, B., & Gawer, A. (2021). Digital platforms for development: Foundations and research agenda. *Information Systems Journal*, 31(6), 869–902.
- Bouncken, R. B., & Reuschl, A. J. (2018). Coworking-spaces: How a phenomenon of the sharing economy builds a novel trend for the workplace and for entrepreneurship. *Review of Managerial Science*, 12, 317–334.
- Brahm, F., & Tarziján, J. (2014). Transactional hazards, institutional change, and capabilities: Integrating the theories of the firm. *Strategic Management Journal*, 35(2), 224–245.
- Breza, E., & Liberman, A. (2017). Financial contracting and organizational form: Evidence from the regulation of trade credit. *The Journal of Finance*, 72(1), 291–324.
- Brown, J. R., Dev, C. S., & Lee, D. J. (2000). Managing marketing channel opportunism: The efficacy of alternative governance mechanisms. *Journal of Marketing*, 64, 51–65.
- Bucheli, M., Ciravegna, L., & Sáenz, L. F. (2023). The octopus that shrank: A historical analysis of how multinationals address policy and contractual uncertainty in a global value chain. *Strategic Management Journal*, 44(13), 3289–3323.
- Carbone, J., & Brinig, M. F. (1990). Rethinking marriage: Feminist ideology, economic change, and divorce reform. *Tul. L. Rev.*, 65, 953.
- Casadesus-Masanell, R., & Spulber, D. F. (2000). The fable of Fisher body. *Journal of Law and Economics*, 43, 67–104.
- Cassiman, B., & Valentini, G. (2016). Open innovation: Are inbound and outbound knowledge flows complementary? *Strategic Management Journal*, 37(6), 1034–1046.

- Caves, R. E., & Bradburd, R. E. (1988). The empirical determinants of vertical integration. *Journal of Economic Behavior and Organization*, 9, 265–279.
- Chen, Y., Richter, J. I., & Patel, P. C. (2021). Decentralized governance of digital platforms. *Journal of Management*, 47(5), 1305–1337.
- Chen, L., Tong, T. W., Tang, S., & Han, N. (2022). Governance and design of digital platforms: A review and future research directions on a meta-organization. *Journal of Management*, 48(1), 147–184.
- Coase, R. H. (1937). The nature of the firm. In idem. *The firm, the market and the law*. University of Chicago Press.
- Coase, R. H. (2000). The acquisition of Fisher Body by general motors. *Journal of Law and Economics*, 43, 15–31.
- Cohen, B., & Kietzmann, J. (2014). Ride on! Mobility business models for the sharing economy. *Organization & Environment*, 27(3), 279–296.
- Cooter, R., Marks, S., & Mnookin, R. (1982). Bargaining in the shadow of the law: A testable model of strategic behavior. *The Journal of Legal Studies*, 11, 225–251.
- Crocker, K. J., & Masten, S. E. (1988). Mitigating contractual hazards: Unilateral options and contract length. *RAND Journal of Economics*, 19, 327–343.
- Crocker, K. J., & Reynolds, K. J. (1993). The efficiency of incomplete contracts: An empirical analysis of air force engine procurement. *RAND Journal of Economics*, 24, 126–146.
- David, R. J., & Han, S. K. (2004). A systematic assessment of the empirical support for transaction cost economics. *Strategic Management Journal*, 25(1), 39–58.
- Davis, G. F., & DeWitt, T. (2021). Organization theory and the resource-based view of the firm: The great divide. *Journal of Management*, 47(7), 1684–1697.
- DeCanio, S. J., & Frech, H. E. (1993). Vertical contracts: A natural experiment in natural gas pipeline regulation. *Journal of Institutional and Theoretical Economics*, 149, 370–392.
- Delios, A., & Henisz, W. J. (2000). Japanese firms' investment strategies in emerging economies. *Academy of Management Journal*, 43, 305–323.
- de Souza Filho, H.M., & Varella Miranda, B. (2019). Asset specificity, intensity of coordination, and the choice of hybrid governance structures. *Journal of Agribusiness in Developing and Emerging Economies*, 9(2), 139–158.
- De Vita, G., Tekaya, A., & Wang, C. L. (2010). Asset specificity's impact on outsourcing relationship performance: A disaggregated analysis by buyer-supplier asset specificity dimensions. *Journal of Business Research*, 63(7), 657–666.
- Diez-Vial, I., & Alvarez-Suescun, E. (2011). The impact of geographical proximity on vertical integration through specific assets: The case of the Spanish meat industry. *Growth and Change*, 42(1), 1–22.
- Ellickson, R. C. (1989). A hypothesis of wealth-maximizing norms: Evidence from the whaling industry. *The Journal of Law, Economics, and Organization*, 5(1), 83–97.
- Ellickson, R. C. (1991). *Order without law: How neighbors settle disputes*. Harvard University Press.
- Fan, D., Yeung, A. C. L., Lo, C. K. Y., & Tang, C. (2022). Impact of the US–China trade war on the operating performance of US firms: The role of outsourcing and supply base complexity. *Journal of Operations Management*, 68(8), 928–962.
- Fein, A. J., & Anderson, E. (1997). Patterns of credible commitments: Territory and brand selectivity in industrial distribution channels. *Journal of Marketing*, 61, 19–34.
- Ferguson, S., & Formai, S. (2013). Institution-driven comparative advantage and organizational choice. *Journal of International Economics*, 90(1), 193–200.
- Fishback, P. V. (1986). Did coal miners 'owe their souls to the company store'? Theory and evidence from the early 1900s. *Journal of Economic History*, 46, 1011–1029.
- Fishback, P. V. (1992). The economics of company housing: Historical perspectives from the coal fields. *Journal of Law, Economics, and Organization*, 8, 346–365.
- Fisher, S. L., & Connelly, C. E. (2017). Lower cost or just lower value? Modeling the organizational costs and benefits of contingent work. *Academy of Management Discoveries*, 3(2), 165–186.
- Forbes, S. J., & Lederman, M. (2009). Adaptation and vertical integration in the airline industry. *American Economic Review*, 99(5), 1831–1849.

- Foss, N. J., Klein, P. G., Lien, L. B., Zellweger, T., & Zenger, T. (2021). Ownership competence. *Strategic Management Journal*, 42(2), 302–328.
- Freeland, R. F. (2000). Creating holdup through vertical integration: Fisher Body revisited. *Journal of Law and Economics*, 43, 33–66.
- Gallick, E. C. (1984). Exclusive dealing and vertical integration: The efficiency of contracts in the tuna industry. In *Federal Trade Commission Bureau of economics staff report*. Federal Trade Commission.
- Gambardella, A., & Panico, C. (2014). On the management of open innovation. *Research Policy*, 43(5), 903–913.
- Gartenberg, C., & Pierce, L. (2017). Subprime governance: Agency costs in vertically integrated banks and the 2008 mortgage crisis. *Strategic Management Journal*, 38(2), 300–321.
- Gatignon, H., & Anderson, E. (1988). The multinational Corporation's degree of control over foreign subsidiaries: An empirical test of a transaction cost explanation. *Journal of Law, Economics, and Organization*, 4, 305–336.
- Gawer, A. (2021). Digital platforms' boundaries: The interplay of firm scope, platform sides, and digital interfaces. *Long Range Planning*, 54(5), 102045.
- Gayle, P. G., & Weisman, D. L. (2007). Are input prices irrelevant for make-or-buy decisions? *Journal of Regulatory Economics*, 32, 195–207.
- Ghoshal, S., & Moran, P. (1996). Bad for practice: A critique of the transaction cost theory. *Academy of Management Review*, 21, 13–47.
- Gifford, A., Jr. (1993). The economic organization of 17th- through mid-19th-century whaling and shipping. *Journal of Economic Behavior and Organization*, 20, 137–150.
- Gil, R., & Ruzzier, C. A. (2018). The impact of competition on “make-or-buy” decisions: Evidence from the Spanish local TV industry. *Management Science*, 64(3), 1121–1135.
- Globerman, S., & Schwindt, R. (1986). The organization of vertically related transactions in the Canadian Forest products industries. *Journal of Economic Behavior and Organization*, 7, 199–212.
- González-Díaz, M., Arruñada, B., & Fernández, A. (2000). Causes of subcontracting: Evidence from panel data on construction firms. *Journal of Economic Behavior and Organization*, 42, 167–187.
- Gray, J. V., Roth, A. V., & Tomlin, B. (2009). The influence of cost and quality priorities on the propensity to outsource production. *Decision Sciences*, 40(4), 697–726.
- Guadalupe, M., & Wulf, J. (2010). The flattening firm and product market competition: The effect of trade liberalization on corporate hierarchies. *American Economic Journal: Applied Economics*, 2(4), 105–127.
- Hallwood, P. C. (1991). On choosing organizational arrangements: The example of offshore oil gathering. *Scottish Journal of Political Economy*, 38, 227–241.
- Halonon-Akatwijuka, M., & Hart, O. D. (2013). *More is less: Why parties may deliberately write incomplete contracts*. No. w19001. National Bureau of Economic Research.
- Hamilton, G. (1999). Property rights and transaction costs in marriage: Evidence from prenuptial contracts. *Journal of Economic History*, 59, 68–103.
- Harrigan, K. R. (1986). Matching vertical integration strategies to competitive conditions. *Strategic Management Journal*, 7, 535–555.
- Hart, O. D. (1995). *Firms, contracts, and financial structure*. Oxford University Press.
- Hart, O. (2009). Hold-up, asset ownership, and reference points. *The Quarterly Journal of Economics*, 124(1), 267–300.
- Hart, O., & Moore, J. (2008). Contracts as reference points. *The Quarterly Journal of Economics*, 123(1), 1–48.
- He, D., & Nickerson, J. A. (2006). Why do firms make and buy? Efficiency, appropriability and competition in the trucking industry. *Strategic Organization*, 4(1), 43–69.
- Heide, J. B., & John, G. (1992). The role of dependence balancing in safeguarding transaction specific assets. *Journal of Marketing*, 52, 20–35.
- Heide, J. B., Dutta, S., & Bergen, M. (1998). Exclusive dealing and business efficiency: Evidence from industry practice. *Journal of Law and Economics*, 41, 387–407.
- Hein, A., Schreieck, M., Riasanow, T., Setzke, D. S., Wiesche, M., Böhm, M., & Krcmar, H. (2020). Digital platform ecosystems. *Electronic Markets*, 30, 87–98.

- Helfat, C. E., & Campo-Rembado, M. A. (2016). Integrative capabilities, vertical integration, and innovation over successive technology lifecycles. *Organization Science*, 27(2), 249–264.
- Helper, S., MacDuffie, J. P., & Sabel, C. F. (2000). Pragmatic collaborations: Advancing knowledge while controlling opportunism. *Industrial and Corporate Change*, 9, 443–483.
- Henisz, W. J. (2000). The institutional environment for multinational investment. *Journal of Law, Economics, and Organization*, 16, 334–364.
- Henisz, W. J., & Zelner, B. A. (2001). The institutional environment for telecommunications investment. *Journal of Economics and Management Strategy*, 10, 123–147.
- Hennart, J.-F. (1988). Upstream vertical integration in the aluminum and tin industries. *Journal of Economic Behavior and Organization*, 9, 281–299.
- Hennart, J.-F. (1989). The transaction cost rationale for countertrade. *Journal of Law, Economics, and Organization*, 5, 127–153.
- Hu, M. Y., & Chen, H. (1993). Foreign ownership in Chinese joint ventures. *Journal of Business Research*, 26, 500–513.
- Hubbard, T. N. (1999). *How wide is the scope of hold-up based theories of governance? Shipper-carrier relations in trucking* (Working paper). Department of Economics, UCLA.
- Hughes, R. C. (2022). Regulatory entrepreneurship, fair competition, and obeying the law. *Journal of Business Ethics*, 181(1), 249–261.
- Huurne, T., Maarten, A. R., Corten, R., & Buskens, V. (2017). Antecedents of trust in the sharing economy: A systematic review. *Journal of Consumer Behaviour*, 16(6), 485–498.
- John, G., & Weitz, B. A. (1988). Forward integration into distribution: An empirical test of transaction cost analysis. *Journal of Law, Economics, and Organization*, 4, 337–355.
- Joskow, P. L. (1985). Vertical integration and long term contracts: The case of coal-burning electric generating plants. *Journal of Law, Economics, and Organization*, 1, 33–80.
- Joskow, P. L. (1987). Contract duration and relationship-specific investments: Empirical evidence from coal markets. *American Economic Review*, 77, 168–185.
- Joskow, P. L. (1988a). Asset specificity and the structure of vertical relationships: Empirical evidence. *Journal of Law, Economics, and Organization*, 4, 95–117.
- Joskow, P. L. (1988b). Price adjustment in long-term contracts: The case of coal. *Journal of Law and Economics*, 31, 47–83.
- Joskow, P. L. (1990). The performance of long-term contracts: Further evidence from the coal markets. *RAND Journal of Economics*, 21, 251–274.
- Joskow, P. L. (1991). The role of transaction cost economics in antitrust and public utility regulatory policies. *The Journal of Law, Economics, and Organization*, 7(special issue), 53–83.
- Kapoor, R. (2013). Persistence of integration in the face of specialization: How firms navigated the winds of disintegration and shaped the architecture of the semiconductor industry. *Organization Science*, 24(4), 1195–1213.
- Kapoor, R., & Adner, R. (2012). What firms make vs. what they know: how firms' production and knowledge boundaries affect competitive advantage in the face of technological change. *Organization Science*, 23(5), 1227–1248.
- Kenney, R. W., & Klein, B. (1983). The economics of block booking. *Journal of Law and Economics*, 26, 497–540.
- Kermani, A., & Ma, Y. (2023). Asset specificity of nonfinancial firms. *The Quarterly Journal of Economics*, 138(1), 205–264.
- Klein, B. (1988). Vertical integration as organized ownership: The Fisher Body-general motors relationship revisited. *Journal of Law, Economics, and Organization*, 4, 199–213.
- Klein, P. G. (2005). The make-or-buy decision: Lessons from empirical studies. In *Handbook of new institutional economics* (pp. 435–464). Springer Nature.
- Klein, P. G., & Saldenberg, M. R. (2010). Organizational structure and the diversification discount: Evidence from commercial banking. *The Journal of Industrial Economics*, 58(1), 127–155.
- Klein, B., Crawford, R. A., & Alchian, A. A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21, 297–326.
- Klein, S., Frazier, G. L., & Roth, V. J. (1990). A transaction cost analysis model of channel integration in international markets. *Journal of Marketing Research*, 27, 196–208.

- Kolympiris, C., Hoenen, S., & Klein, P. G. (2019). Learning by seconding: Evidence from national science foundation rotators. *Organization Science*, 30(3), 528–551.
- Lafontaine, F., & Masten, S. E. (2002). *Contracting in the absence of specific investments and moral hazard: Understanding carrier-driver relations in U.S. trucking* (Working paper). University of Michigan Business School.
- Lafontaine, F., & Raynaud, E. (2002). The role of residual claims and self enforcement in franchise contracting (NBER working paper 8868).
- Lafontaine, F., & Shaw, K. (1999). The dynamics of franchise contracting: Evidence from panel data. *Journal of Political Economy*, 107, 1041–1080.
- Lafontaine, F., & Shaw, K. L. (2005). Targeting managerial control: Evidence from franchising. *RAND Journal of Economics*, 36(1), 131–150.
- Lahiri, A., Doğan, O. B., & Kumar, V. (2023). Nurturing resource availability of ridesharing platforms by priming goal-oriented operational effectiveness. *Journal of Business Research*, 168, 114121.
- Langlois, R. N., & Robertson, P. L. (1989). Explaining vertical integration: Lessons from the American automobile industry. *Journal of Economic History*, 49, 361–375.
- Lee, R. S. (2013). Vertical integration and exclusivity in platform and two-sided markets. *American Economic Review*, 103(7), 2960–3000.
- Levy, D. (1985). The transaction cost approach to vertical integration: An empirical examination. *Review of Economics and Statistics*, 67, 438–445.
- Li, H.-L., & Tang, M.-J. (2010). Vertical integration and innovative performance: The effects of external knowledge sourcing modes. *Technovation*, 30(7–8), 401–410.
- Lieberman, M. B. (1991). Determinants of vertical integration: An empirical test. *Journal of Industrial Economics*, 39, 451–466.
- Liston, P., Byrne, J., Byrne, P. J., & Heavey, C. (2007). Contract costing in outsourcing enterprises: Exploring the benefits of discrete-event simulation. *International Journal of Production Economics*, 110(1–2), 97–114.
- Loredo, E., & Suárez, E. (2000). The governance of transactions: Joskow's coal-burning generating plants example revisited. *Energy Policy*, 28, 107–114.
- Lyons, B. R. (1995). Specific investment, economies of scale, and the make-or-buy decision: A test of transaction cost theory. *Journal of Economic Behavior and Organization*, 26, 431–443.
- Macher, J. T., & Richman, B. D. (2008). Transaction cost economics: An assessment of empirical research in the social sciences. *Business and Politics*, 10(1), 1–63.
- MacMillan, I. C., Hambrick, D. C., & Pennings, J. M. (1986). Uncertainty reduction and the threat of supplier retaliation: Two views of the backward integration decision. *Organization Studies*, 7, 263–278.
- Magelssen, C., Rich, B., & Mayer, K. (2022). The contractual governance of transactions within firms. *Organization Science*, 33(6), 2226–2249.
- Mandy, D. M. (2009). Pricing inputs to induce efficient make-or-buy decisions. *Journal of Regulatory Economics*, 36, 29–43.
- Maskin, E., & Tirole, J. (1999). Unforeseen contingencies and incomplete contracts. *The Review of Economic Studies*, 66(1), 83–114.
- Masten, S. E. (1984). The organization of production: Evidence from the aerospace industry. *Journal of Law and Economics*, 27, 403–417.
- Masten, S. E. (1994). *Empirical research in transaction-cost economics: Challenges, progress, directions*. University of Michigan Business School.
- Masten, S. E. (Ed.). (1996). *Case studies in contracting and organization*. Oxford University Press.
- Masten, S. E., & Saussier, S. (2000). Econometrics of contracts: An assessment of developments in the empirical literature on contracting. *Revue d'Économie Industrielle*, 92, 215–236.
- Masten, S. E., Meehan, J. W., & Snyder, E. A. (1989). Vertical integration in the U.S. auto industry: A note on the influence of specific assets. *Journal of Economic Behavior and Organization*, 12, 265–273.
- Masten, S. E., Meehan, J. W., & Snyder, E. A. (1991). The costs of organization. *Journal of Law, Economics, and Organization*, 7, 1–25.

- Masters, J. K., & Miles, G. (2002). Predicting the use of external labor arrangements: A test of the transaction cost perspective. *Academy of Management Journal*, 45, 431–442.
- Ménard, C. (2004). The economics of hybrid organizations. *Journal of Institutional and Theoretical Economics*, 160, 1–32.
- Ménard, C. (2013). Plural forms of organization: Where do we stand? *Managerial and Decision Economics*, 34(3–5), 124–139.
- Ménard, C., & Klein, P. G. (2004). Organizational issues in the agri-food sector: Toward a comparative approach. *American Journal of Agricultural Economics*, 86, 746–751.
- Monteverde, K., & Teece, D. J. (1982a). Appropriable rents and quasi-vertical integration. *Journal of Law and Economics*, 25, 321–328.
- Monteverde, K., & Teece, D. J. (1982b). Supplier switching costs and vertical integration in the automobile industry. *Bell Journal of Economics*, 13, 206–213.
- Moschuris, S. J. (2007). Triggering mechanisms in make-or-buy decisions: An empirical analysis. *Journal of Supply Chain Management*, 43(1), 40–49.
- Munger, M. C. (2021). *The sharing economy: Its pitfalls and promises*. Do Sustainability.
- Murcia, M. J., Panwar, R., & Tarzijan, J. (2021). Socially responsible firms outsource less. *Business & Society*, 60(6), 1507–1545.
- Muris, T. J., Scheffman, D., & Spiller, P. T. (1992). Strategy and transaction costs: The organization of distribution in the carbonated soft drink industry. *Journal of Economics and Management Strategy*, 1, 83–128.
- Nickerson, J. A., & Silverman, B. S. (2003). Why aren't all truck drivers owner operators? Asset ownership and the employment relation in interstate for-hire trucking. *Journal of Economics and Management Strategy*, 12, 91–118.
- Nickerson, J. A., Hamilton, B. H., & Wada, T. (2001). Market position, resource profile, and governance: Linking Porter and Williamson in the context of international courier and small package Services in Japan. *Strategic Management Journal*, 22, 251–273.
- Oskam, J., & Boswijk, A. (2016). Airbnb: The future of networked hospitality businesses. *Journal of Tourism Futures*, 2(1), 22–42.
- Oxley, J. E. (1997). Appropriability hazards and governance in strategic alliances: A transaction cost approach. *Journal of Law, Economics, and Organization*, 13, 387–409.
- Oxley, J. E. (1999). Institutional environment and the mechanisms of governance: The impact of intellectual property protection on the structure of inter-firm alliances. *Journal of Economic Behavior and Organization*, 38, 283–309.
- Palay, T. M. (1984). Comparative institutional economics: The governance of rail freight contracting. *The Journal of Legal Studies*, 13(2), 265–287.
- Palay, T. M. (1985). Avoiding regulatory constraints: Contracting safeguards and the role of informal agreements. *Journal of Law, Economics, and Organization*, 1, 155–176.
- Parmigiani, A. (2007). Why do firms both make and buy? An investigation of concurrent sourcing. *Strategic Management Journal*, 28(3), 285–311.
- Parmigiani, A., & Mitchell, W. (2009). Complementarity, capabilities, and the boundaries of the firm: The impact of within-firm and interfirm expertise on concurrent sourcing of complementary components. *Strategic Management Journal*, 30(10), 1065–1091.
- Pirrong, S. C. (1993). Contracting practices in bulk shipping markets: A transactions cost explanation. *Journal of Law and Economics*, 36, 937–976.
- Pisano, G. P. (1990). Using equity participation to support exchange: Evidence from the biotechnology industry. *Journal of Law, Economics, and Organization*, 5, 109–126.
- Poppo, L., & Zenger, T. (1995). Opportunism, routines, and boundary choices: A comparative test of transaction cost and resource-based explanations for make-or-buy decisions. *Academy of Management Journal, Best Papers Proceedings*, 38(1), 42–46.
- Poppo, L., & Zenger, T. (1998). Testing alternative theories of the firm: Transaction cost, knowledge-based, and measurement explanations for make-or-buy decisions in information services. *Strategic Management Journal*, 19, 853–877.
- Prete, D., Davide, G. G., & Marvasi, E. (2017). Global value chains participation and productivity gains for north African firms. *Review of World Economics*, 153(4), 675–701.

- Rawley, E., & Simcoe, T. S. (2013). Information technology, productivity, and asset ownership: Evidence from taxicab fleets. *Organization Science*, 24(3), 831–845.
- Regan, L. (1997). Vertical integration in the property-liability insurance industry: A transaction cost approach. *The Journal of Risk and Insurance*, 64, 41–62.
- Reitzig, M., & Wagner, S. (2010). The hidden costs of outsourcing: Evidence from patent data. *Strategic Management Journal*, 31(11), 1183–1201.
- Rindfleisch, A., & Heide, J. B. (1997). Transaction cost analysis: Past, present, and future applications. *Journal of Marketing*, 61, 30–54.
- Roy, R., & Sarkar, M. B. (2016). Knowledge, firm boundaries, and innovation: Mitigating the incumbent's curse during radical technological change. *Strategic Management Journal*, 37(5), 835–854.
- Sako, M., Chondrakis, G., & Vaaler, P. M. (2016). How do plural-sourcing firms make and buy? The impact of supplier portfolio design. *Organization Science*, 27(5), 1161–1182.
- Sanders, N. R., et al. (2007). A multidimensional framework for understanding outsourcing arrangements. *Journal of Supply Chain Management*, 43(4), 3–15.
- Saussier, S. (2000). Transaction costs and contractual incompleteness: The case of Electricité de France. *Journal of Economic Behavior and Organization*, 42, 189–206.
- Shelanski, H. A., & Klein, P. G. (1995). Empirical research in transaction cost economics: A review and assessment. *Journal of Law, Economics, and Organization*, 11, 335–361.
- Shi, C., Chen, Y., You, J., & Yao, H. (2018). Asset specificity and contractors' opportunistic behavior: Moderating roles of contract and trust. *Journal of Management in Engineering*, 34(5), 04018026.
- Silverman, B. S. (1999). Technological resources and the direction of corporate diversification: Toward an integration of the resource-based view and transaction cost economics. *Management Science*, 45, 1109–1124.
- Silverman, B. S., Nickerson, J. A., & Freeman, J. B. (1997). Profitability, transactional alignment, and organizational mortality in the U.S. trucking industry. *Strategic Management Journal*, 18(Special Issue), 31–52.
- Spiller, P. (1985). On vertical mergers. *Journal of Law, Economics, and Organization*, 1, 285–312.
- Stallkamp, M., & Schotter, A. P. J. (2021). Platforms without borders? The international strategies of digital platform firms. *Global Strategy Journal*, 11(1), 58–80.
- Stuckey, J. (1983). *Vertical integration and joint ventures in the aluminum industry*. Harvard University Press.
- Takata, H., & Parry, M. E. (2022). Human asset specificity, physical asset specificity, and direct distribution. *Industrial Marketing Management*, 105, 515–531.
- Teece, D. J. (2010). Forward integration and innovation: Transaction costs and beyond. *Journal of Retailing*, 86(3), 277–283.
- Ulset, S. (1996). R&D outsourcing and contractual governance: An empirical study of commercial R&D projects. *Journal of Economic Behavior and Organization*, 30, 63–82.
- Uygun, Y., Gotsadze, N., Schupp, F., Gzirishvili, L., & Nana, B. S. T. (2023). A holistic model for understanding the dynamics of outsourcing. *International Journal of Production Research*, 61(4), 1202–1232.
- Uzunca, B., Coen Rigtering, J. P., & Ozcan, P. (2018). Sharing and shaping: A cross-country comparison of how sharing economy firms shape their institutional environment to gain legitimacy. *Academy of Management Discoveries*, 4(3), 248–272.
- Vannoni, D. (2002). Empirical studies of vertical integration: The transaction cost orthodoxy. *International Review of Economics and Business*, 49, 113–141.
- Von Hirschhausen, C., & Neumann, A. (2008). Long-term contracts and asset specificity revisited: An empirical analysis of producer–importer relations in the natural gas industry. *Review of Industrial Organization*, 32, 131–143.
- Walker, G., & Weber, D. (1984). A transaction cost approach to make-or-buy decisions. *Administrative Science Quarterly*, 29, 373–391.

- Weiss, A. M., & Kurland, N. (1997). Holding distribution channel relationships together: The role of transaction-specific assets and length of prior relationships. *Organization Science*, 8(6), 612–623.
- Whinston, M. D. (2000). *On the transaction cost determinants of vertical integration* (Working paper). Department of Economics, Northwestern University.
- Wiggins, S. N., & Libecap, G. D. (1985). Oil field unitization: Commercial failure in the presence of imperfect information. *American Economic Review*, 75, 368–385.
- Williamson, O. E. (1975). *Markets and hierarchies: Analysis and antitrust implications*. Free Press.
- Williamson, O. E. (1985). *The economic institutions of capitalism*. Free Press.
- Williamson, O. E. (1991a). Comparative economic organization: The analysis of discrete structural alternatives. *Administrative Science Quarterly*, 36, 269–296.
- Williamson, O. E. (1991b). Strategizing, economizing, and economic organization. *Strategic Management Journal*, 23, 75–94.
- Williamson, O. E. (1996a). Economic organization: The case for candor. *Academy of Management Review*, 21, 48–57.
- Williamson, O. E. (1996b). *The mechanisms of governance*. Oxford University Press.
- Williamson, O. E. (2000). The new institutional economics: Taking stock, looking ahead. *Journal of Economic Literature*, 38, 595–613.
- Wilson, J. A. (1980). Adaptation to uncertainty and small numbers exchange: The New England fresh fish market. *Bell Journal of Economics*, 4, 491–504.
- Yarbrough, B. V., & Yarbrough, R. M. (1987). Institutions for the governance of opportunism in international trade. *Journal of Law, Economics, and Organization*, 3, 129–139.
- Zanarone, G. (2013). Contract adaptation under legal constraints. *The Journal of Law, Economics, & Organization*, 29(4), 799–834.
- Zhang, Q., Yan, B., & Huo, X. (2018). What are the effects of participation in production outsourcing? Evidence from Chinese apple farmers. *Sustainability*, 10(12), 4525.
- Zhao, H., Luo, Y., & Suh, T. (2004). Transaction cost determinants and ownership-based entry mode choice: A meta-analytical review. *Journal of International Business Studies*, 35, 524–544.
- Zhou, Y. M., & Wan, X. (2017). Product variety and vertical integration. *Strategic Management Journal*, 38(5), 1134–1150.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Hybrids and Other Challenges in Organizational Economics

20

Claude Ménard

Contents

20.1	Introduction	476
20.2	The Classic NIE Approach to the Tradeoff Among Organizations Revisited	477
20.2.1	Landmarks	478
20.2.2	The Heuristic Model: A Reminder	478
20.2.3	Extending the Initial Model: Integrating Rights More Explicitly	481
20.3	Hybrids: A Visit to the Zoo	484
20.3.1	The Variety of Species	484
20.3.2	A Typology	485
20.3.3	Why Go Hybrid?	487
20.4	Governance and Coordination of Hybrids	489
20.4.1	Contracts	490
20.4.2	Administrative Regimes	492
20.4.3	From Managerial Provisions to Managerial Practices	493
20.5	Some Challenges	494
20.5.1	Plural Forms	494
20.5.2	Competition Policies	495
20.5.3	Methodology	496
20.6	Conclusion	497
	References	498

Abstract

How can we explain the variety and resilience of non-standard arrangements through which agents organize the transactions that support their economic activities? This chapter proposes an extended and inclusive transaction cost framework that goes beyond the classic tradeoff between markets and hierarchies. Through the case of hybrid arrangements, this chapter argues that what differentiates hybrids and other organizational solutions are the allocation of rights in the organization of transactions and the constraints it imposes on their governance.

C. Ménard (✉)

Centre d'Economie de la Sorbonne, Pantheon-Sorbonne University, Paris, France

e-mail: claud.menard@univ-paris1.fr

© The Author(s) 2025

C. Ménard, M. M. Shirley (eds.), *Handbook of New Institutional Economics*,

https://doi.org/10.1007/978-3-031-50810-3_20

475

Special attention is paid to the mechanisms implemented to face disturbances and unexpected events and, more generally, to the many challenges the diversity of organizational solutions poses to theory as well as policy-making.

Keywords

Institutions · Organizations · Hybrids · Transaction costs · Property rights · Decision rights · Governance · Contracts · Management

20.1 Introduction

The analysis of the variety of modalities to organize transactions in a market economy has come a long way since the celebrated but long-ignored article by Coase on “The Nature of the Firm” (1937). True, “firms have invented far more ways to work together than organizational economics has so far expressed” (Baker et al., 2008: 146). However, there is an increasing acknowledgment of the richness in a market economy of “micro-institutions,” of which organizations form a substantial part. The multiplication of competing explanations for this variety sends a positive signal in that respect. “Transaction costs,” “agency,” “property rights,” “resource-based,” and “evolutionary” theories lead this now crowded field of investigation.

Within this diversity of approaches, the research program initiated by Coase (1937, 1960) and Williamson (1975, 1985, 1996) remains central and the source of continuing progress in our understanding of this complex world. Although alternative explanations have delivered significant contributions (see Klein and Mazzoni, this *Handbook*, ► Chap. 19), this chapter builds primarily on transaction cost economics (TCE) and its embeddedness in New Institutional Economics (NIE). This choice is motivated by TCE’s exceptional set of concepts integrated in a unified model that has opened the way to exploration of the variety of organizations.

The core concepts of property rights, transactions costs, and contracts delineate the golden triangle of NIE. Property rights are about the right to use goods and services or to transfer these rights (Alchian, 1965: 818). The existence of institutions to establish, delineate, and implement these rights is a prerequisite to making their transfer possible and effective. Hence, property rights are embedded into specific institutional settings (see Sect. 20.3 of this *Handbook*). The world of transactions is defined by the transfer of these rights “across technologically separable interface,” providing support to all economic activities (Williamson, 1996: 379). The resulting transaction costs are central to understand how parties organize these transfers. Finally, contracts occupy a special status in the NIE approach to organizations because they provide support to transactions, relaxing the constraints of bounded rationality, fixing schemes of references for future actions, and checking potential opportunistic behavior of participants to transactions. Contracts comprise “. . . two distinct elements: (a) rational planning of the transaction with careful provision for as many future contingencies as can be foreseen; and (b) the existence or use of actual or potential sanctions to induce performance of the exchange or to compensate for non-performance” (Macaulay, 1963: 31).

The integrative model built on these concepts to capture the variety of organizations that operate in a market economy was initially developed by Williamson (1975, 1985) and is well summarized in his *Mechanisms of Governance* (1996, Chap. 4). In the NIE perspective, organizations are understood as institutional arrangements through which a transaction or a set of interdependent transactions is decided and implemented. The terminology regarding these arrangements has somewhat fluctuated. Davis and North (1971: 6) introduced this concept through their now classic distinction between the “institutional environment” (the rules of the game) and “institutional arrangements” (the players of the game). Williamson (1985) disentangled this last concept, making a sharp distinction between markets, integrated organizations (“hierarchies”), and hybrids, a distinction later encapsulated as “mechanisms of governance” (1996). Recent contributions (e.g., Kunneke et al., 2021) revisited the concept of institutional arrangements to better integrate the growing set of organizational solutions identified in a rich empirical literature. They have qualified these arrangements as “micro-institutions” to signal that they are part of a general theory of institutions.

Notwithstanding their differences, these conceptualizations all acknowledge the many alternative ways to organize transactions in order to take advantage of the division of labor while economizing on bounded rationality and providing safeguards against contractual hazards. They also concur in their efforts to explain the variety and tradeoffs among alternative solutions through the concepts of the golden triangle. This chapter sticks to the term “organizations,” understood as designating those components of the micro-institutional dimension within which transactions are established, implemented, and enforced. And it mainly focuses on those “non-standard arrangements” (“neither markets nor hierarchies”) that Williamson (1996, Chap. 4) qualified as hybrids, thus complementing other contributions to this *Handbook* (particularly Williamson, ► Chap. 4; Joskow, ► Chap. 18; Klein and Mazzoni, ► Chap. 19).

The analysis is developed as follows. Section 20.2 briefly summarizes the now classic transaction cost explanation to the tradeoff among alternative organizations and extends the initial model to better integrate arrangements distinct from “markets” and “hierarchies.” Sect. 20.3 focuses on hybrids, pointing out the properties that make them a specific category of organizations. Section 20.4 pushes further the analysis of the modalities of governance and coordination that make hybrids specific and sustainable. Section 20.5 discusses some challenges this approach to organizations faces. Section 20.6 concludes.

20.2 The Classic NIE Approach to the Tradeoff Among Organizations Revisited

The NIE approach to organization took off in the 1970s and at the beginning of the 1980s (Sect. 20.2.1). It led to the heuristic model developed by Williamson (1985, Chap. 4; 1996, Chap. 3), briefly reviewed and discussed in Sect. 20.2.2. An extended model is then submitted to make explicit the types of rights involved in the organization of transactions, leading to a representation of alternative organizational solutions that integrates the three concepts delineating the golden triangle of NIE (Sect. 20.2.3).

20.2.1 Landmarks

Coase's pioneering article (1937) initiated the idea of a tradeoff between organizing transactions through markets versus within firms. Almost simultaneously, Barnard (1938) emphasized the coordinating role of "authority" as distinguishing firms from the price mechanism that characterizes markets. Simon (1951) modeled the role of authority in the employment relationship, while Coase (1960) showed how "the legal delimitation of rights," the transfer of these rights through transactions, and the costs associated with the way these operations are organized shape economic activities.¹

In the 1970s these ideas took roots in economics, shaping what progressively became known as the TCE branch of NIE. Williamson's "Vertical Integration" (1971) was a turning point, combining the role of transaction costs in the tradeoff between "markets" and "hierarchies" with contracts as a key organizational device. Not long after, Alchian and Demsetz (1972) challenged the Coasian approach and argued that firms do not fundamentally differ from markets: they both rely on a nexus of contracts and the resulting allocation of property rights. Almost simultaneously, a less quoted but influential article by Richardson (1972) emphasized the need to pay attention to the variety of ways through which firms coordinate. Arrow (1974) took a different stance, pushing issues of internal organization of firms higher on the agenda of economists. However, it is *Markets and Hierarchies* (Williamson, 1975) that really signaled the emergence of a new paradigm which became the micro-analytical branch of NIE, parallel to the macro-branch personified by Douglass North.² Williamson's 1975 book, complemented by a key article (1979), operationalized the intuition of Coase through a model linking transaction costs, contracts, and types of organization. Not long after, Klein et al. (1978) closed the decade by challenging the view that vertical integration and more generally the tradeoff between "make" and "buy" would be determined by a transaction costs minimizing strategy. In Klein et al. (1978), the motivation would rather be to prevent risks of hold-up when one party has to make highly specific investments that create interdependencies. A stream of research and controversies was born.

20.2.2 The Heuristic Model: A Reminder

Already sketched in Williamson (1975, 1979), what became the model of reference for most institutional analyses of organizations can be found in his *Economic Institutions of Capitalism* (1985), complemented by *Mechanisms of Governance* (1996).³

¹Numerous other authors have been influential in shaping this movement, e.g., Commons (1934), Hayek (1945), Malmgren (1961), Macaulay (1963), etc. References in this section (and in this chapter) do not pretend to exhaustivity.

²Significantly, Davis and North (1971) were published the same year as Williamson (1971).

³More on this model and its empirical implementation in ► Chaps. 18 (Joskow) and ► 19 (Klein and Mazzoni) of this *Handbook*.

The motivation for the model is the fundamental question raised by Coase: how can agents take advantage of the division of labor without losing the potential advantages of cooperation? Division of labor implies decomposition of tasks, which raises the issues of coordination, organizational modalities, and cost.⁴ Cooperation has to do with the behavior of agents and relates to incentives, i.e., those mechanisms that can make diversified agents efficiently complementing each other. The two concepts are distinct: even when cooperation prevails, coordination remains an issue.

To meet both challenges, agents must organize transactions: to take advantage of specialization, they need to trade rights to use goods and services. However, there are different ways to do so. Identifying, analyzing, and comparing costs of alternative organizational solutions are fundamental to understanding how economic systems work. Choosing the most appropriate and cost-minimizing solution is also a key concern for decision-makers. Indeed, in the world of positive transaction costs, all devices that are adopted to transfer rights consume resources. For instance, elaborating, negotiating, monitoring, and enforcing contracts involve significant costs (Dahlman, 1979).

Williamson's model (1985, Chap. 3) made the Coasian concept of transaction costs operational. It did so by identifying the sources of these costs, rooted in the "attributes" of a transaction: the specificity of assets required to make it possible, the uncertainties surrounding its implementation, and the frequency at which it happens. *Specificity of assets* comes from the value of investments that would be lost in alternative usages; it also creates mutual dependence, opening the possibility of ex post "hold-up" over the quasi-rent by one or some transactors (Klein et al., 1978; Alchian & Woodward, 1987: 114).⁵ *Uncertainties* may come out of agents' behavior, organizational deficiencies, flawed institutions, or the physical environment; uncertainties also generate costs through the constraints they impose on the establishment and enforcement of a transaction. Last, the *frequency* at which a transaction is repeated also likely affects its costs, although its effects are ambiguous.⁶ The impact of these attributes on the costs of organizing a transaction is summarized in Equation (20.1), where the signs indicate the predicted impact on transaction costs of a positive variation of each attribute:

⁴This Coasian-Williamsonian sequence differs from other stories, e.g., the one proposed by Demsetz (1988, 2011) who went so far as to qualify the Coasian perspective as misleading. In Demsetz's view, which shares similarities with the resource-based view, economies of scale and particularly those resulting from managerial knowledge are the main explanation to why firms may overcome markets. Demsetz considers mainstream economists to be wrong in seeing prices as a coordination mechanism: prices do not coordinate, they signal opportunities that allow decision-makers to coordinate (see also Hayek, 1945). Therefore the tradeoff is not between markets and hierarchies, but between firms and households: without economies of scale, production would be carried on by households; with economies of scale, firms prevail in organizing production, while markets connect households and firms.

⁵Coase vigorously challenged this emphasis on the risk of hold-up, and it remains a highly controversial issue, as well summarized in Coase (2000) versus Klein (2000).

⁶According to Williamson (1985: 76), "The frequency of a transaction matters because the more often it takes place, the more widely spread are the fixed costs establishing a non-market governance system," thus favoring integration. On the other hand, frequency might lower the costs of organizing transactions through markets.

$$TC = F \left(\underset{+}{AS}, \underset{+}{U}, \underset{-}{F} \right) \quad (20.1)$$

The underlying assumption is that although transaction costs might not be measurable directly, they can be approximated through these attributes. Most transactions involve a mix of them, the resulting costs depending on their respective weight. Taking on board these determinants and their variations led Williamson to explain the tradeoff among alternative organizational solutions through the “discrete alignment principle” (Williamson, 1985, Preface; this *Handbook*, ► Chap. 4). That is, the attributes of a transaction determine its costs, which leads transactors to seek the most appropriate, cost-minimizing solution. Hence, the selected organizational solution is ultimately a function of the attributes of the transaction (Figure 20.1).

Besides these costs, endogenous to the attributes of a transaction, other costs may emerge due to factors not inherent to a specific transaction. Thus, the “bounded rationality” of agents and their propensity to behave opportunistically or environmental uncertainties coming out of natural factors (e.g., climate changes) or institutional flaws (e.g., poorly defined contract law) may also generate hazards that push transaction costs upward.

In developed market economies, contracts play a key role in dealing with these issues and providing support to transactions and are a significant source of transaction costs. Williamson (1971: 115 sq.; 1985, Chap. 3) pioneered the introduction of contracts in the economic literature. He pointed out their role in reducing hazards stemming from the factors mentioned above but also their limitation and flaws. Indeed, the more complex is a transaction, the higher the costs of encapsulating its characteristics and constraints into clauses defined *ex ante* and that could facilitate *ex post* adaptations. Hence, the great majority of contracts are incomplete. Two major consequences follow. (1) The presence of non-contractibilities creates the need for complementary devices and mechanisms of governance: in other words, contracts tell only part of the story. (2) Since organizational solutions differ according to the intensity of transaction costs, contractual arrangements will differ as well. Hence the adoption and adaptation by Williamson (1985, Chap. 3, Sect. 2; 1996: 95 sq.) of a typology of contracts first introduced by Macneil (1974), and that distinguishes between classical contracts, typically commercial ones, which dominate market relationships; “forbearance,” which prevails in firms and makes hierarchy its “own court of ultimate appeal,” exercising “fiat, whereas markets cannot, to manage transactions” (Williamson, 1996: 27);⁷ and relational contracts, which complement

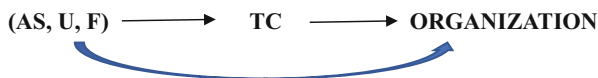


Fig. 20.1 From the attributes of a transaction to its organization

⁷See the early and nice modeling of the employment relationship in Simon (1951).

formal clauses with informal rules and mechanisms facilitating adaptability, making these contracts central to hybrids although they also play a role in other arrangements (more on contracts in Parts 3 and 4).

The resulting model, clearly expressed in Williamson (1996, Chap. 4), was intended to make Coase's intuition operational. Using it to test the "selection process" among alternative organizational solutions was a major challenge for this theory and is considered "a success story" (see already Whinston, 2003), substantiated by hundreds of contributions to what has been routinely identified as the "make-or-buy" tradeoff. Most initial empirical studies adopted reduced forms, focusing mostly on asset specificity as the independent variable to explain the choice of vertical integration (the dependent variable) as opposed to markets to organize the transaction at stake. Monteverde and Teece (1982) and Joskow (1985) initiated pioneering econometric studies, while Palay (1984, 1985) delivered nice qualitative studies. Shelanski and Klein (1995), Joskow (this *Handbook*, ► Chap. 18), and Klein (2005) delivered early surveys of this literature, while Masten et al. (1991) provided a clever technical discussion of the underlying research strategy and of some of the problems it raised.⁸

20.2.3 Extending the Initial Model: Integrating Rights More Explicitly

In the mid-1990s, two sides of the golden triangle, transaction costs and contracts, were already high on the research agenda of new institutionalists. However, although Coase (1960) had already emphasized the importance of rights and their legal dimension for understanding how a transactional system works and Williamson (1996: 379) had explicitly defined transactions as the transfer of rights, and notwithstanding some pioneering contributions on the importance of the mechanisms transferring rights (e.g., Barzel, 1989; Libecap & Wiggins, 1985), investigations of the role of property rights in the tradeoff among organizational solutions lagged behind. It is only recently, partially in relation to the booming literature on contracts, that the analysis of rights involved in transactions and their role has attracted more systematically the attention of NIE researchers (e.g., Allen, 2000; Arrunada, 2003; Baker et al., 2008; and for a partial review, Ménard & Shirley, 2022, Chap. 3).

Most new institutionalists endorse the Alchian view of property rights (1965: 818 sq.) as the socially enforced holding by an agent (or a set of agents) of rights to use, benefit, and transfer the usage of specific assets. As such, these rights determine the residual claimant over the value generated by this usage.⁹ This economic version

⁸ Another issue, discussed in Klein and Mazzoni (this *Handbook*, ► Chap. 19), concerns measurement in the transactional approach (see already Barzel, 1982).

⁹ The "old" PR approach focuses on the role of asset holders as coordinators, processing information to use resources efficiently, making them a residual claimant (Alchian & Demsetz, 1972). The "new" PR perspective rather emphasizes the incentive to invest that depends on the control through ownership over related assets and their usage by those not holding PR (Holmstrom, 1999: 76).

differs from the legal one in that by definition legal rights are formal rights that benefit from third-party enforcement (Smith, 2021), while economic rights are broader and capture different types of property rights (private, state, communal restricted to a community, or communal with open access) including situations in which parties may hold *de facto* rights not enforceable by the law or in which legal rights do not even exist because they are too difficult to establish and/or enforce.

In a NIE perspective, well-defined and enforceable property rights are critical to the efficient organization of transactions because they create incentives to use resources efficiently, reduce transaction costs, and diminish risks of conflicts among parties to the transaction. Property rights determine *who* controls what to do with available resources and the conditions under which rights over these resources can be transferred. By contrast, decision rights, which might be embedded in but are not identical to property rights, are about *how* these rights can actually be used, by whom, following what procedures (Bernstein, 2015; Gibbons & Henderson, 2012). Hence, decision rights are about delegating the exercise of property rights, which characterize and differentiate administrative mechanisms of governance (Williamson, 1996, Chap. 4), such as the right delegated to managers to reorganize tasks within the rules and constraints defined by holders of property rights (Aghion & Tirole, 1997; Roe, this *Handbook*, ► Chap. 21).

Building on these approaches, Ménard (2013a, 2018) proposed a model correlating both types of rights to transaction costs, therefore to the attributes of a transaction, thus determining the tradeoffs among alternative organizational solutions. If transaction costs are high, due to highly specific assets and/or high uncertainty, transactors will have a strong incentive to tightly control property rights and centrally monitor decision rights. Symmetrically, if transaction costs are low, thanks to assets that are easily redeployable and/or substitutable and to low levels of uncertainty, transactors have incentives to relax their hold on property rights and adopt decentralized solutions to the exercise of decision rights. In Fig. 20.2, the horizontal and vertical axes, respectively, reflect this impact of transaction costs on the incentives to centralize control over rights within an integrated organization (a “hierarchy” in Williamson vocabulary) or decentralize control over rights by rather relying on markets. Between these two polar situations is the wide area of hybrid arrangements. In this figure, (a, b) corresponds to the “Pareto frontier” of possible combinations of these modalities of allocation of rights, above which no transaction is implemented. The Macneil-Williamson typology of contracts mentioned above can enrich this representation. When classic contracts that characterize markets become “unfit” for “unprogrammed adaptation” and/or when the specificity of investments expose parties to significant contractual hazards, “a different contracting relation that preserves trading but provides for additional governance structure might be devised” (Williamson, 1985: 70). Parties to the transaction will turn to forbearance arrangements, in which “hierarchy is its own court of ultimate appeal” (Williamson, 1996: 98) if highly specific assets combine with significant uncertainties; with milder

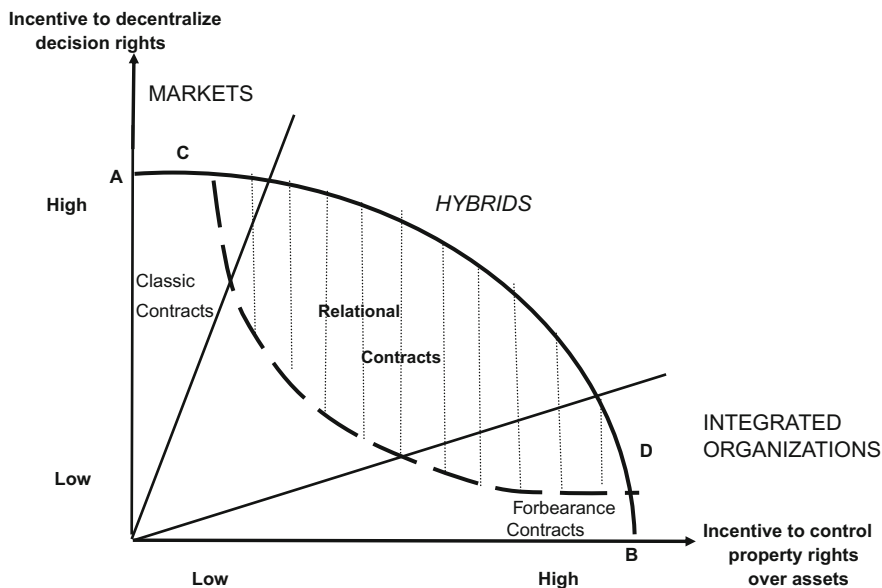


Fig. 20.2 Connecting decision rights and property rights to organizational arrangements. (Adapted from Ménard, 2013a, 2018)

specificity and/or moderate uncertainty, relational contracts will be adopted that permit smooth adaptations.¹⁰ The lower bound (c, d) then suggests the limits of the area below which relational contracts do not play a significant role (if any) since they would induce unstable relationships. Rays indicate boundaries between alternative organizational solutions and supporting contractual agreements, which are determined by the transaction costs coming out of the specificity of assets and the uncertainties characterizing a transaction. Thus, the figure and its underlying model suggest the respective role of contracts, property rights, and administrative control mechanisms that characterize and differentiate organizations (more on this in the next sections).

The two polar cases, “markets” and “hierarchies,” have focused most of the attention on TCE, at least until the beginning of this century. However, the multiplication of case studies on “non-standard” arrangements that do not fit these polar cases led to the increasing acknowledgment of the importance of hybrids and to the analysis of their properties (see surveys in Ménard, 2004, 2013a). I now turn to the analysis of these “non-standard” arrangements.

¹⁰ As suggested by Figure 20.2, relational contracts may play a role in markets that are imperfect or in integrated organizations when dealing with non-contractibilities (e.g., the employment relationship). Our model hypothesizes that relational contracts are particularly crucial to transactional situations in which parties can benefit from sharing part of their rights over the usage of some assets while they remain distinct and even compete on other dimensions, which is typical of hybrid.

20.3 Hybrids: A Visit to the Zoo

Until the late 1980s–early 1990s, only a handful of contributions developed a TCE perspective on inter-firm agreements (Klein et al., 1978; Rubin, 1978) or “non-standard arrangements” (Joskow, 1985; Masten, 1984; Palay, 1984). Simultaneously, a very rich empirical literature was developed in managerial publications, challenging this benign neglect. The publication of “Comparative Economic Organization” (Williamson, 1991/1996, Chap. 4) changed this situation by acknowledging and theorizing the status of hybrids as a specific category among the “generic forms of governance” (1996: 93). These arrangements would differ from markets as well as hierarchies by their “contract law regime,” their mode of adaptation to disturbances, the modalities of their administrative controls, and the intensity of their incentives. In what follows I review more explicitly what makes hybrids distinct, while the next section digs deeper in their modalities of governance and coordination.

20.3.1 The Variety of Species

The multiplication of publications on “non-standard” solutions to the organization of transactions in the 1980s–1990s dramatically changed the landscape in the economics of organization by showing the variety of arrangements involved. Beyond markets and “hierarchies” emerged a rather heteroclite set of “mechanisms of governance,”¹¹ among which (1) subcontracting (Eccles, 1981; Aoki, 1988, Chap. 6; Bajari & Tadelis, 2001), (2) networks of firms operating in similar or adjacent domains (Podolny & Page, 1998; Powell, 1990; Thorelli, 1986), (3) alliances among parties that remain partially rivals (Baker et al., 2002; Oxley, 1999), (4) franchising (Lafontaine & Slade, 1997; Rubin, 1978), (5) collective trademarks (Dwyer & Oh, 1988; Ménard, 1996), (6) partnerships (Farrell & Scotchmer, 1988; Powell, 1996), and (7) cooperatives (Cook, 1995). Many of these initial contributions connected only loosely (if at all) to TCE, at least until Williamson and his followers acknowledged that “the hybrid form of organization is not a loose amalgam of market and hierarchy but possesses its own disciplined rationale” (1996:119).

Subsequent developments substantiated what these arrangements share. What characterizes all of them is that they connect partners who remain legally distinct and keep their hold on key assets while they simultaneously jointly control decision rights and eventually some property rights over assets of a variable strategic value. Hence, hybrids can be understood as arrangements in which two or more partners pool strategic decision rights as well as some property rights while simultaneously keeping distinct ownership over key assets through which they may compete, so that they require specific devices to coordinate their joint activities and arbitrate the allocation of related payoffs (adapted from Ménard, 2013a). The strategic alliance between *Air France-KLM*, *Delta*, and *Virgin* provides an excellent illustration: it covers only some routes and with limited assets shared while the monitoring of

¹¹ For the first extensive survey, see Ménard (2004), updated in (2013).

the arrangement and of its resulting value-added requires complex mechanisms of coordination and control (Ménard & Damergy, 2023).

Hybrids therefore differ from markets in the combination of interdependent rights and tight coordination on some assets and related activities on the one hand and of competition on other segments of activities within the arrangement as well as with other hybrids and/or integrated firms on the other hand. They also differ in the strategic role of selecting partners (identity matters a lot) and implementing relational contracts and information-sharing procedures, all conditions essential to their stability. Dyer (1996) provided an early illustration with the case of the Japanese auto-makers and their well-structured relationships with suppliers of “first rank,” distinct from their suppliers of less significance. Simultaneously, Raynaud (1997) showed the complex organizational architecture monitoring the relational agreement between French millers who developed tight cooperation on a jointly owned brand while remaining strong competitors on other segments of their activities, with goods often produced using the same assets. As a consequence of their characteristics, there are blurred boundaries delineating participating entities and the need for specific mechanisms of governance that can monitor joint decisions, discipline partners, and solve conflicts while preventing free riding (see Sect. 20.4).

Hybrids also differ from integrated firms in that the legal as well as economic autonomy of partners with respect to key property rights prevents them from being subject in the last resort to a unified and centralized control, a strong contrast with, for example, the divisions of a firm. Even in their most tightly coordinated form, hybrids do not fit with the mainstream characterization of the firm as “a unitary profit-maximizing entity defined by a technologically determined production function” (Yarbrough & Yarbrough, 1988: 2). They also differ from the TCE concept of the firm-as-governance structure operating as a unique legal entity controlling the transfer of rights and relying in the last resort on the predominance of hierarchy for final decisions (see already Barnard, 1938; Simon, 1951; Beckmann, 1988; Radner, 1992). Hence, while firms face principal-agent problem, hybrids rather need to monitor principal-principal issues.

In sum, hybrids differ from firms by adapting to consequential disturbances through their “elastic contracting mechanisms” (Williamson, 1996: 96), and they differ from markets by their capacity to coordinate beyond the price mechanisms. However, in many cases they differ only by a small margin. For instance, Bernstein and Peterson (2023) provide rich illustrations about repeated interactions between buyers and suppliers leading beyond contractual clauses to the development of administrative provisions embedded in documents (e.g., supplier handbooks, scoreboards, etc.) and making room for inter-firm managerial adaptation, even in some cases to joint investments in specific assets.

20.3.2 A Typology

This comparative approach suggests the diversity of hybrid solutions to the organization of transactions, which led to the search for classifications that would capture and integrate this variety into a unified framework. Different criteria have been

explored to do so. Among them are distinction based on non-equity versus equity alliances (Oxley, 1997; Sampson, 2004), the degree of specificity of assets (Oxley & Silverman, 2008), the intensity of coordination among parties (US DoT and European Commission, 2010), and the types of decisions involved (Reuer & Klijn, 2018).

Building on these various contributions and elements reported in Sect. 20.2, Ménard (2013a, 2018) suggested an integrative typology, differentiating types of hybrids according to the intensity of their control over property rights to monitor joint assets and the intensity of their control over the exercise of decision rights to face uncertainties. On the left end of the hybrid section of the spectrum (see Fig. 20.3) are arrangements close to markets. Typical are *information-based networks* (Grandori & Soda, 2009; Powell, 1990, 1996) or the “Marshallian” industrial districts, built around “transactional reciprocity” (Park, 1996). These hybrids rely on relatively decentralized control over assets and high decentralization of decision rights. The “glue” sticking together partners to these relatively loose arrangements comes out of the trust developed through regular exchange of information, the benefits of complementary competences, and/or social connections (Powell, 1996; Robinson & Stuart, 2007). At the other end of the spectrum, hybrids come close to full integration, with partners delegating significant rights to *strategic centers* statutorily empowered to exercise tight control over shared assets, while decision rights become more centralized to face non-negligible uncertainties. The alliance between *Air France-KLM* and *Delta-Virgin* is an example. In between are the polymorphic (and often polycentric) hybrids in which each partner keeps full control over hard core assets while simultaneously developing relation-specific investments that require coordination of decision rights to face challenging uncertainties. Examples are hybrids coordinated by a *leader* (Hart et al., 2019; Sauvé et al., 2005), by a

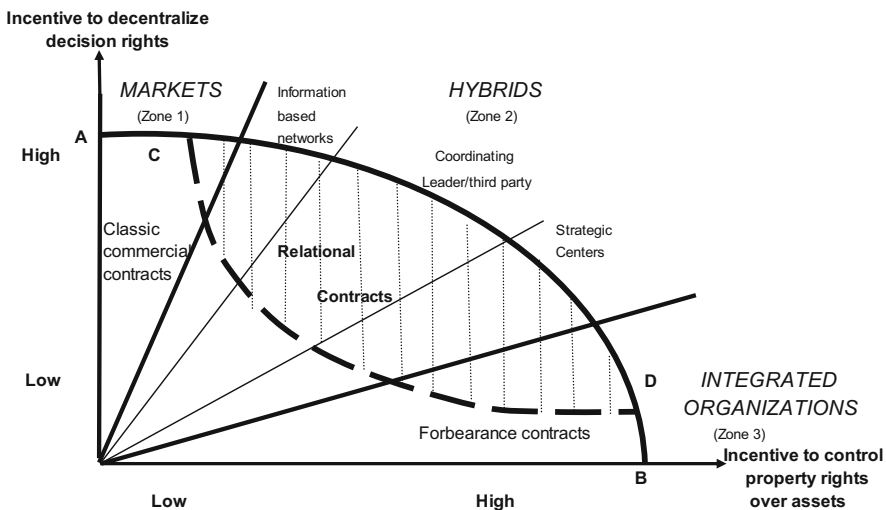


Fig. 20.3 Typology of hybrids

third party appointed to do so (e.g., “project managers,” as in Palay, 1985 or Bernstein, 2015), or through “private voluntary standards” (di Marcantonio & Ciaian, 2017). Figure 20.3 summarizes this typology.

This figure enriches Fig. 20.2 by differentiating the types of hybrids according to their governance while better qualifying the extension and limits of relational contracts (more on this in Sect. 20.4.1). The area delineated by (a, b) and (c, d) corresponds to the domain of effective relational contracts. Below the boundary (c, d), relational contracts would perform so poorly that they rather leave way to pure forbearance agreements (Zone 3) or to pure classic market contracts (Zone 1), depending on the degree of asset specificity and uncertainty and the associated intensity of control over the types of rights. In Zone 2, below (c, d), relational contracts are so inefficient that either partners give up or face high instability.

20.3.3 Why Go Hybrid?

Beyond their variety, hybrids raise the question: why do autonomous firms decide to engage in arrangements that create strong interdependencies and require continuing adaptation through interactions among partners that are often also competitors? Why choose the complexity of hybrids rather than the more standard solutions of markets (and the price mechanism) or integration (and coordination through hierarchy)? Stating as did Klein (1996) that hybrids exist because participants expect benefits exceeding the transaction costs involved does not tell the whole story. In an uncertain world, neither transaction costs nor benefits can be well assessed for the long-term relationship that characterizes hybrids. So, why go hybrid?

20.3.3.1 Pooling Strategic Resources

Building on lessons from the “make-or-buy” literature helps push the analysis further: joint investments in specific assets might allow partners to exploit their competitive capacities while reducing their costs of organizing transactions through markets. Two research strategies inherited from TCE can help explain the decision of firms to jointly invest in specific assets. One strategy emphasizes the advantages (as well as risks) of *complementary investments* by firms that maintain full control over their property rights while monitoring the interdependencies of their decision rights through contracts (Joskow, 1985; Klein et al., 1978; Masten, 1984, 1996; Palay, 1985). The famous controversy between Coase and Klein about why *General Motors* acquired *Fisher Body* in the mid-1920s rather than sticking to the existing contractual agreement comes out of this approach.¹² It raises the question of why contracts do not allow parties to cooperate while respecting their full autonomy. A different research strategy emphasizes the prospective value-added from *co-investments* in specific assets (Hennart, 1988; Lafontaine & Shaw, 1999; Powell,

¹²Extensive arguments are developed (including by Coase and Klein) in the *Journal of Law and Economics*, 2000: 43 (1):15–142.

1996). For the transactors, it means accepting more interdependencies, including over property rights, and building specific devices going beyond contracts to secure commitments, guarantee quality, and control risks of opportunism (Dwyer & Oh, 1988; John & Weitz, 1988; Ménard, 1996; Dyer, 1996). The co-investment views progressively prevailed, with an increasing interest in the modalities of governance adopted to deal with the resulting interdependency (see Sect. 20.4; and Ménard, 2013a).

An alternative explanation emerged at the intersection of TCE and the resource-based view (e.g., Gulati, 1998; Powell, 1996; Teece, 2010). In a changing environment in which important innovations require significant human, technical, and financial resources, sharing investments and competences can alleviate the burden of acting separately and deliver spillover effects. This strategy also involves risks. As already noted by Arrow (1962), sharing information or knowledge risks hold-up, whereby the receiving party benefits from the irreversibility of this transfer. As illustrated by the transfer of advanced technologies to a partner, this risk of predatory practices is particularly high when the institutions protecting property rights are weak or even non-existent (Oxley, 1999; Oxley & Silverman, 2008; Hennart & Verbeke, 2022: 1562–63).¹³ As the intensity of related uncertainties increases, parties will opt for increasingly centralized hybrid governance (see Oxley, 1997, Gambardella & Panico, 2014, for examples; also Fig. 20.2 and Sect. 20.4).

20.3.3.2 Monitoring Uncertainties

This brings up the issue of uncertainties. The discussion above pointed out the negative externalities of pooling strategic resources because of the risks of opportunistic behavior or because parties may misunderstand or misinterpret contractual clauses defining mutual responsibilities (Argyres & Zenger, 2012). However, going hybrid might also help partners circumvent uncertainties and deal with non-contractibilities, in the following ways: (1) A transactor can select high-valued partners through sophisticated screening procedures, repeated transactions, or reputation. (2) Parties can adopt modalities of governance to share information and facilitate unprogrammed adaptation without infringing on those activities on which they remain competitors. (3) Parties can implement devices and mechanisms to support joint planning and coordination (see Sect. 20.4). In sum, parties can gain from going hybrid by building coordination through a well-designed venture that provides “buffers” to external shocks and eases adaptation to unprogrammed events while securing cooperation by framing co-commitments.

An important source of potential uncertainties within hybrids comes from the difficulty of establishing rules to allocate rents arising from the joint use of pooled resources that concerns only part of their activities (and their assets). The incentive to accept limited dependence comes from estimations that markets cannot adequately bundle the relevant resources and capabilities (Teece & Pisano, 1994), while

¹³ One advantage of this approach is to connect organizational issues with their broad institutional environment.

integration would increase administrative costs, reduce flexibility, create irreversibility, and weaken incentives. However, cooperating on some dimensions while competing on others might easily lead to destabilizing decisions or open room to capture rents generated by the agreement, particularly in multilateral partnerships (e.g., supply chains, producers' associations). This problem formally shares similarities with the problem of remuneration in a multi-task model (Holmstrom & Milgrom, 1991). However, the hybrid case differs from the case of multi-task within a firm in that it is not a principal-agent issue (with the related problem of appropriate incentives) but rather a principal-principal problem (with the related problem of appropriate governance). To deal with the non-contractibilities in joint-ownership and usage of specific assets while keeping transaction costs low enough, hybrids implement rules and procedures about how to assess and share rents in order to guarantee the continuity of cooperation. The literature on this issue remains limited, partially because of the reluctance of participating firms to communicate information on such sensitive questions (but see Knoeber, 1989; Lafontaine & Bhattacharyya, 1995; Ménard, 1996).

The difficulties to establish such rules might explain why partners tend to adopt very simple sharing rules and delegate their implementation to managers with a lot of discretionary power. For example, the *Air France-KLM*, *Delta*, and *Virgin* alliance adopted a simple 40-40-20 sharing rule for the revenues from the North Atlantic routes concerned by their joint agreement (Ménard & Damergy, 2023). Similar observations have been made with franchising: franchisors adopt linear formulae based on observable components, typically a lump sum plus royalties based on sales (Lafontaine & Slade, 2007, 2013). Similarly, joint research projects in advance technology in the defense industry often rely on a 50-50 rule to share risks...and gains (Ménard & Oudot, 2009).

Other factors may play a role in the decision to go hybrid, for example, restrictive regulations prompt hybridization in air transportation. Although much remains to be done to explain the complex logic pushing parties to endorse hybrid solutions, what is striking is their resilience. Indeed, notwithstanding transaction costs coming out of these obstacles and challenges, hybrids show remarkable adaptability, as illustrated by the capacity of air alliances to survive the financial crisis of 2008 and the even more dramatic consequences of the COVID-19 crisis on air traffic or by the adaptability of franchising to different institutional environments.

20.4 Governance and Coordination of Hybrids

Hybrid resilience largely depends on the rich modalities of governance adopted to monitor agreements. Governance in this context is understood as the set of devices (e.g., joint committees) and mechanisms (e.g., protocols) through which parties implement their decisions, monitor their joint activities, and oversee or even control areas overlapping with independent transactions developed by partners with third parties. In that respect it is no surprise that the governance structure of many hybrids is quite

complex, going far beyond the inter-firm coordination mentioned by Richardson in his pioneering article (1972) or even in more recent research (e.g., Bernstein, 2015).

To allow unprogrammed adaptation without threatening their sustainability, hybrids must meet demanding conditions. First, if co-investments are to be made, rights must be allocated *ex ante* in ways that provides appropriate incentive, hence the importance of contracts as “legally binding promises” (Macneil, 1974: 693). Second, shared rights must be monitored *ex post* through forms of administrative controls in a way that allows adaptation to unexpected disturbances. Beyond the stabilizing effect of expected gains (Baker et al., 2008; Klein, 1996), which remain uncertain and challenged by non-contractible factors, there is a need for enforcement mechanisms embedded in specific devices, as suggested by Fig. 20.3.

20.4.1 Contracts

Almost simultaneously, Williamson (1971) and Alchian and Demsetz (1972) introduced the role of contracts as central to support transactions and coordinate decisions, making them consubstantial to all types of organizations. However, Alchian and Demsetz pushed further, arguing that all contracts share the same fundamental properties, which led Klein to conclude that “[T]hinking of all organizations as a group of explicit and implicit contracts among owners of factors of production represents a fundamental advance” (1983: 373). Although most new institutional economists share this view that contracts provide a fundamental support to transactions, whatever their mode of organization, they do not endorse the idea that all contracts are the same. Following Macneil-Williamson, it has been argued that in continuity with the differences between markets, hierarchies, and hybrids, contracts differ, whether they are the classic ones prevailing on spot markets, the forbearance forms characterizing firms, or relational contracts that, although transversal, are of particular significance in hybrids.¹⁴

There is some divergence among scholars about the nature of relational contracts, between an interpretation that focuses on *ex ante* contractual design, with the resulting contractual agreements being self-enforcing (Baker et al., 2002;¹⁵ Klein, 1996), and the Williamsonian view that *ex post* adaptation remains central for most contracts that are “unavoidably incomplete.” This chapter endorses the Macneil-Williamsonian interpretation, according to which contracts “are subject to (1) disturbances (or contingencies) during the contract implementation phase for which (2) defection from the spirit of cooperation to insist on the letter of the contract can be projected for outliers (where the stakes are great), and (3) inability of the courts to fill gaps and settle disputes in a timely, knowledgeable, and efficient fashion” (Tadelis & Williamson, 2013: 165). To face these challenges, a relational

¹⁴For detailed arguments about contracts in hybrids, see Ménard, 2013a. For a well-documented argument about the differences between market (commercial) contracts and employment contracts within the firm, see Masten, 1988.

¹⁵This interpretation is developed by Gil and Zanmarone, this *Handbook*, ► Chap. 25.

contract differs from the more formal “classic” commercial contract in that it “(1) contemplates unanticipated disturbances for which adaptation is needed, (2) provides a tolerance zone . . . within which misalignments will be absorbed, (3) requires information disclosure and substantiation if adaptation is proposed, and (4) provides for arbitration in the event voluntary agreement fails” (Williamson, 1996: 96). As already pointed out by Macaulay (1963), the adoption and adaptation of contracts of the relational type (Macaulay did not use this term) rely on a combination of formal clauses and informal commitment established through “handshakes,” “word commitments,” etc., leaving plenty of room for non-contractible adjustments. Indeed, although relational contracts share with other types of contracts the purpose of organizing and securing transactions, they differ in that they combine formal clauses with ex post governance exercised through devices designed to “informally” monitor the joint planning and coordination of shared assets and decision rights. As such, relational contracts are “a substitute for enforcement by courts, not a substitute for careful planning.” (Malcolmson, 2013: 1059). In hybrids, the resulting governance intends to facilitate unprogrammed adaptation without infringing too much on the internal practices of partners who remain independent or even also competitors. In sum, relational contracts provide essentially a *blueprint* to ease cooperation and secure the expected gains from co-specialized investments, opening the way to “relational governance” (Poppo & Zenger, 2002) built on an intricate set of jointly defined rules and norms to be implemented through specific governance devices.

Behind this concept of contracts as a *blueprint*, there is an assumption that has fed controversies: contracts are “unavoidably incomplete.” The debate goes back to the development of general equilibrium models and the conditions to be fulfilled to reach that state. Following Debreu (1959) and Arrow and Hahn (1971), numerous models were developed that argued that contracts can be complete, either by making them short-term renewable and renegotiable, making them adaptable, or introducing provisions to face unpredictable events, for example, automatic adjustments formula. By contrast, Williamson (e.g., 1971, 1996) and followers consider that such mechanisms involve significant transaction costs, making most contracts incomplete and in need of being complemented by other modalities of governance.¹⁶

This debate is not purely academic. Completeness suggests that contracts can provide the adequate unit for analyzing organizations, with all relevant properties specified in the ex ante clauses, making ex post governance irrelevant as long as contracts are enforced. Defining an optimal (complete) contracts then becomes the priority for decision-makers, whether in private organization or for public authorities (e.g., in privatizing utilities). However, if contracts are unavoidably incomplete, the institutional conditions of their monitoring and enforcement become central: governance implies much more than following what is in the contract. Most contracts

¹⁶The “incomplete contract” theory *à la Hart* (e.g., Hart, 1989) also acknowledges incompleteness but essentially ex ante so that optimal contracts can be drawn that allow ex post adjustments, ultimately making them self-fulfilling. For an early challenge of the “incomplete” hypothesis, see Maskin & Tirole, 1999.

become inherently “relational,” with an active role for managers in continuously adapting organizations to internal changes or to unexpected external events (Gibbons & Henderson, 2012; Ménard & Damergy, 2023).

20.4.2 Administrative Regimes

Indeed, if contracts are incomplete, the governance of different types of organizations requires devices and mechanisms that define different “administrative regimes” (Williamson, 1996, Chap. 4). This administrative component, complementing the role of contracts, refers to the nature, status, and role of management as one of the criteria differentiating markets, hybrids, and hierarchies. Drawing lessons from the analysis of *alliances*, which is actually a typical hybrid, Reuer and Klijn (2018: 7) made a nice distinction between three types of decisions: *alliance investment decisions*, which are about the decision to go hybrid (or not); *alliance type decisions*, which are about the type of hybrid to select; and *alliance design decisions*, which are about the administrative regime to be implemented (italics are from the authors).

These contributions (among others) raise the questions of what are the administrative regimes of hybrid and how they differ from those in integrated organizations. Building on indications by Klein et al. (1978), Williamson (1996), and numerous contributions of the late 1980s–early 1990s on “non-standard” arrangements, Ménard (1997, 1996, 2004) provided early evidence of various regulating devices as core elements in the organizational architecture of hybrids. Figure 20.3 synthesized these indications by emphasizing three types of administrative regimes in hybrids: information-sharing mechanisms, coordination by a leader or through a distinct third party, or coordination and monitoring of joint activities through a strategic center. Empirical studies have shown different facets of these regimes, from joint committees meeting regularly to trade information to permanent “working groups,” the creation of steering committees or of autonomous “joint strategic center” (see Hennart & Verbeke, 2022; Li & Matouschek, 2013; Ménard & Damergy, 2023; Reuer & Devaradonka, 2016). Notwithstanding their differences, these devices share one characteristic: they signal the transfer by partners of sub-classes of decisions to coordinating entities while retaining control over more or less significant rights.

Thus, hybrids develop well-structured administrative regimes based on *intentionality* and *mutuality*, maintaining a formal symmetry among parties. This makes hybrid agreements different from employment relationships (Masten, 1988) as well as from buyer-supplier relationships (Bernstein & Peterson, 2023), both of which depend on centralized control in the final resort. Part of these administrative regimes can be embedded in the formal contract, e.g., clauses creating a steering committee, imposing constraints on the selection, monitoring of suppliers, etc. (Bernstein, 2015; Hart et al., 2019; Ménard, 2013a). For example, in order to secure quality and maintain the reputation of its premium brand, a large corporation pushed members of its supply chain of fresh vegetables to form a well-structured organization that would select delegates to negotiate contracts and monitor their adjustments

(Sauvée et al., 2005). In this case and similar one (see Bernstein & Peterson, 2023; Gibbons & Henderson, 2013; Ménard & Damergy, 2023), managers play a pivotal role. These cases raise the question, rarely studied so far (but see Gil and Zanaron, this *Handbook*, ► Chap. 25), of the articulation between the formal and relational dimensions of contracts.

20.4.3 From Managerial Provisions to Managerial Practices

Indeed, managers are central within hybrids as they are in integrated firms, because of their role in monitoring transactions when contracts are incomplete, with “blanks” to be filled in order to adapt. However, the managers’ role has some aspects specific to hybrids because they must deal with counterparts that depend on different holders of rights to accomplish their tasks, thus operating in a principal-principal relationship. Notwithstanding rare exceptions (e.g., Demsetz, 1988; Radner, 1992; Milgrom & Roberts, 1995; and more recently Gibbons & Henderson, 2012, 2013; Gibbons & Roberts, 2013), it is striking how little attention leading economists (including neo-institutionalists) have paid to managerial functions, “as if ” managers were a minor factor in understanding the sources of productivity and performance of all types of organizational arrangements. A notable and early exception in NIE is Palay (1985) who pointed out how the acquisition of specific knowledge to monitor inter-firm relationships requires time and efforts, making the depositories of this knowledge (the “go-betweens”) highly regarded in inter-firm agreements as problem-solvers and key to secure the continuity of the relationship.

To capture the complex role of managers, it might be useful to distinguish between managerial provisions and managerial practices. Managerial provisions correspond to those formal devices and mechanisms designed *ex ante*, often in the contract between property rights holders and managers, specifying tasks and responsibilities of the latter (e.g., defining working groups and their codes of conduct in the *Air France-KLM/Delta-Virgin* agreement, Ménard & Damergy, 2023; establishing scoreboards to assess managers’ performance, Gibbons & Henderson, 2012; etc.). By contrast, managerial practices relate to those informal “initiatives” (Gibbons & Henderson, 2013) developed over time by managers for “preventing non-performance or detecting it early through buyer regulation of and involvement in suppliers’ design, sourcing, testing, and production processes rather than on merely deterring non-performance or giving the buyer the right to sue for damages in court” (Bernstein & Peterson, 2023: 12; Hart et al., 2019).

An interesting aspect of these links between formal clauses and relational relationships is the continuity pointed out by some empirical analyses between what is going on within the firm and the governance of external relations, either in procurement or in a hybrid relation. The now often-quoted case of *Lincoln Electric Co.* illustrates. Milgrom and Roberts (1995) referred to this example to discuss the role of organizational incentives to solve the principal-agent problem. Decades later, Bernstein and Peterson (2023) went a step further showing how the company extended its

successful internal governance to its relation with its suppliers, raising the interesting question of how much internal contracts and rules permeate relations with outsiders (Blader et al., 2020; Gibbons & Henderson, 2012). Indeed, the company developed practices summarized in a document on “Supplier Guidelines & Expectations” that provides criteria based on its internal experience to measure performance and determines its business with external suppliers. These indicators are primarily used for discussion at regular business review meetings, bringing the buyer-supplier relationship close to a hybrid arrangement. Besides providing indications on managerial practices, the *Lincoln* case suggests that organizational culture plays an important role in determining the openness to alternative ways of doing things as well as the capacity to build trust (Bernstein, 2015; Dyer & Chu, 2000). This may help explain the stability and resilience of hybrids, an issue that requires more investigation. However, lessons coming from the buyer-supplier relationship remain limited because in the last resort, decisions in these arrangements are made under “unified control,” thus falling under the rule of “forbearance.”

In hybrids, in the absence of forbearance, in which a hierarchy ultimately makes the decision, and with ownership distributed over autonomous partners, non-contractible factors and/or unexpected events raise problems distinct from the principal-agent relationship. Among them are the following: (1) How does the organization of tasks and allocation of responsibilities in the polycentric decision structure that characterizes hybrids differ from the decision process in integrated/hierarchical firms? (2) What are the decision mechanisms, for example, voting rules, specific to organizational solutions in which parties share some assets while using other assets (and in some cases the same assets) to compete? (3) Since administrative costs likely differ across organizational arrangements, under what conditions do hybrids outperform integrated firms? (4) Does going hybrid improve access to financial resources, which could explain this outperformance? Again, much research is needed on these issues.

20.5 Some Challenges

There are many other challenges raised by the nature of hybrids as well as by their role and sustainability in a competitive market economy. I briefly refer to three of these challenges here.

20.5.1 Plural Forms

One difficult issue concerns the explanation of what Bradach and Eccles (1989: 112; also, Bradach, 1997) have identified as “plural forms” in the organizational landscape. This expression might designate two very different situations. In one configuration, competing organizational solutions (e.g., a firm and a hybrid) persistently coexist in the same sector, delivering homogeneous products, using similar technologies, and mobilizing comparable assets so that the transactions at

stake are quite similar (Ménard, 2004: 26). Since these arrangements operate in competitive markets, what could explain their respective resilience? In another configuration, the same firm organizes for identical activities part of its transactions in-house and part through inter-firm contracts, a phenomenon well-documented in the case of franchising (Lafontaine & Slade, 2007). What is the incentive to do so?

The plural forms can be found in many different sectors and different institutional environments (Ménard, 2013b). It challenges the “either...or” approach of TCE without a convincing alternative explanation (see the discussion in Lafontaine & Slade, 2007; also 2013). The question of why firms operating in the same institutional environment choose these complex arrangements still requires in-depth exploration.¹⁷

20.5.2 Competition Policies

The very existence of hybrids as well as of plural forms and the resulting complexity of a micro-institutional landscape populated by this variety of organizational arrangements also challenge competition policies (for an early discussion, see Ménard, 1998). Indeed, at the theoretical level, the existing policies remain largely based on the dualism of markets versus integrated firms (or “hierarchies”) as illustrated by the usual reference to the criteria of “market power” as well as “freedom of entry.”¹⁸

Hybrids do not fit well in the standard approach of regulatory authorities that assess anti-competitive cases essentially through the examination of the behavior of targeted firms (e.g., is a merger or acquisition motivated by a strategy to control the relevant market?) and/or of the existing market structure (e.g., does it unduly favor one firm, creating obstacles to entry of newcomers?). Because participants in hybrids share some rights and cooperate accordingly (e.g., accepting geographical restrictions) while simultaneously competing along other dimensions (e.g., fighting to attract buyers from the same market), these arrangements challenge well-established policies. Fluctuating decisions by competition authorities about how to determine unfair practices (see di Marcantonio & Ciaian, 2017) or even the rules that should apply are well illustrated by the changing rules of the European Union regarding producers’ organization in the agri-food sector.

Making room for hybrids as an unavoidable part of the organizational landscape and adapting competition policies accordingly should become high on the agenda of policy-makers as well as researchers.

¹⁷These situations differ from that of firms choosing different forms because they operate in different institutional environments (see Oxley, 1997).

¹⁸And so does the approach to platforms. See Peitz, this *Handbook*, ► Chap. 23.

20.5.3 Methodology

Lastly, this brief review of some challenges faced by the economics of organization, particularly when hybrids are introduced into the landscape, cannot avoid pointing out some key methodological ones.¹⁹

20.5.3.1 Choosing an Organizational Solution

As noted by Klein and Mazzoni (this *Handbook*, ► [Chap. 19](#)), the initial empirical studies focused on the decision to “make,” “buy,” or (to a lesser degree) go “hybrid,” in line with the model summarized in Sect. 20.2. Challenges to this research strategy mostly focused on the nature, role, and measurement of “asset specificity” as the most frequently used independent variable (for a critical discussion, see de Vita et al., 2011). Among the most questioned issue has been the reliance on datasets built on surveys, interviews, or case studies. However, notwithstanding their limitations, results across industries and countries from hundreds of tests based on such data have been remarkably consistent with the core argument (Klein and Mazzoni, id.; Hennart & Verbeke, 2022).

Another difficulty was raised early on by Masten et al. (1991) who pointed out the challenge of using the same variable(s) to compare organizational solutions that widely differ or, even worse, to compare a chosen form to a hypothetical one that could have been chosen. More recent contributions have extended the discussion to issues like making causal inference based on a reduced-form model or assuming it is appropriate to consider which organizational solution is most efficient for a transaction having distinct attributes (Klein and Massoni, id.).

Nevertheless, recent developments that are well illustrated in several chapters of this *Handbook* show that TCE remain central to the analysis of the tradeoffs among organizational solutions and more generally to the research agenda of the economics of organization (see Hennart & Verbeke, 2022; Ménard, 2022; Reuer & Klijn, 2018).

20.5.3.2 Contracting

The importance of relational contracts in the organization of transactions is now well-acknowledged (see Macchiavello, this *Handbook*, ► [Chap. 24](#)). This chapter has emphasized their particular significance for hybrids as regulators of transactions among parties that share some strategic rights while keeping separate full control over others. What makes these arrangements even more complex to understand is that they are most of the time multilateral. Nevertheless, empirical studies of relational contracts in hybrids face problems similar to those in other organizational solutions. Among them are the difficulties pointed out in recent contributions (this *Handbook*, Macchiavello, ► [Chap. 24](#); Gil and Zanarone, ► [Chap. 25](#)), e.g., the difficulty of differentiating relational transactions from repeated ones, of understanding the articulation between the formal and the relational dimension, of assessing the expected value (and its sustainability) that motivates partners to stick to their mutual

¹⁹Other methodological challenges are discussed in this *Handbook*, particularly in Section VIII.

commitment, and of identifying the opportunity cost of deviating. And there is of course the core problem of getting appropriate data, one obstacle specific to the economics of organization being business' "secrets." New methodologies, many initially developed outside economics, may help circumvent these difficulties in the future (see Section VIII of this *Handbook*).

20.5.3.3 Administrating

One last challenge that deserves particular mention comes from the lack of attention to and information on the internal governance of hybrids. This issue, although at the core of "administrative regimes" and their differences, nevertheless remains largely neglected by economists, notwithstanding its importance in understanding why and how certain arrangements outperform others. Theoretical progress has recently been made in the economics of organization, partially inspired by game theory, about how to explore more systematically the role and impact of managerial practices on performance (Gibbons & Henderson, 2012, 2013). On the empirical side, Ménard (1996, 1998) early on examined the importance of coordinating devices in hybrids (e.g., through the case of the French poultry industry). However, studies limited to one case or a handful of cases raise issues of generalization (for a re-examination of the value of case studies, see Skarbek, 2020; and Bennett, this *Handbook*, ► Chap. 40). Recent contributions with more extensive datasets should also help overcome this limitation (see Bernstein & Peterson, 2023).

These examples are only illustrative, to suggest the numerous avenues opening for young and entrepreneurial researchers.

20.6 Conclusion

The introduction of hybrids in the world of micro-institutions has made the representation of the ways in which market economies organize transactions much richer and complex than the traditional micro-economics of "markets" and "firms." What makes hybrids distinct is that they develop relationships among legally separate parties that share assets and in many cases some property rights while remaining autonomous and even competitors on other segments of their activities. In sum, partners enter into a principal-principal relationship, distinct from a principal-agent one. To monitor and secure this relationship over time, hybrids rely on complex decision-making processes, with a crucial role for relational contracts and management. This complexity also helps explain the "impossibility of selective intervention" (Williamson, 1996: 150), that is, the impossibility of replicating a hybrid form within an integrated firm or of replicating market properties within a hybrid (or a firm for that matter) because of the "fundamental difference in how ownership is allocated" (Foss, 2003: 346).

A lot remains to be done to better understand what determines the boundaries between the different arrangements and *under what circumstances* one organizational solution outperforms another one. Answering these questions requires a thorough investigation of the internal characteristics of alternative arrangements.

The richness of the resulting research agenda is a tribute to the contributions of NIE to the economics of organization. These contributions have also raised new questions that likely require going beyond TCE. In that respect, NIE remains a progressive research program.

Whether the term hybrid will survive the filter of time should not obscure that Coase, Williamson, and TCE opened a resilient field of investigations: there is a whole set of transactions that cannot be understood through the traditional lenses, which makes investigating the domain of micro-institutions both complex and exciting.

Acknowledgments I am grateful to two reviewers and Mary Shirley for their very stimulating comments. And I have a long-term debt to all the students and professionals who have raised so many relevant and challenging questions over the years.

Acknowledgment of Funding Open access is made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Aghion, P., & Tirole, J. (1997). Formal and real authority in organizations. *Journal of Political Economy*, 105(1), 1–29.
- Alchian, A. A. (1965). Some economics of property rights. *Il Politico*, 30(4), 816–819.
- Alchian, A. A., & Demsetz, H. (1972). Production, information costs and economic organization. *American Economic Review*, 62(5), 777–795.
- Alchian, A. A., & Woodward, S. (1987). The firm is dead. Long live the firm. *Journal of Economic Literature*, 26(1), 65–79.
- Allen, D. W. (2000). Transaction costs. In B. Bouckaert & G. de Geert (Eds.), *Encyclopedia of law and economics* (pp. 893–926). Edward Elgar Publishing.
- Aoki, M. (1988). *Information, incentives, and bargaining in the Japanese economy*. Cambridge University Press.
- Argyres, N., & Zenger, T. (2012). Transaction costs and firm boundaries. *Organization Science*, 23, 1463–1667.
- Arrow, K. (1962). The economic implications of learning-by-doing. *Review of Economic Studies*, 29(3), 155–173.
- Arrow, K. (1974). *The limits of organization*. Norton and Co.
- Arrow, K., & Hahn, F. (1971). *General competitive analysis*. Holden Day.
- Arrunada, B. (2003). Property enforcement as organized consent. *Journal of Law, Economics and Organization*, 19(2), 401–444.
- Bajari, P., & Tadelis, S. (2001). Incentives versus transaction costs: A theory of procurement contracts. *RAND Journal of Economics*, 32(3), 387–407.
- Baker, G., Gibbons, R., & Murphy, K. (2002). Relational contracts and the theory of the firm. *Quarterly Journal of Economics*, 117(1), 39–84.
- Baker, G., Gibbons, R., & Murphy, K. (2008). Strategic alliances: Bridges between “islands of conscious power”. *Journal of the Japanese and International Economies*, 22(2), 146–163.
- Barnard, C. (1938). *The functions of the executive*. Harvard University Press.
- Barzel, Y. (1982). Measurement costs and the organization of markets. *Journal of Law and Economics*, 25(2), 27–48.
- Barzel, Y. (1989). *Economic analysis of property rights*. Cambridge University Press.

- Beckmann, M. J. (1988). *Tinbergen lectures on organizational theory*. Springer Verlag.
- Bernstein, L. (2015). Beyond relational contracts: Social capital and network governance in procurement contracts. *Journal of Legal Analysis*, 7(2), 561–621.
- Bernstein, L., & Peterson, B. (2023). Managerial contracting. *Journal of Legal Analysis*, 14(1), 176–243.
- Blader, S., Gartenberg, C., & Pratt, A. (2020). The contingent effect of management practices. *The Review of Economic Studies*, 87(2), 721–749.
- Bradach, J. (1997). Using the plural form in the management of restaurant chains. *Administrative Science Quarterly*, 42(2), 276–303.
- Bradach, J., & Eccles, R. (1989). Price, authority, and trust: From ideal types to plural forms. *Annual Review of Sociology*, 15, 97–118.
- Coase, R. H. (1937). The nature of the firm. *Economica*, 2(1), 386–405.
- Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 3(October), 1–44.
- Coase, R. H. (2000). The Acquisition of Fisher Body by general motors. *Journal of Law and Economics*, 43(1), 15–32.
- Commons, J. R. (1934). *Institutional economics. Its place in political economy*. Macmillan.
- Cook, M. (1995). The future of US agricultural cooperatives: A neo-institutional approach. *American Journal of Agricultural Economics*, 77(December), 1153–1159.
- Dahlman, C. (1979). The problem of externality. *Journal of Law and Economics*, 22(1), 141–162.
- Davis, L., & North, D. (1971). *Institutional change and American economic growth*. Cambridge University Press.
- de Vita, G., Tekaya, A., & Wang, C. (2011). The many faces of asset specificity: A critical review of key theoretical perspectives. *International Journal of Management Reviews*, 13(4), 329–348.
- Debreu, G. (1959). *The theory of value: An axiomatic analysis of economic equilibrium*. Princeton University Press.
- Demsetz, H. (1988). *Ownership, control, and the firm*. Basil Blackwell.
- Demsetz, H. (2011). The problem of social cost: What problem? A critique of the reasoning of A.C. Pigou and R.H. Coase. *Review of Law & Economics*, 7(1), 1–13.
- di Marcantonio, F., & Ciaian, P. (2017). *Unfair trading practices in the food supply chain: A literature review on methodologies, impacts and regulatory aspects*. European Commission, Joint Research Centre. <http://publications.jrc.ec.europa.eu/repository/handle/JRC108394>
- Dwyer, R., & Oh, S. (1988). A transaction cost perspective on vertical contractual structure and interchannel competitive strategies. *Journal of Marketing*, 52, 2–21.
- Dyer, J. (1996). Does governance matter? Keiretsu alliances and asset specificity as sources of Japanese competitive advantage. *Organization Science*, 7(6), 644–666.
- Dyer, J., & Chu, W. (2000). The determinants of trust in supplier-buyer relations in the U.S., Japan, and Korea. *Journal of International Business Studies*, 31(2), 259–285.
- Eccles, R. (1981). The Quasifirm in the construction industry. *Journal of Economic Behavior and Organization*, 2(4), 335–357.
- Farrell, J., & Scotchmer, S. (1988). Partnerships. *Quarterly Journal of Economics*, 103(May), 279–297.
- Foss, N. (2003). Selective intervention and internal hybrids: Interpreting and learning from the rise and decline of the Oticon spaghetti organization. *Organization Science*, 14(3), 331–349.
- Gambardella, A., & Panico, C. (2014). On the management of open innovation. *Research Policy*, 43(5), 903–913.
- Gibbons, R., & Henderson, R. (2012). Relational contracts and organizational capabilities. *Organization Science*, 23(5), 1350–1364.
- Gibbons, R., & Henderson, R. (2013). What do managers do? In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics* (pp. 680–730). Princeton University Press.
- Gibbons, R., & Roberts, J. (Eds.). (2013). *Handbook of organizational economics*. Princeton University Press.
- Grandori, A., & Soda, G. (2009). Governance and organization design: A negotiation and network analytic approach. *Corporate Ownership and Control*, 6(3), 489–503.

- Gulati, R. (1998). Alliances and networks. *Strategic Management Journal*, 19, 293–317.
- Hart, O. (1989). Incomplete contracts. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *The new Palgrave: Allocation, information and markets* (pp. 163–179). Macmillan.
- Hart, O., Frydlienger, D., & Vitasek, K. (2019, September–October). A new approach to contracts. *Harvard Business Review*. <https://hbr.org/2019/09/a-new-approach-to-contracts>
- Hayek, F. (1945). The use of knowledge in society. *American Economic Review*, 35(4), 519–530.
- Hennart, J.-F. (1988). A transaction cost theory of equity joint ventures. *Strategic Management Journal*, 9, 361–374.
- Hennart, J.-F., & Verbeke, A. (2022). Actionable and enduring implications of Oliver Williamson's transaction cost theory. *Journal of International Business Studies*, 53, 1557–1575.
- Holmstrom, B. (1999). The firm as a subeconomy. *Journal of Law, Economics and Organization*, 15(1), 74–102.
- Holmstrom, B., & Milgrom, P. (1991). Multi-task principal-agent analyses: Incentive, contracts, asset ownership, and job design. *Journal of Law, Economics and Organization*, 7, 24–52.
- John, G., & Weitz, B. (1988). Forward integration into distribution: An empirical test of transaction cost analysis. *Journal of Law, Economics and Organization*, 4(2), 337–355.
- Joskow, P. (1985). Vertical integration and long-term contracts: The case of coal-burning electric generating plants. *Journal of Law, Economics, and Organization*, 1(1), 33–80.
- Klein, B. (1996). Why hold-ups occur? The self-enforcing range of contractual relationships. *Economic Inquiry*, 34, 444–463.
- Klein, B. (2000). Fisher-general motors and the nature of the firm. *Journal of Law and Economics*, 43(1), 105–141.
- Klein, P. (2005). The Make-or-buy decision: Lessons from empirical studies. In C. Ménard & M. Shirley (Eds.), *Handbook of new institutional economics* (pp. 435–464). Springer.
- Klein, B., Crawford, R., & Alchian, A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21(2), 297–326.
- Knoeber, C. (1989). A real game of chicken: Contracts, tournaments, and the production of broilers. *Journal of Law, Economics and Organization*, 5(2), 271–292.
- Kunneke, R., Ménard, C., & Groenewegen, J. (2021). *Network infrastructures: Technology meets institutions*. Cambridge University Press.
- Lafontaine, F., & Bhattacharyya, S. (1995). Double-sided moral hazard and the nature of share contracts. *The Rand Journal of Economics*, 26(4), 761–781.
- Lafontaine, F., & Shaw, K. (1999). The dynamics of franchise contracting: Evidence from panel data. *Journal of Political Economy*, 107(5), 1041–1080.
- Lafontaine, F., & Slade, M. (1997). Retail contracting: Theory and practice. *Journal of Industrial Economics*, 45(March), 1–25.
- Lafontaine, F., & Slade, M. (2007). Vertical integration and firm boundaries: The evidence. *Journal of Economic Literature*, 45(3), 629–685.
- Lafontaine, F., & Slade, M. (2013). Inter-firm contracts. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics* (pp. 958–1012). Princeton University Press.
- Li, J., & Matouschek, N. (2013). Managing conflicts in relational contracts. *American Economic Review*, 103(6), 2328–2351.
- Libecap, G., & Wiggins, S. (1985). The influence of private contractual failure on regulation: The case of oil field unitization. *Journal of Political Economy*, 93(4), 690–714.
- Macaulay, S. (1963). Non-contractual relations in business: A preliminary study. *American Sociological Review*, 28(1), 55–67.
- Macneil, I. (1974). The many futures of contracts. *Southern California Law Review*, 47(May), 691–816.
- Malcolmson, J. (2013). Relational incentive contracts. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics* (pp. 1014–1065). Princeton University Press.
- Malmgren, H. B. (1961). Information, expectations, and the theory of the firm. *Quarterly Journal of Economics*, 75(3), 399–421.

- Maskin, E., & Tirole, J. (1999). Unforeseen contingencies and incomplete contracts. *The Review of Economic Studies*, 66(1), 83–114.
- Masten, S. (1984). The organization of production: Evidence from the aerospace industry. *Journal of Law and Economics*, 27(2), 403–417.
- Masten, S. (1988). A legal basis for the firm. In O. E. Williamson & S. Winter (Eds.), *The nature of the firm* (pp. 196–212). Oxford University Press.
- Masten, S. (Ed.). (1996). *Case studies in contracting and organization*. New York-Oxford.
- Masten, S., Meehan, J., & Snyder, E. (1991). The costs of organization. *Journal of Law, Economics, and Organization*, 7(1), 1–25.
- Ménard, C. (1996). On clusters, hybrids, and other strange forms. The case of the French poultry industry. *Journal of Institutional and Theoretical Economics*, 152(1), 154–183.
- Ménard, C. (1997/2004). The Governance of hybrid organizational forms. In C. Ménard (Ed.), *Modes of organization in the new institutional economics. The international library of the new institutional economics* (Vol. 4, pp. 105–113).
- Ménard, C. (1998). The maladaptation of regulation to hybrid organizational for. *International Review of Law and Economics*, 18(4), 403–417.
- Ménard, C. (2004). The economics of hybrid organizations. *Journal of Institutional and Theoretical Economics*, 160(3), 345–376.
- Ménard, C. (2013a). Hybrid modes of organization. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics* (pp. 1066–1105). Princeton University Press.
- Ménard, C. (2013b). Plural forms of organization: Where do we stand? *Managerial and Decision Economics*, 34(3–5), 124–139.
- Ménard, C. (2018). Finding our way in the jungle: Insights from organization theory. In G. Martino, K. Karantininis, S. Pascucci, L. Dries, & J.-M. Codron (Eds.), *It's a jungle out there! The strange animals of economic organization in Agri-food value chains* (pp. 27–50). Wageningen Academic Publishers.
- Ménard, C. (2022). Hybrids: Where are we? *Journal of Institutional Economics*, 18(2), 297–312.
- Ménard, C., & Damergy, G. (2023). Hybrid-within-hybrid: Governance issues within international airline alliances. *European Journal of International Management*, 1, 1. <https://doi.org/10.1504/EJIM.2022.10055799>
- Ménard, C., & Oudot, J.-M. (2009). L'évaluation préalable dans les contrats de partenariat. *Revue Française d'Administration Publique*, 130, 349–364. <https://doi.org/10.3917/rfap.130.0349>
- Ménard, C., & Shirley, M. (2022). *Advanced introduction to new institutional economics*. Edward Elgar Pub.
- Milgrom, P., & Roberts, J. (1995). *The economics of organization and management*. Prentice Hall.
- Monteverde, K., & Teece, D. (1982). Suppliers switching costs and vertical integration in the automobile industry. *Bell Journal of Economics*, 13(1), 206–213.
- Oxley, J. (1997). Appropriability hazards and governance in strategic alliances: A transaction cost approach. *Journal of Law, Economics and Organization*, 13(2), 387–409.
- Oxley, J. (1999). Institutional environment and the mechanism of governance: The impact of intellectual property protection on the structure of inter-firm alliances. *Journal of Economic Behavior and Organization*, 38(3), 283–309.
- Oxley, J., & Silverman, B. (2008). Interfirm alliances: A new institutional economics approach. In E. Brousseau & J. M. Glachant (Eds.), *New institutional economics: A guidebook* (pp. 209–234). Cambridge University Press.
- Palay, T. (1984). Comparative institutional economics: The governance of the rail freight contract. *The Journal of Legal Studies*, 13(June), 265–288.
- Palay, T. (1985). Avoiding regulatory constraints: Contracting safeguards and the role of informal agreements. *Journal of Law, Economics, and Organization*, 1(1), 155–175.
- Park, S. H. (1996). Managing an interorganizational network: A framework of the institutional mechanism for network control. *Organization Studies*, 17(5), 795–824.
- Podolny, J., & Page, K. (1998). Network forms of organization. *Annual Review of Sociology*, 24, 57–76.

- Poppo, L., & Zenger, T. (2002). Do formal contracts and relational governance function as substitutes or complements? *Strategic Management Journal*, 23, 707–725.
- Powell, W. (1990). Neither market nor hierarchy: Network forms of organization. In L. L. Cummings & B. Staw (Eds.), *Readings in organizational behavior* (pp. 295–336). JAI Press.
- Powell, W. (1996). Inter-organizational collaboration in the biotechnology industry. *Journal of Institutional and Theoretical Economics*, 152(1), 197–215.
- Radner, R. (1992). Hierarchy: The economics of managing. *Journal of Economic Literature*, 30(3), 1382–1415.
- Raynaud, E. (1997). *Propriété et exploitation partagée d'une marque commerciale: aléas contractuels et ordre privé*. Ph.D. Université de Paris (Panthéon-Sorbonne).
- Reuer, J., & Devaradonka, S. (2016). Mechanisms of hybrid governance: Administrative commitments in non-equity alliances. *Academy of Management Journal*, 59(2), 510–533.
- Reuer, J., & Klijn, E. (2018). Governance of hybrid organizations. *Annals of Corporate Governance*, 3(1), 1–81.
- Richardson, G. B. (1972). The organization of industry. *Economic Journal*, 82(327), 383–396.
- Robinson, D., & Stuart, T. (2007). Network effects in the governance of strategic alliances. *Journal of Law, Economics and Organization*, 23(1), 242–273.
- Rubin, P. (1978). The theory of the firm and the structure of the franchise contract. *Journal of Law and Economics*, 21(1), 223–233.
- Sampson, R. (2004). Organizational choice in R&D alliances: Knowledge-based and transaction costs perspectives. *Managerial and Decision Economics*, 25(6–7), 421–436.
- Sauvée, L., Raynaud, E., & Valceschini, E. (2005). Alignment between quality enforcement devices and governance structures in the agro-food vertical chains. *Journal of Management and Governance*, 9(1), 47–77.
- Shelanski, H., & Klein, P. (1995). Empirical research in transaction cost economics: A survey and assessment. *Journal of Law, Economics, and Organization*, 11(2), 335–361.
- Simon, H. (1951). A formal theory of the employment relationship. *Econometrica*, 19(3), 293–305.
- Skarbek, D. (2020). Qualitative research methods for institutional analysis. *Journal of Institutional Economics*, 16(4), 409–422.
- Smith, H. (2021). Property beyond flatland. *Brigham-Kanner Property Rights Journal*, 10-(October), 9–56.
- Tadelis, S., & Williamson, O. (2013). Transaction cost economics. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics* (pp. 159–189). Princeton University Press.
- Teece, D. (2010). Forward integration and innovation: Transaction costs and beyond. *Journal of Retailing*, 86(3), 277–283.
- Teece, D., & Pisano, G. (1994). The dynamic capabilities of firms. *Journal of Economic Behavior and Organization*, 27(3), 537–556.
- Thorelli, H. (1986). Networks: Between markets and hierarchies. *Strategic Management Journal*, 7(1), 37–51.
- US DoT and European Commission joint report. (2010). *Transatlantic Airline alliances: Competitive Issues and Regulatory Approaches*. Joint report, November 16. https://ec.europa.eu/competition/sectors/transport/reports/joint_alliance_report.pdf
- Whinston, M. (2003). On the transaction costs determinants of vertical integration. *Journal of Law, Economics, and Organization*, 19(1), 1–23.
- Williamson, O. (1971). The vertical integration of production: Market failure considerations. *American Economic Review*, 61(May), 112–123.
- Williamson, O. (1975). *Markets and hierarchies: Analysis and antitrust implications*. Free Press.
- Williamson, O. (1979). Transaction cost economics: The governance of contractual relations. *Journal of Law and Economics*, 22(2), 3–61.
- Williamson, O. (1985). *The economic institutions of capitalism*. The Free Press-Macmillan.

- Williamson, O. (1996). *The mechanisms of governance*. Oxford University Press.
- Williamson, O. (2009). Transaction cost economics: The natural progression. *Nobel Lecture*, 455–476. https://www.nobelprize.org/uploads/2018/06/williamson_lecture.pdf
- Yarbrough, B., & Yarbrough, R. (1988). The transactional structure of the firm: A comparative survey. *Journal of Economic Behavior and Organization*, 10(1), 1–28.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



The Persistent Institutions of Corporate Governance

21

Mark J. Roe

Contents

21.1	Introduction and Scope: The Organization at the Top of the Large Business Firm	506
21.2	The Core Problems of Corporate Governance	507
21.2.1	Vertical Corporate Governance: Managerial Agency Costs	509
21.2.2	Horizontal Corporate Governance: Taming Dominant Shareholders	511
21.2.3	External Corporate Governance: Corporate Legitimacy	511
21.2.4	Enduring or Ephemeral?	513
21.3	The Basic Institutions of Corporate Governance	513
21.3.1	Markets	513
21.3.2	The Board of Directors	516
21.3.3	Information Distribution and Gatekeeping	517
21.3.4	Coalescing Shareholders: Takeovers, Proxy Fights, Hedge Funds, Institutional Investors, and Shareholder Voice	517
21.3.5	Executive Compensation	519
21.3.6	Professionalism and Norms	520
21.3.7	Corporate Lawsuits	520
21.3.8	Complements and Substitutes	521
21.4	Contract Law as the Corporate Law “Primitive”?	522
21.4.1	Corporate Law as Standardized Corporate Contract	522
21.4.2	Contract Law’s Limits	523
21.5	Institutional Legitimacy: CSR, ESG, and Corporate Purpose	523
21.5.1	American Legitimacy Light: Populism and Anti-Takeover Laws	524
21.5.2	Legitimacy Heavy: European Left-Right Politics	524
21.5.3	Corporate Social Responsibility	527
21.6	Corporate Institutions in Their Setting	527
21.6.1	Contract	527
21.6.2	Property Rights	530
21.7	Summary	531
	References	531

M. J. Roe (✉)
 School of Law, Harvard University, Cambridge, MA, USA
 e-mail: mroe@law.harvard.edu

Abstract

In this chapter, I examine the institutions of corporate governance. These institutions primarily deal with two distinct problems, one of vertical governance (between distant shareholders and managers) and another of horizontal governance (between a close, controlling shareholder and distant shareholders). If institutions cannot deal well with both of these problems, then large public firms with dispersed ownership will be relatively more costly and more difficult for an economy to sustain.

Some institutions deal well with vertical corporate governance but do less well with horizontal governance. The institutions interact as complements and substitutes, and many can be seen as developing out of a “primitive” of contract law. Arguably a system must first have satisfactory contract enforcement institutions, as well as basic property rights that shield private actors from the state before it embarks on more sophisticated corporate governance institutions.

Keywords

Corporate law · Corporate governance · Board of directors · Concentrated ownership · Corporate social responsibility

21.1 Introduction and Scope: The Organization at the Top of the Large Business Firm

I outline here the institutions of decision-making in the large public firm in the wealthy West, emphasizing those that try to thwart decision-making from going awry. While that has been the long-standing traditional focus of corporate governance, new forces in recent years have pressured large firms to suppress profit for social purpose. The firm faces increased pressure from social forces seeking that the firm contributes more to public well-being and that it reduces profitable externalities such as those that contribute to climate change.

By corporate governance, I mean the relationships among the players at the top of the firm—the board of directors, the senior managers, and the stockholders. By institutions, I mean those repeated mechanisms that allocate authority among the three and that affect, modulate, and control the decisions made by these parties at the top of the firm.

By taking governance to mean the relationships among the triumvirate at the top—and not taking it to mean their relations with, say, the firm’s employees or its labor unions—I right away give the analysis an American cast, not a European one. Although my primary focus here is on American corporate governance institutions, I look at other nations’ corporate governance institutions primarily by way of contrast (as opposed to, say, finding the deep functions that all corporate systems must have). Thus, I focus on the division of authority between the board and the CEO on the one hand and the shareholders on the other, where the primary task of the institutions of

corporate governance is to make managers inside the firm run that firm well and be loyal to shareholders. At the end of this paper, I look at how these relationships relate to basic institutions of contract and political organization in Europe and the United States and how these institutions have been under increasing social pressure in recent years.

I focus on features that are at the heart of recent decades of academic legal inquiry. Others could, and would, emphasize other governance features: an organizational theorist might look to how a leader motivates the people in a large organization. A psychologist or a sociologist might emphasize how discussion in the boardroom is vibrant, supportive, and inquiring, instead of stale, formal, and useless. A technology theorist might emphasize the mechanisms by which the firm innovates. Another might emphasize how ideas are transformed into products. One type of economist might relentlessly analyze what goes on inside the firm (and thereby is “governed” within) and what goes on outside it (and is thereby “governed” by contract). These are all worthwhile modes of inquiry. But they are not mine, at least not here.

In Sect. 21.1, I sort out the central problems of corporate governance. In Sect. 21.2, I catalog the basic institutions of corporate governance, from markets to organization to contract. In Sect. 21.3, I consider contract law as corporate law’s “primitive” building block. In Sect. 21.4, I briefly examine issues of corporate legitimacy that could, and do, affect corporate governance. The institutions of corporate governance are usually seen as separated from the institutions of political organization. But they should not be. In recent years the public has come to expect more of governance institutions than it did before. In Sect. 21.5, I re-examine corporate governance in terms of economies of scale, contract, markets, and property rights. And then I summarize and conclude.

21.2 The Core Problems of Corporate Governance

The corporate governance triumvirate—the board, the managers, and the stockholders—traditionally has had a vertical and a horizontal dimension. The vertical dimension is between senior managers and distant shareholders (see Fig. 21.1). The focus there is on keeping the CEO and the top people (the board and senior officers) loyal to shareholders and competent for the task of managing the firm. It’s this vertical dimension that’s been especially relevant in the United States.

The horizontal dimension is between dominant stockholders and dispersed stockholders (see Fig. 21.2). Instead of the primary friction being to align managers with the firm’s owners, the horizontal focus sketched in Fig. 21.2 is on preventing or minimizing the shifts in value from dispersed outsiders to controlling inside stockholders. The frictions are between shareholders instead of between shareholders and managers. That horizontal dimension of inquiry has historically been less important in the United States than it has been in Europe. Two reasons for this difference are the institutions of corporate jurisprudence which historically has limited insider self-dealing more effectively in the United States and, second, the basic fact that large American firms have historically tended to have more dispersed ownership. Without

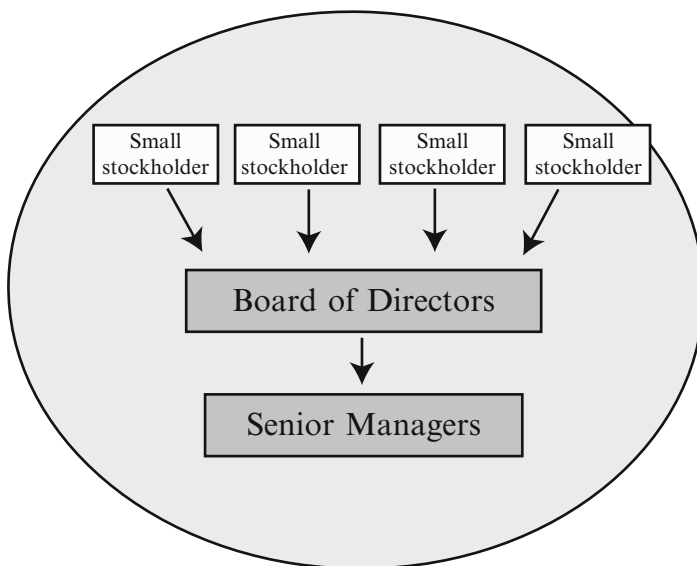


Fig. 21.1 Vertical governance: diffuse ownership

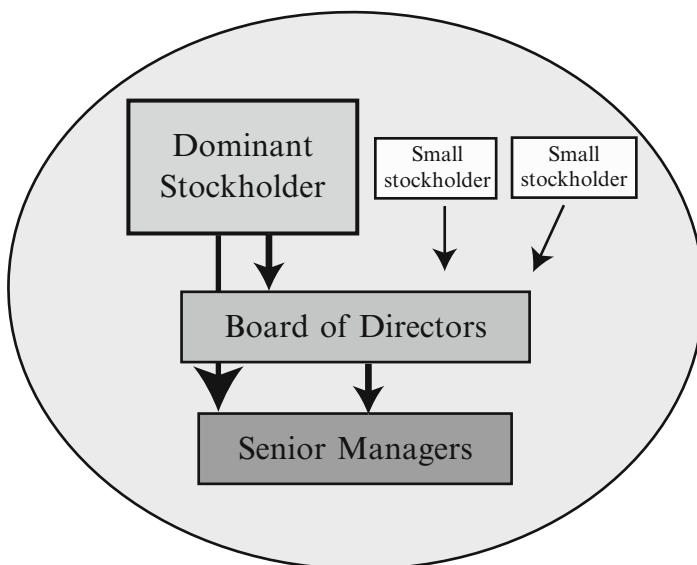


Fig. 21.2 Horizontal governance: controlling vs. minority stockholders in the public firm

a controlling insider, there are fewer controlling-insider issues (and more managerial motivation issues).

Most large public firms in the United States lack a single controller, although those that do—like Tesla's Elon Musk—have such controller frictions. See Tornetta

v. Elon Musk (Del. Chancery 2018-0408, Jan. 30, 2024) (invalidating Elon Musk's \$50 billion compensation plan). Lacking a single shareholder-controller, the typical American firm has fewer horizontal problems but more vertical problems: dispersed ownership diminishes the horizontal dimension but brings to the forefront the firm's vertical weaknesses. Foreign nations also have legitimacy as a core problem of corporate governance, a muted issue in the United States.

These two corporate governance problems are similar in one dimension—in each a controller extracts rents or private benefits—but less so in other, perhaps more critical dimensions. First, the centrality of each problem differs around the world—horizontal issues dominate in most of the world, vertical issues in the United States. Second, the means by which the controller extracts benefits differ between the two systems. And, third, the means to mitigate the costs of each system differ. These distinctions are not always made, but should be.

Finally here, the third dimension can be thought of as traditional legitimacy; the corporation cannot impose externalities on the rest of society, even if it is profitable to do so and even if public regulation has not barred the activity.

21.2.1 Vertical Corporate Governance: Managerial Agency Costs

Public firms with full ownership separation have no dominant shareholder—the shareholders are small relative to the firm, are not deeply informed about the firm's operations, and have limited influence on the firm's operations. With shareholders dispersed, the task of keeping managers working primarily in shareholders' interests becomes critical (Jensen & Meckling, 1976). And ownership in the largest American firms is dispersed: Mark Zuckerberg's dominance of Facebook is an exception; General Motors, which lacks a dominant stockholder, is the norm. This diffused ownership is, in the classical analysis, layered over a basic principal-agent problem: the stockholders' agenda—typically maximizing shareholder value—can be at odds with managers' agendas. In that setting, aligning stockholders' interests becomes the typical problem of corporate governance. And the institutions of corporate governance are those repeated mechanisms that tie, or fail to tie, managers to stockholders. In recent years, problems of aligning executives' interests with those of shareholders have come to be seen as less costly than they once were. In part, the small stockholder model has moved to a more concentrated institutional investor model; traditional corporate governance has shifted away from the problem of dispersed shareholders facing centralized management and toward the relationship between larger institutional investors and corporate boards.

This “agency cost,” principal-agent, stockholder-manager alignment problem comes in two main varieties. One variety is of diversion, while the other is of competence, “stealing and shirking” in its alliterative form. Managers could divert value from the firm into their own hands: they could have the firm transfer funds to their own bank accounts (or a relative's) or more surreptitiously have the firm sell goods at low prices to (or buy at high prices from) entities that the managers control or more transparently pay themselves (excessively) high salaries.

Agency costs come in a second variety. Managers might not be up to running the firm, either because one or another manager never was up to the task (their selection was a mistake) or more plausibly because changed circumstances made the incumbent manager no longer the best-equipped person to run this company. Well-functioning corporate governance institutions put the appropriately skilled executive in place and give that executive incentives and constraints designed to maximize returns.

The alliterative “stealing and shirking” hides a couple of issues. The first is that each specific corporate governance institution does not uniformly reduce both. Some affect one, some affect the other. “Stealing and shirking” are both costs to shareholders, and so it is correct to lump them together as costs, but they cannot be lumped together when evaluating the institutions of corporate governance because the institutions affect the two differently. For example, corporate lawsuits are geared to handling observed “stealing” but less good at reducing “shirking” or managerial error. (In fact, via the “business judgment rule,” the American corporate judge won’t listen to complaints about managerial error.) Corporate disclosure rules, in contrast, are probably better at affecting “shirking”—distant stockholders can see the bottom line results that managers produce but less good at handling “stealing” (because for very large firms, lots of personal wealth could be obtained from a large firm without destroying the corporate bottom line).

The second submerged element is that “shirking” doesn’t convey the subtlety of the vertical problem. Much of the power of an organization comes from routines and embedded information. These give the organization great power and efficiency. Bureaucracy often has a pejorative connotation, but bureaucracies exist for a good reason: they are effective at routinizing action to produce powerful output. But when markets or technologies change, the routines and embedded information can weaken the organization. The old routines repeat, and the organization works via its embedded information, even though neither the old routines nor the embedded information correspond to the demands of the market or the new technologies (Henderson & Clark, 1990). Organizational change is necessary, but insiders often cannot refashion the firm, because they are subject to its old routines and its embedded information. In a sense, it’s not so much that the managers “shirk” in the primary sense of the word—they desperately *want* to work—but they don’t know what to do. Their efforts may be intense but misdirected.

So, the vertical corporate governance problem is one of vertical agency costs. That problem interacts with a shareholder free-rider problem, due to the large firm’s ownership by small, distant shareholders. Some corporate governance institutions seek to overcome the free-rider problem, so that large shareholders can focus their ownership and thereby deal better with managerial agency costs. These focusing institutions include a more responsive board of directors (as a focal point for shareholders’ impact on managers), takeovers, blockholders capable of reining in some managerial agency costs, proxy voting, and so on. Institutional investors owning 5% of the public firm’s stock have the potential for action beyond that of the historically small individual investor, but for the most part, these institutions are not directly active—their fee structure does not support activity—but they do

sometimes act to support (or oppose) hedge fund and other activists seeking corporate change. And this possibility of focused shareholder action, especially large blockholder action to mitigate the vertical agency cost problem, leads us into the second core corporate problem: horizontal governance.

21.2.2 Horizontal Corporate Governance: Taming Dominant Shareholders

Large firms with a dominant stockholder have a different core problem, one between stockholders, with a focus on a dominant stockholder's potential to shift value from the minority stockholders to itself. Horizontal corporate relationships have traditionally tended to be the focus of corporate governance in continental Europe, Asia, and Latin America. Again, the alliterative "stealing and shirking" fits, but for these types of firms, the emphasis is less on the "shirking" by managers than on the "stealing" by the dominant stockholders. Early twentieth-century analysis saw the inability to stop controllers' self-dealing as central to developing securities markets and the means to finance large firms (see especially the La Porta et al.'s series of articles (La Porta et al., 1997, 1998, 1999) and the literature summarized in Denis and McConnell (2003)). A consensus arose that in the transition nations, especially Russia, the failure to develop institutions that would stop dominant stockholders' self-dealing fundamentally flawed the corporate transition (see, e.g., Black et al. (2000)). And the flawed corporate transition may well have impeded political transitions.

21.2.3 External Corporate Governance: Corporate Legitimacy

Corporate governance can have another dimension. Neither necessarily horizontal nor vertical, nor necessarily governing the relationships among owners and managers, this dimension is of governing the firm so that it is legitimate in its society. One could see it as purely defensive: the inside players act, or build institutions, that they would not otherwise have done or built, to deflect outside incursions into the firm. Popular narrative in the media and elsewhere affect how the corporation is perceived (Roe & Shapira, 2021). Or one could see this legitimacy dimension as determined by the outsiders: the polity demands that the firm structure its internal governing institutions in this way or that way, to implement some public policy.

This dimension is shallow for modern corporate governance in the United States (at least at the time I write) but has not been completely absent: mechanisms to detect and prevent foreign bribery in the 1970s and 1980s could be seen as having been one such governance goal in the United States. One might see anti-takeover laws as resulting in part from the antagonism they attracted from the polity (Romano, 1988).

This dimension is deeper in many other nations. For example, in the United States, corporate governance vis-à-vis employees is absent from the governance institutions at the top. Employees are typically seen in the traditional American academic literature as inputs (like suppliers, capital, etc.), ones that managers

transform into sales and profits (Williamson, 1985). Employees have to be paid well and think well of their work environment, but this is not usually seen as a corporate governance issue, at least not in terms of vertical or horizontal corporate governance at the “top” of the firm. Not so in other countries, where social considerations and the role of employees have much more impact. Sometimes the impact is formal: German-style codetermination, in which employees have a voice and vote inside the boardroom, is one such example. Sometimes the impact is indirect: owners might seek to structure the firm so that the outside incursions, or their impact, are minimized. Some of this may be a function of the labor market and general mobility. If people can readily “walk with their feet”—move from, say, Indiana to California if they become disgruntled—they may demand less influence inside the local firm than if they and their ancestors have always lived in the Ruhr Valley and expect to spend their lives working for one big local firm.

These issues of legitimacy also tie into modes of production. By defining the corporate governance inquiry in such stark terms—vertical, horizontal, and legitimacy—I have excluded other perspectives, as I noted in the Introduction. For example, one might view the problem of corporate governance in organizational terms, as how to motivate a team, how to keep it flexible but effective, etc. That kind of inquiry focuses less on relationships between shareholders and managers or among shareholders. It could focus on how to make the team as effective as possible. Or it could focus on the distribution of rewards inside the firm. It could have a “legitimacy” angle, as some structures command the respect of voters and political leaders and other structures do not. (On the idea of legal institutions as reflecting some organization of managers, employees, and capital, see Blair and Stout (1999). On the idea of national systems revolving around varying organizational modes, see Hall and Soskice (2001).)

Pressures for a public purpose for the large corporation have always been present and debated among academics. In recent years, these pressures have grown, with greater expectations—or hope—that corporations will promote proponents’ notions of the public good. Important institutional investors have pushed these views forward and sought that boards better reflect gender and minority diversity and that firms reign in externalities that contribute to climate change. Considerable debate has arisen on how (and whether) these notions of public good can be effective (e.g., Hart & Zingales, 2017). The efforts have recently generated a backlash from interests and those with ideologies opposed to such efforts (Puzder, 2023).

This issue of corporate governance and legitimacy can be viewed more abstractly: how, normatively can one organize a business entity that survives politically and socially? And, how as a positive matter have corporate institutions adapted to be legitimate in society? In a divided polity, like that in the United States, the divisions have impeded the political capacity to decide on how to handle multiple social, economic, and climate issues. These divisions in the overall polity are now spilling over into the institutions of corporate governance.

21.2.4 Enduring or Ephemeral?

What seems important in corporate governance is not enduring, but tends to arise from the problems and issues of the moment. “Core” may mean “current” and “transitory” and not be a permanent and enduring “core.”

In the 1980s in the United States, the vertical corporate governance issues of agency costs from managers who misdirected their firms’ operations were center stage. By the late 1990s and early twenty-first century, these seemed less central, and the focus was on scandals of (a few?) managers’ frauds, not on widespread misdirected operations, perhaps because corporate governance institutions attract most attention when economic performance is weak. As we move into the second quarter of the twenty-first century, the issues of corporate social responsibility have become more central than previously.

In the late 1990s in Europe and the transition nations (principally the nations that had been part of communist, Russian-dominated eastern Europe), the horizontal corporate governance issues of minimizing controlling shareholders’ abuses have been at center stage. Especially in the transition nations, and particularly in Russia, the inability to control insider machinations was a primary gap in the institutions of corporate governance (Black et al., 2000; La Porta et al., 1999), although it has come to be over-applied to the restricted sample of rich nations for reasons we’ll briefly see.

21.3 The Basic Institutions of Corporate Governance

We now have some texture for the basic traditional notion of corporate governance. Some of it is vertical: whether managers stay loyal to shareholders. Some of it is horizontal: whether dominant stockholders shift firm value into their own pockets. And some of it is external, based on issues of legitimacy: what does the rest of society force the firm to do, or how do players inside the firm react to outside pressures?

Which institutions align the players? In particular, given the focus of this chapter on the United States, these are the institutions that align American managers with their shareholders and facilitate the firm running well: competitive markets, boards, compensation, gatekeeping transparency, coalescing shareholders (via ownership, activism, and otherwise), information distribution, lawsuits, and capital structure.

21.3.1 Markets

Markets are often the most important institution of corporate governance. Three markets are central: the firm’s product market, its capital market, and the managerial labor market. If the firm cannot sell its product, in time it disappears. If the firm cannot raise capital, it cannot grow. If the firm cannot get and retain good managers (and employees), it will fritter away its resources.

This sometimes leads commentators to belittle the importance of other corporate governance institutions, because one or all of the three markets punish deviant firms

and reward well-performing firms. But using the impact of the market and competition to belittle institutional analysis can go too far, as we see next:

1. *Markets vs. institutions?* Even if markets roughly aligned the players, a markets-are-everything idea—academically popular in some circles, from time to time—would still face three substantial defects. First, each market is imperfect, sometimes substantially so. Corporate governance institutions might pick up the slack where markets cannot reach. Markets can sometimes deal with extremely gross deviations, while the institutions of corporate governance then deal with middle-sized deviations. Markets often draw an outer limit, but not a tight constraint.

Second, the corporate governance institutions can “prime” or charge up each one of the market institutions. Good internal decision-making can make the firm react well to product market changes, economize on capital, or make sure good managers come, stay, and perform. Good governance institutions can feed back and become one of the bases for marketplace competition in all three markets. They can facilitate or retard the firm’s adaptation to the three markets’ pressures.

Third, markets may be good for some governance tasks, while weak for others. Markets may be good at limiting the most egregious types of “shirking” but be less good at limiting “stealing,” especially if the stealing represents a small part of the firm’s total value. (If a low percentage of a firm’s assets is stolen in relation to forgone market opportunities, the market may not deter the manager. The manager may never get another job, but that manager will leave rich.) Law and other institutions are probably more important here than markets. That is, sometimes the nonmarket institutions are better, cheaper, and faster at governing the firm than any of the market constraints. And here—defeating stealing—corporate law may be important to corporate legitimacy.

2. *What markets cannot do.* So, each market-based institution is incomplete in optimizing the firm’s organizational structure. Let us take them each in turn, glancing at their strengths and limits.

Product markets can discipline managers. If there is no marketable product, there will shortly be no firm. But oligopolistic and monopolistic product markets leave slack. The product market doesn’t punish managers severely until the product cannot be sold profitably. That means that managers in non-competitive markets can “lose” some of the monopoly profit, but as long as customers do not flee from the firm and as long as the firm’s weakened performance does not trigger entry of new competitors, managers will only be mildly affected.

Product markets have a second serious limit in disciplining managers. Many firms have deep invested capital. Once that capital is invested, the managers need to recover “only” the firm’s variable costs to survive for the life of that capital (which may be longer than their own working lives). When costs are sunk and long-lived—distinctive features of the “old” economy of steel mills and factories—the constraints on managers weaken. Eventually the badly governed firm withers and disappears, but that eventuality could be a long time in coming; in the meanwhile, good corporate

governance could maximize the value obtained from a firm with heavy invested capital.

Capital market constraints are also incomplete. The importance of the constraint is simple: firms must raise capital for new projects. Firms with debt must return to capital markets and must pump cash out of operations to pay off that debt (Jensen, 1986). And firms and their managers that perform poorly cannot return easily to capital markets: poor performers will face a higher cost of capital. Extremely poor performers cannot get new capital. They wither and disappear, while the strong performers are getting more capital more cheaply and thereby expanding.

The limit here is that of retained earnings and, again, sunk costs. For some firms, capital once invested, if invested badly, means they cannot eventually return easily to capital markets. But again, that eventuality could be a very long run if the sunk capital is substantial, if the retained earnings are adequate, or if monopoly profits give them slack. This scenario is especially relevant for the firm that has large fixed investments that cannot be used in other businesses, i.e., once the automotive plant is built, the machinery has no obvious alternative use. True, such badly governed firms will wither in the long run, but improved corporate governance institutions—given the market's limits here—can make managers better stewards of the resources they control in the medium run. Moreover, given the depth of invested capital, governance institutions can be seen as the constraints on ex post bargaining and maneuvering over the quasi-rents created. When governance institutions are good, they facilitate ex ante investment (cf. Williamson, 1985; Zingales, 1997). When they're bad, an economy may have to forgo a key class of investments.

The perception that managerial agency costs are not as severe a problem as they once were in the United States could come from the evolution of product markets. Technological change is faster than ever. Firms cannot sit on their laurels for as long as they once could—because a new technology could quickly eliminate their business nowadays, in contrast to the firm running down its invested capital over many years, as once was more possible.

Managerial labor markets also have their limits. True, directors usually want to be on boards, and managers want to be compensated and promoted. If they do badly, they'll jeopardize their compensation and their future career prospects. But many senior managers are at the end of their careers. They aren't usually moving anywhere except into retirement. Labor markets may be important in selecting the "right" people for these final positions, but once "selected" for their final jobs, managers face weak labor market constraints. And, although the following idea is not boldly highlighted in academic analyses, many senior managers at large firms are already very rich. While they would surely welcome yet more money, more money is not their chief motivation.

3. *Limits vs. irrelevance?* The point here is *not* that these markets utterly fail to constrain managers, but that they are not tight, perfect constraints. Each leaves lots of "play at the joints." Eventually if the firm gets seriously out of line, it will hit a limit in one or another of these markets. But it can, depending on its market setting, waste major corporate resources before hitting that market-imposed limit. In a sense, corporate governance institutions fill in the gaps and detail of these three primary colors of the markets of corporate governance.

21.3.2 The Board of Directors

The board is the quintessential vertical corporate governance institution. It's the board that hires and fires the CEO, makes key business decisions, and reviews the work of the firm's senior managers. (Because a dominant stockholder typically controls the board, the board is less important as an institution of horizontal governance—less important in reigning in the controlling shareholder.) Indeed, one could see the other corporate governance institutions as primarily interacting with the board. For example, there are multiple institutions that affect how well shareholders coalesce. That concentration, when it occurs, primarily affects corporate governance through the concentrated shareholders' capacity to select a new board. Law affects these means through several channels. The law-regulating proxy contests and takeovers can make it easier or harder for outsiders to replace the board of directors. Those legal structures in turn are strengthened or weakened by the degree of information distribution and transparency that allow or impede outside stockholders to see what's happening inside the board.

In the abstract, it's simple: shareholders elect the board. Distant shareholders lack information and focus; they can neither run the company nor understand its business in any deep sense nor select or motivate the CEO. So the board manages the company in general, hiring managers to do the job day to day. In practice though, the board's role has been in flux for decades. Historically, the board was often seen as captive to senior managers who "suggested" people for vacancies and, through their control of information, were thought to dominate the board (Davis, 1993; Lorsch & MacIver, 1989). The era of the independent director followed, with inside directors declining in numbers and with many boards getting active audit committees, and with some having membership committees that took the nomination function away from the CEO (or at least shared it) (Useem, 1992). Evidence of the effectiveness of independent directors is mixed (Bhagat & Black, 1999; Franks et al., 2001; Gilson & Kraakman, 1991; Hermalin & Weisbach, 1998). Activists now seek minority members on boards to shake things up, (Spamann, 2022) with as many as one out of ten public companies being subject to such pressures in some years. Securities regulation adapted over the years to accommodate (and sometimes to restrict) these efforts.

There are enduring understandings that deal with the very basics of organizational governance: Organizations are thought to improve their decision-making when the people who make proposals are separated from those who approve them (Fama & Jensen, 1983: 303–04, 308). And there is an understanding of the value of committees in making decisions.¹ But beyond that, there's empirical uncertainty about the significance of board size, the degree of independence, and other board characteristics, despite the fact that independence, size, and related issues have been "planks" in corporate governance reforms (Eisenberg et al., 1998; Yermack, 1996).

¹ Alan B. Krueger, *Economic Scene: A study shows committees can be more than the sum of their members*, N.Y. TIMES, Dec. 7, 2000, at C2, citing Alan Blinder & John Morgan, Are Two Heads Better Than One? (available at www.princeton.edu/rjmorgan/working.htm).

21.3.3 Information Distribution and Gatekeeping

Distant shareholders need information about their companies. They (not all of them, of course) need information so that they can price their securities (and, via that pricing, *other* corporate governance institutions like incentive compensation or board action, which can be triggered by bad pricing results, come into play). They need information so that they can decide whether a corporate control transaction (ousting the board, taking over the company, engaging in a proxy fight) makes sense. And they need information so that they can decide whether legal action makes sense. If the managers engage in a related party transaction but the shareholders know nothing about it, a legal remedy might not arise, because no shareholder would know to sue. More generally, they need information so that they can help capital markets function properly in corporate governance (by denying or rewarding firms with new capital).²

The institutions here are the securities law mandates of periodic disclosure (and the penalties for noncompliance). The gatekeepers are those not deeply embedded inside the firm who verify or sometimes warrant the information about the firm: they are the lawyers, accountants, securities analysts, underwriters, and outside directors (Kraakman, 1986).

21.3.4 Coalescing Shareholders: Takeovers, Proxy Fights, Hedge Funds, Institutional Investors, and Shareholder Voice

If the core problems in American vertical corporate governance arise from the dispersion of stockholders, then a core solution could be to coalesce stockholders. Building big blocks of stock has its costs in owner liquidity and diversification, but one would expect coalesced interests to arise when the costs from dispersion are so high that they dwarf the costs of coalescing shareholders.

The obvious American mode of coalescing stockholders, overcoming free-rider issues, and then directing and controlling managers was the takeover—a common phenomenon of the 1980s. An outsider, or another firm, offers to buy up the stock of the target firm. If the target firm was badly run, then its stock price should have sagged. The offering company can buy up the stock, run the firm better, and profit from the transaction.

Takeovers are not the only means of coalescing shareholders. Proxy fights have one team seeking the votes of other shareholders with a view (usually) to gaining control of the board. Leveraged buyouts are another: a management team (or a group of outsiders) borrows heavily to buy out a division of a conglomerate (or an entire firm). Ownership concentrates.

²Allocative efficiency may be a more important economic task facilitated by good information distribution. Accurate information flow helps both allocative efficiency and corporate governance.

Institutional investor voice is another potential means of coalescing shareholders. Institutional investors could complain to boards and send around proposals for better governance. Generally speaking, institutional investors are reluctant to act on governance specifics. They tend to be passive in governance (Rock, 1991), either because their compensation does not reward them for activism, because they do not have enough information or skill to be effective, because they do not want to offend boards and executives, or because their blocks of stock (although bigger than they once were) are still not big enough to incentivize action—they suffer from the same free-rider disincentives that affect scattered individual stockholders. However, institutional investors do consider whether to support or oppose activist investors (Bratton, 2007). Often the institutions' support is crucial to success (Gilson & Gordon, 2013). This structure, which has become important, resembles in the abstract the propose-and-ratify decision-making mentioned earlier in the discussion about boards.

Blockholding also coalesces shareholder interests. A large investor takes a big block of stock, then often sits in the boardroom, and thereby has the means and the motivation to make managers work better and smarter for shareholders (Bethel et al., 1998). Blockholding would then raise the issues of horizontal governance: it would be most efficacious if the system minimized the horizontal problems so that the blockholders could then act on the vertical issues. For very large firms, only very rich people or very large financial institutions could play this role. In the United States, the “supply” of very rich people is (surprisingly) relatively small against the number of firms that could have blocks. In fact, many of the richest Americans *already* have blocks (think of Musk, the Waltons, and Zuckerberg). But there are *many* other diffusely owned firms available for blockholding. Among the largest American firms, only a few Americans could take big blocks, but the few who could are *already* occupied with their own blocks at their own firms. Odd as it is to think so, there's a relative “shortage” of very rich potential individual blockholders relative to the number of very large American firms.

Financial institutions were taken off the table as American blockholders early on, due to historical, political, and social considerations. At the time of the rise of the large American firm, at the end of the nineteenth century and the beginning of the twentieth century, the financial institutions in play were the banks and the large life insurance companies. But American populism stopped this form of blockholding (common in modern Germany and Japan), by denying these institutions' national scope (such as via cross-state branching restrictions), restricting their panoply of products (such as the nineteenth National Bank Act and the more famous Glass-Steagall stock ownership and trading restrictions), and by barring their basic authority to own stock (such as via the laws resulting from the 1906 Armstrong investigation for life insurers and via the banking regulatory statutes generally) (see Roe, 1994).

Blockholders also raise legitimacy problems. Historically the American polity did not tolerate large financial institutions. The takeover coalesced scattered shareholders in the 1980s, and it's plausible that takeovers are not as widespread as they otherwise would be in the United States because of legitimacy issues and a negative public reaction (see Roe, 1994; Romano, 1988). (There were structural

reasons why takeovers diminished as well: many 1980's takeovers dismantled conglomerates or dealt with operating issues in the petroleum industry. Once they were dismantled or resolved, there was less reason for takeovers.) And even today a negative reaction to perceptions of institutional investor power may make some governance structures less legitimate than they otherwise would be (Coates, 2023).

21.3.5 Executive Compensation

Compensation, especially stock-based compensation, would seem a quite promising way to align senior managers with shareholders and thereby reduce the vertical corporate governance problem.

The theory is simple: managers get, say, options to buy the company's stock. If their management of the firm induces the firm's stock price to rise, the managers cash in valuable options and make money. If their management of the firm doesn't induce its stock price to rise, then the managers' options are without value. (Note the interplay with the institutions of information disclosure and gatekeeping: for incentive compensation to be effective, stock analysts must have good information so that the stock on which the options are written is accurately priced. Or for it to work well, the board must be sufficiently functional to make incentive-compatible compensation deals and then monitor performance.)

This is the theory, and incentive compensation certainly is an important institution of corporate governance. But it too has its limits. First, with shareholders scattered and not themselves writing the incentive contract, the options may not fully motivate managers but may successfully enrich them. Since the typical options contracts seem to fall short of an ideal contract, there's a good reason to think that the options actually used don't resolve the vertical governance problem as well as they could. (They have typically been based on a general rise in the company's stock price and not on its performance compared to that of other firms.) (Bebchuk et al., 2002; Jensen & Murphy, 1990). True, others conclude that incentive compensation over the long run approximately correlates with performance. Over approximately a decade, managers at strongly performing companies do indeed earn much more than managers at weakly performing companies (Hall & Liebman, 1998; Murphy, 2002) (on the general issues, see Aggarwal & Samwick, 1999; Murphy, 1999; Core et al., 2003).

A second problem with the incentive compensation institution is not noted in the literature but ought to be. Even if the compensation is nicely attuned to performance, senior managers quickly become rich. But the successful managers do not, once they become multimillionaires, turn the firm over to a new crew of hungry but impoverished managers. They instead continue, while wealthy, running the firm. For these people surely more money remains better than less money, but one can question how strongly incentive compensation motivates truly wealthy people. (Its role might then be more to motivate managers *seeking* to be truly wealthy—those lower down in the firm's hierarchy—*not* those already there.)

21.3.6 Professionalism and Norms

Managers do not act solely for remuneration but also for the satisfaction of doing a good job. And “doing a good job” is defined by circumstance, psychology, and culture. Professionals want the firm to do well. They are often acculturated to work for shareholders and often do so even when the other institutional constraints are not tight. Or, managers who get a new product out, who succeed, who innovate, and who turn a division around feel good about themselves and in turn shareholders profit. Or, organizations depend on managers’ capacity for collaboration, cooperation, and trust. These notions of norms and professionalism are softer and less well understood than the other institutions, but that doesn’t mean that they’re absent or that they’re unimportant.

21.3.7 Corporate Lawsuits

Directors can be sued. In the United States, there are two broad bases for stockholders to sue directors and managers: breach of state law fiduciary duties and breach of federal securities law obligations. The latter relate most basically to the company’s quality of information disclosure and can involve SEC enforcement actions as well as private lawsuits. Significant resources in the United States are spent on *private* lawsuits, a feature not matched even in nations like Britain that are usually seen to have similar legal structures. Many directors dread the aggressive adversarial questioning of a long deposition in a securities lawsuit.

In the United States, lawsuits arising from state corporate law largely grow from fiduciary duties. Controllers who steal from the firm have typically violated one of those duties. Controllers who divert business opportunities from the firm to themselves will typically violate one of those duties.

It’s here that market restraints are not likely to strongly reduce insiders’ misbehavior, because the value moved might not be large in relation to the size of the firm or because the value even if large is a one-shot deal that the market will not adequately punish. Moreover, markets cannot work well if they do not know about the wrongdoing, and law-based institutions typically force the information to be disseminated. In the United States, these disclosure-forcing rules primarily come from federal securities law, not state law.

State lawsuits don’t go much further though. Managers may make poor decisions on behalf of shareholders, but judges won’t hear their complaints. The American business judgment rule precludes shareholder actions against directors for mistakes. Since managers can lose much money for shareholders by mistakes, corporate lawsuits are less important than the other institutions of corporate governance. Thus key aspects of vertical corporate governance—“shirking”—aren’t governed by shareholder lawsuits.

One might see the institutions, especially markets and lawsuits, as specializing. Lawsuits are more effective in controlling conflicts of interest than are markets; market institutions, coalescing institutions, and the incentive institutions are better

primed to channel managers toward pro-shareholder decision-making. The first problem of conflicts of interest are addressed by American fiduciary duties; the latter problem of directing managers toward profitability is something that corporate law does not much deal with directly and often ignores. Thus, the domain of the once-burgeoning “law and finance” literature (e.g., La Porta et al., 1999) may be more limited than generally seen in the finance literature, as basic corporate law deals with one core problem of corporate governance—that of diversions or “stealing”—but only indirectly with “shirking” and bad business decisions or with managers who lose sight of shareholder interests. (That is, profitability issues are addressed indirectly, via institutions such as who elects the directors and not via corporate lawsuits.) Other institutions beyond basic corporate law must be key in reducing “shirking.”

21.3.8 Complements and Substitutes

Are these imperfect corporate governance institutions complements or substitutes? Or both?

Clearly each can independently affect the quality of management. True, some primarily affect insider diversions of value, and others align managers with shareholders. But each of those institutions that induce managers to produce the goods for shareholders can operate independently and substitute for one another. Substitution effects influence institutional arrangements: if one institution is demeaned in a nation or becomes too costly for ancillary reasons, then the demand for the other corporate governance institutions should rise.

Consider product markets. Product markets confine managerial discretion, and, in that sense, product markets and the other governance institutions are substitutes. But better product market competition can enhance the other institutions: fierce competition can sharply lower corporate profitability if the managers misstep. Lower profits should lower the firm’s stock price, make the incentive compensation less remunerative, and signal takeover entrepreneurs that an opportunity is brewing. Thus, in the United States, which historically has been seen as having had fairly strong product market competition (at least as compared to pre-EU Europe), some large firms could prosper without really strong constraints on managers because product market competition kept managers working effectively.

Or, take the opposite causal direction. Markets might be structured competitively, but *all* incumbent managers could be lackadaisical. Charge up one institution or another of corporate governance in some of the competing firms—say, incentive compensation or a more dynamic board—and then managers in those firms ramp up production and innovation, grabbing market share away from other firms. To survive, the other competing firms must react. They react by innovating as well, and some of that innovation is, say, better than R&D, but some of it is also to improve *their* own internal corporate governance systems, with better boards, better compensation, and so on.

Or take the complementarity of information disclosure. Better information leads investors to price securities more accurately than if they had inferior information. Better information could then “prime” the pump of other institutions. Managers who are professionals would be chagrined to see their stock price decline. Managers who are looking at the value of their stock options would be motivated to make the value of those options rise. Boards who got the feedback of declining stock prices would be on notice that action might be needed. Takeover entrepreneurs might also be primed to act.

Or better boards could design better compensation plans. They could foster the professionalism of managers. Better boards could induce a capital structure geared toward the firm’s product markets and management.

And so on.

21.4 Contract Law as the Corporate Law “Primitive”?

In a handbook on institutions, one ought to inquire whether corporate law institutions can be broken down further toward a “primitive”—a basic institution on which the corporate governance institutions rest, like chemists looking inside the gas to find the molecule, then inside the molecule to find the atom. Are there more basic building blocks than corporate law and similar institutions, and are corporate governance institutions built atop these more basic blocks?

21.4.1 Corporate Law as Standardized Corporate Contract

Consider contract. It’s a basic exercise in teaching corporate law to analyze how most of the corporate charter—the document governing relations among shareholders and between the board and its shareholders—could be created out of contract law, without a separate corporate law. (And to then analyze those aspects of corporate law that would be exceedingly expensive to create out of contract law (Hansmann & Kraakman, 2000).) The rules governing, say, meeting frequency, voting rights, the mechanism of vote solicitation, the question of who pays for the proxy solicitation, and so on could all be created via contract. And, indeed the contractarian view of corporate law is an important theoretical strain here: in its normative form, contractarians maintain that corporate law should be a pure contract. Corporate law should be the contract that the shareholders and managers would have come to naturally if bargaining and transaction costs were cheap. Or, it should be the charter that the parties would find easiest to bargain around (Easterbrook & Fischel, 1991).

Moreover, the institutions that minimize the opportunism in the corporate charter could be seen as contract-based institutions. Fiduciary duties, in this view, are the terms that the parties, had they anticipated the questionable transaction, would have negotiated toward (Easterbrook & Fischel, 1991; but cf. Brudney, 1985).

Hence, if contract law is good, then (a) much that is useful can be done by contract and (b) the primitive institution that could make for good corporate law is in place.

21.4.2 Contract Law's Limits

But contract law has its limits in making the corporation work here. Corporate law has its criminal component. Insider trading can destabilize share ownership. It can also land the perpetrator in jail. A contract against insider trading cannot send a violator to jail.

And corporate contracting entails multi-party contracts that the players cannot immediately standardize. Corporate contract could, one supposes, lead to audit firms and audit firm liability. But building a system up from the two-party contract would not immediately lead to standardized financial information. Even here though, standardized formats may eventually arise without more than contract law. Private associations, such as stock exchanges in the United States, can standardize rules and formats. There's evidence that stock exchanges and other organizations came first, followed by corporate (securities) laws later (Cheffins, 2001; Coffee, 2001; Macey & Kanda, 1990; Mahoney, 1997; Miwa & Ramseyer, 2002; Roe, 2000).

21.5 Institutional Legitimacy: CSR, ESG, and Corporate Purpose

Thus far we have focused on the “internal” effects and purposes of corporate governance institutions, principally those that align managerial and shareholder interests and secondarily those that prevent massive diversions of value from dispersed shareholders to the insiders. These institutions are needed to stabilize the corporation.

Corporate governance institutions have another role. Players outside the corporation can affect the corporation's internal dynamics: if corporate arrangements appear unfair, insiders can appeal to outsiders in political institutions to intervene to alter the internal relationships of corporate governance. They can ban some arrangements, raise the costs of others, and subsidize others. They can, and they do, everywhere in the world. When they do so, they can deeply affect the institutions of corporate governance, as illustrated in Fig. 21.3.

And players inside the corporation who cannot get what they want via internal arrangements might appeal to the outsiders, making external political alliances that then press the firm internally. They leverage their positions inside the firm by calling on institutions outside the firm.

The institutions of corporate governance are not just organizational and technical. They are, or they are affected by, political institutions as well. Three examples follow: one from American history, the second from European post-World War II left-right politics, and the third from modern corporate purpose pressures for the large firm to be more socially responsible than profit orientation would ordinarily induce.

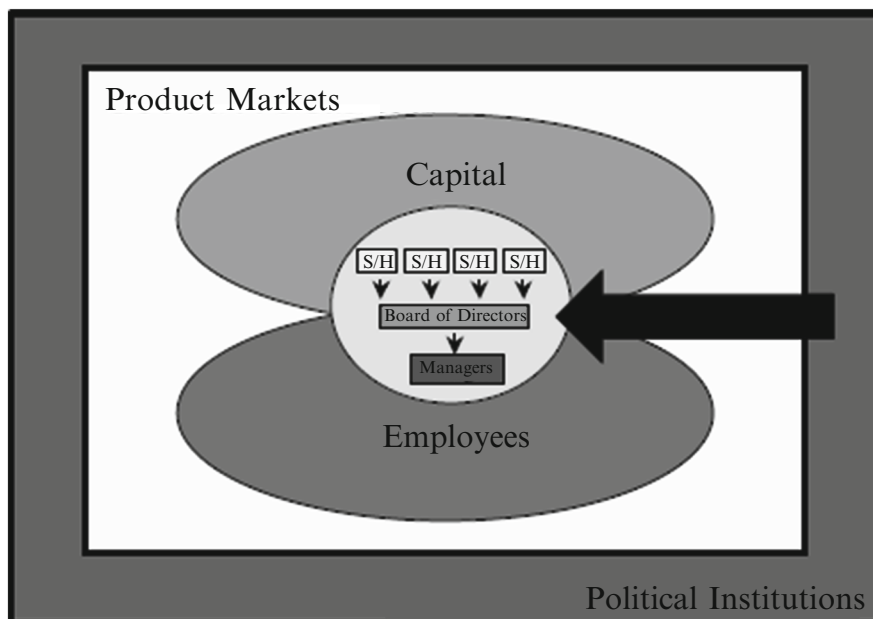


Fig. 21.3 The corporate governance environment

21.5.1 American Legitimacy Light: Populism and Anti-Takeover Laws

Takeovers constrained managers in the 1980s. But managers called on political allies—labor at target firms, bystander voters who sympathized with target firms and their employees, and politicians who saw more votes in opposing takeovers than in promoting them (Roe, 1993, 1994; Romano, 1988).

More fundamentally, and further back in American history, American populism made visibly powerful financial institutions such as J.P. Morgan's end-of-the-nineteenth century investment bank or a Japanese-style main bank or a German-style universal bank incompatible with American political culture (of an earlier era) (Roe, 1994). The consequence was that one form of concentrated ownership—that of financial institutions—was largely removed from the menu by American politics. This made the diffusely owned firm inevitable in the United States.

21.5.2 Legitimacy Heavy: European Left-Right Politics

Left-right political orientation affects corporate governance institutions. Ownership by diffuse, dispersed shareholders calls for tools that keep managers loyal to those

distant shareholders. (Or at least shareholders do better if they get some of these tools.) Incentive compensation, shareholder primacy norms, and other shareholder-enhancing mechanisms need to be efficacious. But if a nation was lodged toward the left of the modern political spectrum, political institutions tended not to provide these supports.

Nations lodged toward the left disrupted managerial-shareholder alliances and promoted activities that favor employees with existing jobs. Firms were encouraged to expand without regard to profitability, to delay downsizing when their production is misaligned with the product market, and to go slow in taking profitable but disruptive risks. These pressures, and the denigration of some pro-shareholder tools, would lead firms to have more concentrated ownership than otherwise, so that managers could be more directly controlled (Roe, 2003). Moreover, these are the kinds of managerial actions that cannot readily be contained by corporate law, which focuses on diversions of value more than pro-shareholder operating decisions (Roe, 2002, 2003). These kinds of political pressures—to expand, to avoid downsizing, and to avoid profitable risks—map exactly onto the kinds of agency costs that are thought to be managers' natural tendencies unless otherwise checked (Jensen, 1986).

In nations where labor institutions—whether via social democracy or corporatist power-sharing or other cooperative arrangements—were strong, one would expect managerial agency costs to shareholders to often have been higher for firms that had ownership and control divided than in nations where such labor institutions were weaker.

Two channels were in play, one through the firm and the other through institution-building: First, through the firm, the polity would encourage the firm to expand, irrespective of profitability, and would impede it from downsizing when its capabilities are misaligned with markets. And, there'd be more bargaining over the surplus, with some of that bargaining at the national political level and some inside the firm. Concentrated ownership would be more profitable for shareholders than in other polities, because the concentrated owner could bargain more effectively and resist important political pressures.

Second, nations where labor held significant political power could be unwilling to build the institutions that facilitate distant shareholding. They would be unwilling to devote resources to building good securities regulation, to promoting profit-building institutions, to facilitating shareholder control over (or influence on) managers, and to enhancing shareholder primacy norms that induce managers to align themselves with stockholders.

If one or both of these channels is strong, then one could hypothesize a basic model with testable implications. Greater labor protection should predict weaker ownership separation. Consider these results from OECD data that indexed the level of job protection in the world's wealthiest nations.

Substantial evidence exists for the strong pressure of politics on corporate governance institutions. Look at the world's richest nations, and compare the degree of ownership separation with the strength of labor protection. The correlation is

powerful. And even when controlling for the quality of legal institutions and the size of the firms involved, the correlation is quite robust.

In Fig. 21.4, I show the relationship between national employment protection laws in the wealthy West in 1995 and the degree of ownership separation in a sample of similarly sized firms from each nation. The figure shows the strength of employment protection laws—a rough measure of the left-right shift described earlier in this subsection—in predicting ownership separation. Other measures correlate similarly (see Roe, 2000, 2003).

For efforts to link political and corporate institutions, see Roe (1994, 2003), Pagano and Volpin (2002), Perotti and von Thadden (2003), Rajan and Zingales (2003).

This figure and the technical table show that the OECD’s measure of employment protection negatively correlates with the 1990’s ownership dispersion at a high level of significance. The R-squared result indicates that a large portion of the variation in ownership is potentially explained by the variation in employment protection.

Technical data: med20 v. employment protection	
Regression	$y = -0.04x + 0.65$
Adjusted R-squared	0.64
t-stat	-5.24 ^a

^aSignificant at the 0.0005 level. Sources: OECD (1994) (employment protection); La Porta et al. (1999) (ownership dispersion of mid-sized firms: percentage of sample of medium-sized firms without a dominant, 20% + stockholder)

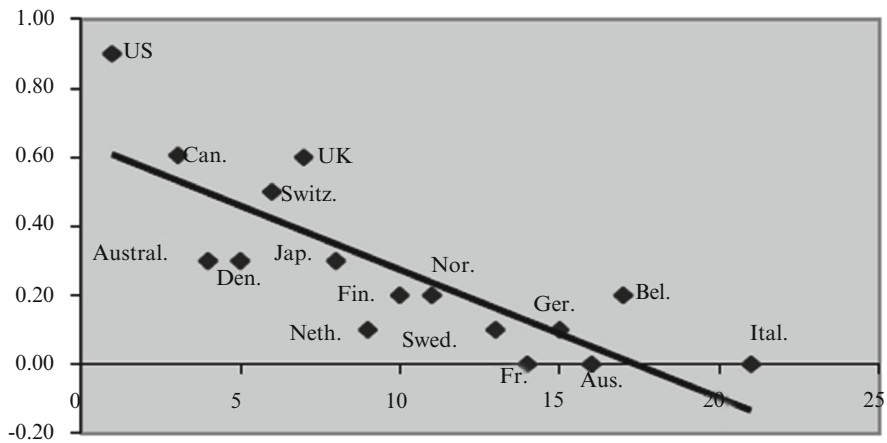


Fig. 21.4 Employment protection vs. ownership dispersion. Y-axis is dispersion; X-axis is employment protection

21.5.3 Corporate Social Responsibility

In the past decade or so, corporate social responsibility pressures on corporate governance institutions have intensified, with a central document being the 2018 letter from the chair of BlackRock—the huge institutional shareholder—pushing public companies to act on “the societal impact of your business as well as the ways that broad, structural trends [such as] climate change” impact the economy and society. Said BlackRock’s chair, Larry Fink: “Society is demanding that companies . . . serve a social purpose. . . . Companies must benefit all of their stakeholders . . .”

Proxy fights have had hydrocarbon burning as a central consideration (e.g., ExxonMobil). Institutional investors in the 2010s pressured firms to reduce carbon emissions and to have women and minorities better represented on boards. Institutional investors often pushed forward the CSR/ESG (for environment, social, and governance) agenda in the early 2010s. Markets and competition limit how far some firms can go down this road (Roe, 2021). Moreover, as interests and opinion struck back, the institutional investors became less animated in pushing this agenda forward.

The SEC brought forward climate disclosure rules that had two justifications: one, some proponents said that with climate change having become such a central economic issue in upcoming years, investors need the information to judge a company’s exposure to climate change risks. Others thought that the proposed disclosure rules were backdoor ways to pressure firms into being greener and to act on climate change. Some thought this result was proper; others thought it illegitimate.

Some CSR/ESG proponents couched their criticisms in terms of capitalist goals: companies that failed to deal with social issues were too short-term and deficient as capitalist firms; firms that dealt with these social issues were long-term representatives of sound capitalism. Some showed that the core issue here was not one of time horizons (i.e., not whether the firm was short- or long-term oriented) but how willing a firm was to externalize environmental and climate costs on the rest of society (Roe, 2022). Externalities are not principally time issues; externalities come from a business’s capacity to profitably push costs of a business onto outsiders.

21.6 Corporate Institutions in Their Setting

Thus far we have examined the corporate institutions of corporate governance. In addition, the strength, weakness, and nature of corporate governance institutions can be, and often is, derivative of more foundational institutions of contract and property.

21.6.1 Contract

1. *Firm size.* Corporate governance institutions have technological predicates. Corporate governance institutions govern the interface between the capital

providers and the firm's management (at least as we've defined the essential corporate governance problem). For there to be a problem to be governed, especially for vertical corporate governance between shareholders and distant managers, there must be a demand for firms where not all capital providers are inside the firm. For that condition to arise, there must be a heavy demand for capital inside the firm, and there must be technologies that demand a size of operation beyond that which managers would ordinarily be able to control directly with their own capital.

In its simple form, there thus must be economies of scale. But in addition, those economies of scale must induce capital needs greater than that which a single investor or a very small group of investors could provide.

The historical story usually told is that advances in communication, production, and transportation technologies facilitated the growth of the large, vertically integrated firm (Chandler, 1990). These new organizations had capital demands that led to the large public firm and its concomitant corporate governance problems, raising the twentieth-century demand for better standard corporate governance institutions. In its theoretical form, one would cite Coase's *Nature of the Firm* article: Economic activity can be organized inside the firm or across firms. A single firm may do the transaction of mining the raw material and then manufacturing it. Or one firm might mine the raw material and then sell it to the manufacturer. The same task is performed but via different transactions. The relative cost of the transaction determines whether the task will be performed inside a single firm or across a market involving more than one firm. When the transaction costs of bringing the next transaction inside the firm exceed the costs of contracting through the market, the firm stops growing (Coase, 1937; Williamson, 1985).

2. *Decentralization vs. centralization.* This transaction cost story can be turned around. If new technologies *reduce* the costs of transacting in the market *across* firms—new Internet and information technologies come to mind—then the demand for large integrated organizations would decline. If they declined enough, the needed capital might be within reach of a handful of managers, making the separation problem less acute. One would still have institutional issues—those say of relations among a network (cf. Piore & Sabel, 1984), perhaps of a contractual dimension—but they wouldn't be the classic vertical and horizontal problems of corporate governance. One might see the computer's advance in the past half-century as having corporate governance implications. In the age of the mainframe, information was best processed at the central headquarters. The centralized corporation, sometimes called the M-form corporation, had much strategic decision-making centralized. It had its heyday decades ago. Centralization fit well with the mainframe. But in the age of the personal computer and the Internet, information is better distributed, and decentralized decision-making becomes favored. Firms decompose, spinning off unrelated operations, as each operation can process its own information well, without the need of a centralized mainframe. Or the management's ability to get good information fast facilitates more centralized

management. Technology, if it strongly tilted one way or the other, could determine whether and which corporate governance issues are important.

3. *The quality of contract.* And one could see another role for law here. If the institutions of contract improve greatly relative to the institutions of corporate governance, then one would expect to see pressure on the size of firms. Corporate governance rises in importance when the transactions to be governed cannot be handled well informally or via contract.³ If production needs to be brought inside a single firm, then the firm must be larger, and that puts pressure on the institutions of corporate governance. (On the need for some transactions and production to be in a single firm, think of the vertical integration that was central to the mid-twentieth-century American economy. The classic Fisher Body problem exemplifies the pressure to bring potentially separate production into a single firm (Klein et al., 1978). But if contractual capacity improves (because of, say, better judges, better information on contractual performance, better third-party verifiers, and so on), then the units of production can be kept in separate firms. When they're in separate firms, the problems of horizontal and vertical corporate governance (to the extent those problems arise from large-sized firms) diminish; contract takes over from the institutions of corporate governance.)
4. *Markets again.* This trade-off between firm and contract as the mode of governance sets us up for another market-oriented perspective. Markets constrain corporate governance failures. Firms must operate within product, capital, and managerial labor markets, each of which we have seen prods the firm toward efficiency and toward using the best corporate governance institutions.

The market can constrain corporate governance failures in another way. The firm's *internal* governance system brings activities into the firm that could have been transacted *across* the market. The firm is not just competing in the product market with other firms that have brought a similar quantity and scope of activities inside the organization (an organization that hence needs governance institutions to make it work well).

The firm *also* competes with disaggregated units that make the same product. If the disaggregated units can produce the same quality product cheaper, etc., then more resources will go to the disaggregated units, fewer resources to the integrated units, until the integrated units learn—presumably via better corporate governance institutions—how to reduce their governance costs, improve their organization, etc. This mechanism is the Coasean theory of the firm, applied to corporate governance institutions.

³Consider this small contract riddle: Part of the technology that favors decentralized production is better contract law. If contract law (and related contract institutions) improved, while holding the quality of corporate law constant, that improvement would favor operational decentralization. But better contract law is the corporate law primitive, and corporate law is one of the institutions that facilitates large organizations. Hence, the two should move in tandem, not separately. And if both improve simultaneously, the improvement's effects on relative decentralization are unpredictable.

21.6.2 Property Rights

Ownership separation has two related underlying property rights problems. The first is simple: for firms to get capital, there must be enough savings somewhere in the system to invest. But if property rights are fragile, savers will be unwilling to save, fearing confiscation (cf. Acemoglu et al., 2002; Mahoney, 2001; North & Weingast, 1989). As Williamson (1985) states: “the suppliers of finance [face the risk that t]he whole of their investment in the firm is potentially placed at hazard. By contrast, the productive assets (plant and equipment; human capital) of suppliers of raw material, labor, intermediate product, electric power, and the like normally remain in the suppliers’ possession.” Some other mechanism of gathering capital would be needed, such as the State, and what now passes for corporate governance problems in the wealthy West would focus on the relationship between the State as capital provider and large enterprises.

The second property rights problem is more nuanced but similar: Separation of ownership from control exposes capital providers. They provide their capital to the firm, but might not get it back. Managers can take it for themselves. Other shareholders might maneuver to grab it. These two risks to investors are the usual focus for corporate governance. But we shouldn’t stop there. The State, or employees, might figure out how and why that capital really belongs, or should belong, in their own pockets. When investors feel unprotected in this dimension, they are likely to adopt institutional arrangements that minimize their losses to the State or to arrangements favored by the State.

Thus, one can find correlations between, say, weak corporate law and weak ownership separation. And while it’s logically possible that it’s the weakness in *corporate* law that directly weakens separation, because investors fear dominant stockholders, it’s equally logical that *it’s a basic weakness in property rights in that society that deters such investment. Investors could fear that others, such as the State or its patrons, will grab the investment.* There must be property rights institutions that make savers relatively comfortable with saving and with visibly investing in distant enterprise. If they fear expropriation, they won’t put their savings into public view even if they trust the firm’s controllers not to steal from them. In much of the world, such property rights are not in place. In some parts of the wealthy West, putting the money out there in some corporate forms—that are public and politically visible—is riskier than in other forms (that are private and not easily observed). Hence, there is less reason in such nations to develop strong institutions of corporate governance, because even if developed they wouldn’t be widely used.

One might go further. Once basic property rights institutions are in place, corporate law institutions can adapt, usually easily. In this sense, corporate law institutions are not central to social and business organization—to the chagrin perhaps of the corporate law professor—but basic property rights are. If property rights are secure, and if the society has the “primitive” foundation on the ground of good basic contract law, then the institutions of corporate governance become (relatively) easy to build. Conversely, if property rights are not secure from state action, then western corporate governance institutions are harder to develop. Recent

thinking about China illustrates this weak property rights aspect. With property rights weak, investors are less likely to invest, and entrepreneurs less likely to devote their energies to building big companies that the state may well take over (Fried & Kamar, 2021; Milhaupt, 1998; Posen, 2023). Corporate governance institutions here are derivative and secondary to the more basic institutions of property and contract.

21.7 Summary

We have reviewed the institutions of corporate governance. The institutions in general respond to two distinct problems, one of vertical governance (between distant shareholders and managers) and another of horizontal governance (between a close, controlling shareholder and distant shareholders).

Some institutions deal well with vertical corporate governance but do less well with horizontal governance. The institutions interact as complements and substitutes, and many can be seen as developing out of a “primitive” of contract law. Arguably a successful market-based economic system must first have satisfactory contract enforcement, as well as basic property rights that shield private actors from the state before the economic system embarks on more sophisticated corporate governance institutions.

Institutional legitimacy affects the institutions of corporate governance. In the United States, intervention inside large firms from powerful financial institutions was seen as politically illegitimate and was largely banned or made more costly. In modern Europe, the institutions of shareholder value have been denigrated. Each legitimacy issue affects corporate institutions, as firms seek substitutes defensively, for example. In recent decades legitimacy has become tied to corporate social responsibility considerations. The interaction between political institutions and corporate governance institutions is an inquiry still in its infancy but promises large returns. Even simple regressions suggest that political institutions may strongly influence the construction and survival of corporate governance institutions.

Acknowledgment of Funding Open access is made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., Johnson, S., & Robinson, J. A. (2002). Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics*, 118, 1231–1294.
- Aggarwal, R., & Samwick, A. (1999). Executive compensation, strategic competition, and relative performance evaluation: Theory and evidence. *Journal of Finance*, 54, 1999–2043.
- Bebchuk, L. A., Fried, J. M., & Walker, D. I. (2002). Managerial power and rent extraction in the Design of Executive Compensation. *University of Chicago Law Review*, 69, 751–846.
- Bethel, J. E., Liebeskind, J. P., & Opler, T. (1998). Block share purchases and corporate performance. *Journal of Finance*, 53, 605–634.

- Bhagat, S., & Black, B. S. (1999). The uncertain relationship between board composition and firm performance. *Business Lawyer*, 54, 921–963.
- Black, B., Kraakman, R., & Tarassova, A. (2000). Russian privatization and corporate governance: What went wrong. *Stanford Law Review*, 52, 1731–1808.
- Blair, M. M., & Stout, L. (1999). A team production theory of corporate law. *Virginia Law Review*, 85, 247–328.
- Bratton, W. W. (2007). Hedge funds and governance targets. *Georgetown Law Journal*, 95, 1375–1433.
- Brudney, V. (1985). Corporate governance, agency costs, and the rhetoric of contract. *Columbia Law Review*, 85, 1403–1444.
- Chandler, A. P. (1990). *Scale and scope: The dynamics of industrial capitalism*. Cambridge, MA.
- Cheffins, B. R. (2001). Law, economics and the UK's system of corporate governance: Lessons from history. *Journal of Corporate Law Studies*, 1, 71–89.
- Coase, R. H. (1937). The nature of the firm. *Economica*, 4, 386–405.
- Coates, J. (2023). *The problem of twelve*. Columbia University Press.
- Coffee, J. C. (2001). The rise of dispersed ownership: Theories of law and the state in the separation of ownership and control. *Yale Law Journal*, 111, 1–82.
- Core, J. E., Guay, W. R., & Larcker, D. F. (2003). Executive equity compensation and incentives: A survey. *FRBNY Economic Policy Review*, (April), 27–50.
- Davis, G. F. (1993). Who gets ahead in the market for corporate directors: The political economy of multiple board memberships. *Academy of Management Best Papers Proceedings*, 1993, 202–211.
- Denis, D. K., & McConnell, J. J. (2003). International corporate governance. *Journal of Financial and Quantitative Analysis*, 38, 1–36.
- Easterbrook, F. H., & Fischel, D. R. (1991). *The economic structure of corporate law*. Harvard University Press.
- Eisenberg, T., Sundgren, S., & Wells, M. T. (1998). Larger board size and decreasing firm value. *Journal of Financial Economics*, 48, 35–54.
- Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. *Journal of Law and Economics*, 26, 301–325.
- Franks, J., Mayer, C., & Renneboog, L. (2001). Who disciplines Management in Poorly Performing Company? *Journal of Financial Intermediation*, 10, 209–248.
- Fried, J., & Kamar, E. (2021). Alibaba: A case study of synthetic control. *Harvard Business Law Review*, 11, 279–309.
- Gilson, R., & Gordon, J. N. (2013). The agency costs of agency capitalism. *Columbia Law Review*, 113, 863–927.
- Gilson, R., & Kraakman, R. (1991). Reinventing the outside director: An agenda for institutional investors. *Stanford Law Review*, 43, 863–906.
- Hall, B. J., & Liebman, J. B. (1998). Are CEOs really paid like bureaucrats? *The Quarterly Journal of Economics*, 63, 653–691.
- Hall, P. A., & Soskice, D. (Eds.). (2001). *Varieties of capitalism: The institutional foundations of comparative advantage*. Oxford University Press.
- Hansmann, H., & Kraakman, R. (2000). The essential role of organizational law. *Yale Law Journal*, 110, 387–440.
- Hart, O., & Zingales, L. (2017). Companies should maximize shareholder welfare not market value. *Journal of Law, Finance, and Accounting*, 2, 247–274.
- Henderson, R. H., & Clark, K. B. (1990). Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Administrative Science Quarterly*, 35, 9–30.
- Hermalin, B. E., & Weisbach, M. S. (1998). Endogenously chosen boards of directors and their monitoring of the CEO. *American Economic Review*, 88, 96–118.
- Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. *American Economic Review Papers and Proceedings*, 76, 323–329.

- Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3, 305–360.
- Jensen, M. C., & Murphy, K. J. (1990). Performance pay and top-management incentives. *Journal of Political Economy*, 98, 225–264.
- Klein, B., Crawford, R., & Alchian, A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21, 297–326.
- Kraakman, R. (1986). Gatekeepers: The anatomy of a third-party enforcement strategy. *Journal of Law, Economics & Organization*, 2, 53–104.
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. (1997). Legal determinants of external finance. *Journal of Finance*, 52, 1131–1155.
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. (1998). Law and finance. *Journal of Political Economy*, 106, 1113–1150.
- La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (1999). Corporate ownership around the world. *Journal of Finance*, 54, 471–517.
- Lorsch, J., & MacIver, E. (1989). *Pawns or potentates: The reality of America's corporate boards*. Harvard Business School Press.
- Macey, J., & Kanda, H. (1990). The stock exchange as a firm: The emergence of close substitutes for the New York and Tokyo stock exchanges. *Cornell Law Review*, 75, 1007–1052.
- Mahoney, P. G. (1997). The exchange as regulator. *Virginia Law Review*, 83, 1453–1500.
- Mahoney, P. G. (2001). The common law and economic growth: Hayek might be right. *The Journal of Legal Studies*, 30, 503–525.
- Milhaupt, C. (1998). Property rights in firms. *Virginia Law Review*, 84, 1145–1194.
- Miwa, Y., & Ramseyer, M. (2002). The value of prominent directors: Corporate governance and Bank access in transitional Japan. *The Journal of Legal Studies*, 31, 273–301.
- Murphy, K. J. (1999). Executive Compensation. In Ashenfelter & Card (Eds.), *Handbook of labor economics* (Vol. 3). North Holland.
- Murphy, K. J. (2002). Explaining executive compensation: Managerial power vs. the perceived cost of stock options. *University of Chicago Law Review*, 69, 847–j869.
- North, D. C., & Weingast, B. R. (1989). Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century England. *Journal of Economic History*, 49, 803–832.
- OECD. (1994). *The OECD jobs study: Evidence and explanations, part II: The adjustment potential of the labour market*. OECD.
- Pagano, M., & Volpin, P. F. (2002). The political economy of finance. *Oxford Review of Economic Policy*, 17, 502–519.
- Perotti, E. C., & von Thadden, E.-L. (2003). The political economy of corporate control and labor rents. *Journal of Political Economy*, 114, 145–174.
- Piore, M. J., & Sabel, C. F. (1984). *The second industrial divide: Possibilities for prosperity*. Basic.
- Posen, A. S. (2023). The end of China's economic miracle. *Foreign Affairs*, 102, 118–130.
- Puzder, A. (2023). Red states have slowed the ESG juggernaut. *Wall Street Journal*, 14(June), 2023.
- Rajan, R. G., & Zingales, L. (2003). The great reversals: The politics of financial development in the twentieth century. *Journal of Financial Economics*, 69, 5–50.
- Rock, E. B. (1991). The logic and (uncertain) significance of institutional shareholder activism. *Georgetown Law Journal*, 79, 445–506.
- Roe, M. J. (1993). Takeover politics. In M. Blair (Ed.), *The Deal decade*. Brookings.
- Roe, M. J. (1994). *Strong managers, weak owners: The political roots of American corporate finance*. Princeton University Press.
- Roe, M. J. (2000). Political preconditions to separating ownership from corporate control. *Stanford Law Review*, 53, 539–606.
- Roe, M. J. (2002). Corporate law's limits. *The Journal of Legal Studies*, 31, 233–271.
- Roe, M. J. (2003). *Political determinants of corporate governance: Political conflict, corporate impact*. Oxford University Press.

- Roe, M. J. (2021). Corporate purpose and corporate competition. *Washington University Law Review*, 99, 223–270.
- Roe, M. J. (2022). *Missing the target*. Oxford University Press.
- Roe, M. J., & Shapira, R. (2021). The power of the narrative in corporate lawmaking. *Harvard Business Law Review*, 11, 233–278.
- Romano, R. (1988). The future of hostile takeovers: Legislation and public opinion. *University of Cincinnati Law Review*, 57, 457–505.
- Spamann, H. (2022). Indirect investor protection: The investment ecosystem and its legal underpinnings. *Journal of Legal Analysis*, 14, 17–79.
- Useem, M. (1992). *Executive defense: Shareholder power and corporate reorganization*. Harvard University Press.
- Williamson, O. (1985). *The economic institutions of capitalism*. The Free Press.
- Yermack, D. (1996). Higher market valuation of companies with a small Board of Directors. *Journal of Financial Economics*, 40, 185–211.
- Zingales, L. (1997). Corporate governance. In P. Newman (Ed.), *The new Palgrave dictionary of economics and the law*. Oxford University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Erin Anderson and Hubert Gatignon

Contents

22.1	Introduction	536
22.2	How Firms Create New Markets Without New Products	538
22.2.1	The Role of Sales and Distribution Activity	539
22.3	How Firms Develop New Products	540
22.3.1	The Firm's Returns to New Products	540
22.3.2	Processes of Internal Development	540
22.3.3	Development Relying on Third Parties	546
22.3.4	Develop by Acquisition or Appropriation	549
22.4	How Firms Create New Markets by Commercializing New Products	550
22.4.1	Diffusion Theory and Models	551
22.4.2	Marketing Entry Strategy Effects	555
22.5	Reprise	559
	References	561

Abstract

Markets are created by firms, who make decisions about the innovation process and the marketing of innovations. New Institutional Economics, particularly Transaction Cost Economics, is well placed to explain the decisions firms make in highly uncertain and risky businesses, where most innovations fail. Decision-makers' capacity to follow a conscious economic logic to the fullest is considerably strained by factors such as bounded rationality, information impactedness, and the tacit nature of knowledge. Therefore, opportunities for self-interest seeking with guile abound (e.g., by suppliers or business partners), and contracts are insufficient to protect firm-specific assets. We present the issues arising in the process of generating innovations and then adopt the demand perspective to introduce the problems managers face in creating new markets. We show that

E. Anderson · H. Gatignon (✉)
INSEAD, Fontainebleau, France
e-mail: hubert.gatignon@insead.edu

the creation of new markets by firms, both when developing and commercializing innovation, is best understood through the lens of New Institutional Economics.

Keywords

Innovation · Marketing · Partnership · Diffusion process · Foreign entry

22.1 Introduction

New markets do not emerge, nor do they appear. They are *made* by the activities of firms. New markets are created when firms correctly sense (by accident or by design) a latent need and communicate their solution to that need: markets spring into being when economic actors shift resources to that firm's solution. The most visible way to create a new market is to offer a product/service that is novel, thereby addressing needs that were not met (and perhaps not even sensed). Much of this chapter focuses on firms' efforts to develop and commercialize new offerings and on how buyers respond, thereby creating new markets. However, new markets are also created when firms cultivate an underserved clientele with established products. Much of marketing is about how to bring new customers into a developed industry (as opposed to rearranging market shares among existing customers). This chapter will also highlight these market-creation activities.¹

Capitalist systems exhibit an astonishing ability to create new markets (and, typically, destroy existing ones) based on developing and commercializing innovations. Schumpeter (1943) argued that large firms innovate so well that they raise a society's general standard of living. In the same vein, Schmookler (1966) argues that long-term economic growth is primarily the result of better knowledge of what goods would be useful and how to make them, i.e., invention. There is little argument that innovation, on the whole, increases public welfare: new markets are thought to arise because buyers recognize they will be better off. Breshnahan and Gordon (1996) document why, with a series of innovations that clearly increase buyers' utility. It should be noted, however, that many innovations are not radical but merely incremental and that their utility is in the eye of the beholder. Although this chapter emphasizes more radical innovation, it will address new products in general.

It is less clear how much firms actually benefit from attempting to create new markets via new products. Firms develop and launch new products with the intention of increasing profitability. Schmookler (1966) shows that invention is not primarily a

¹ We define a market as the set of all actual and potential buyers of a product or service. Following Williamson (1996), we define a firm as a governance structure, an organizational construction. We treat an industry as a group of sellers (firms) serving a market. In the standard economic perspective, the market is the main central institution, and the firms' black-box actions derive from markets. Here, we treat the firm (which is not a black box) as the central institution. And we view markets as an outcome of corporate activities. This reversal of the standard setup follows from New Institutional Economics.

response to intellectual stimuli but is instead an effort to exploit a profit opportunity. But the payoff to the investing firm is highly uncertain. Frequently, developing firms reap heavy costs, but other firms capture the projected benefits—if indeed, these are not competed away in a ruinous race to build new markets and establish dominance. This may explain why Griliches, Hall, and Pakes (1991) find no relationship between a firm's patent counts and its financial performance. Bayus, Erickson, and Jacobson (2003) note that it is an article of faith among businesspeople that new product development (hereafter, NPD) is essential to firm performance but that the evidence is mixed as to why this is true and how strong and long-lasting is the effect.

In short, NPD is inherently uncertain: new markets frequently do not materialize, nor do anticipated profits. We contend that the creation of new markets via the creation of new products is best understood through the lens of New Institutional Economics (NIE), leaning on the twin pillars of evolutionary reasoning and Transaction Cost Economics (hereafter, TCE). From TCE, we adopt the premise that firms intend to be rational but are bounded in their abilities: in particular, critical information is impacted, and decision-makers cannot write complex contingent claims contracts that steer toward optimal outcomes. Instead, executives compare concrete, visible alternatives and attempt to foresee which one does the best job of reducing the total of production and transaction costs long term. Further, opportunism is possible and, given a sufficient scale of operation, worth reducing by employing costly governance mechanisms. The coming sections rely on the TCE mechanisms of asset specificity, environmental uncertainty, and internal uncertainty (difficulty of assessing performance using output measures). The control/commitment continuum from arm's-length market contracting to relational governance through to vertical integration will be invoked frequently to explain how firms develop new products in the hope of creating new markets and how they form relations with economic actors (such as distributors) that are vital to the product's success.

From evolutionary economics (Nelson, 1995; Dosi, 1997), we borrow the premise that markets seldom reach equilibrium and that, if they do so, the equilibrium reached is path dependent at the industry level. Further, firms react to uncertainty by developing routines that are difficult to change, routines that reflect the path of their (unique) history. Learning and imitating feature heavily in our analysis.

Strands of these frameworks also appear in a major part of this chapter that reviews how prospects become converted into customers, thereby calling new markets into existence. We rely here on the study of consumer behavior through the combined lenses of sociology, psychology, and economics. We discard the classical economic assumptions that consumers are well-informed, rational maximizers of their own utility, capable of reducing multiple attributes of a product to the lowest common denominator of net utility. We accept that buyers have complex motives, many of them socially determined, and that utility is subjective and individual (including Veblen effects that could hardly be considered rational). At the same time, we do not portray buyers as capricious, random, or wholly ignorant. We accept that buyers seek to improve their subjective utility, subject to bounds on their rationality, and that they do so by comparing observable outcomes and attempting to foresee which best fits their subjective preferences.

NIE has a number of core approaches, including property rights, the contractual nature of the firm, the tragedy of the commons, and allocation among claimants to common resource pools. While we focus primarily on TCE and on evolutionary reasoning, we acknowledge that other NIE elements can be brought to bear to understand how firms create markets.

This chapter is organized as follows. We begin with the mundane: how do firms create new markets from established products? We then turn to the far more perilous task of creating new markets via new products. We first revisit the problem of why firms bother to try, given that the payoffs of NPD are so uncertain. Three ways to develop new products are then considered in turn: internal development of the sort favored by Schumpeter (1943), third-party development via market contracts of the sort criticized by Williamson (1996), and today's vogue of merely acquiring or appropriating new products developed elsewhere. New product development is a vast topic. We focus on those aspects that determine what sorts of new markets a firm can hope to create and offer references to other NPD issues.

Given the new product, how do prospects convert to buyers? A large literature, following the tradition of Griliches (1957), empirically summarizes how innovations diffuse through a target population. We examine this literature, focusing on how the firm's marketing strategy opens (or sometimes forecloses) new markets.

22.2 How Firms Create New Markets Without New Products

If one accepts bounded rationality and opportunism, it follows that economic actors benefit from sheer information and stimulation and that firms can convert prospects into buyers by offering credible reassurance against both adverse selection and moral hazard. Therefore, *mere marketing activity creates new markets*, and does so even with established products. Pawakapan (2000) shows how firms created a market in a remote Thai village merely by sending salespeople with branded merchandise, a pricing schedule, promotional materials, and the authority to book and fulfill orders. The sales force came from the cities and spoke the national language, not the local dialect. Their novel proposition was so appealing, however, that the villagers undertook to learn the national language. Indeed, Pawakapan (2000) judges salespeople more effective than any other means, including schools and political pressures, in teaching villagers the national language (from which follows national culture). These firms quickly created a market for categories of merchandise the villagers had never seen, let alone considered. Given choice, villagers changed their buying habits. Formerly, they bought essentially on price and exhibited loyalty to selected local traders selling strictly local unbranded products. Rapidly, villagers added variety, convenience, and selected brand preferences to their utility function and took to spreading their business across multiple salespeople. Thus, mere marketing activity not only created new markets but altered how markets function.

Spulber (1999) generalizes this idea to any market, even the most sophisticated. He conceives of firms as intermediaries standing between consumers and the

providers of inputs (components, assembly, finance, delivery, and so forth). In Spulber's conception, problems of information are paramount, transaction costs are high, and markets are uncertain. The mere existence of firms motivates prospects to become buyers: firms are the engine of commerce (money being the oil). Firms do so by standing ready to buy and sell (providing liquidity and immediacy to markets), simplifying exchange, devising pricing and contract information to induce parties to reveal demand and cost information, creating commercial routines (e.g., for getting credit), and making credible commitments backed by a firm's longer life and greater volume of transactions. In Spulber's (1999) conception, managers are important players, and mundane marketing and management activities (such as holding inventory) are essential to bring markets into being—even markets for well-understood products.

22.2.1 The Role of Sales and Distribution Activity

In short, marketing activities (product design, branding, promotion, pricing, sales and distribution) and management activities (product manufacture, inventory, and the like) create markets. One of the most important means of creating a market is simply to offer sales and distribution to it in a credible manner. A major justification for vertical restraints (of which the franchising system is the ultimate expression) is precisely to insure that sales and distribution activities represent and support a brand properly. This alone is thought to open new markets, by overcoming objections to purchase. One justification of selective distribution is that resellers, in return for protection from intra-brand competition, will exert themselves more vigorously on behalf of the brand (Williamson, 1979). For example, they may open outlets in disadvantaged neighborhoods, unquestionably creating new markets.

Chandler (1977) takes this argument to the level of American industry, tracing how large-scale manufacturers in the nineteenth century developed aggressive new methods of marketing their goods, even reaching into isolated rural areas. Many of these distribution methods involved granting territorial protection to dealers in return for brand support and an unusual degree of cooperation with the manufacturer (e.g., McCormick harvesters). Today, we recognize these innovations in distribution as a form of relational governance. It arises when dealers and suppliers make specific investments in each other: dealers need protection against supplier opportunism (e.g., capricious pricing or contrived termination), and suppliers need assurance that dealers will not degrade the brand name or shirk. At the limit of asset specificity, vertical integration into distribution is appropriate (e.g., Singer sewing machines, originally sold in company-owned stores).

In short, a firm's activities, and particularly its marketing efforts, create new markets, even for familiar products. They do so principally by reducing transaction costs (incurred when using the price mechanism). These capabilities are even more useful when the product itself is new. We now turn to new product development (NPD).

22.3 How Firms Develop New Products

22.3.1 The Firm's Returns to New Products

As noted earlier, business executives believe that NPD is critical. Indeed, aggregate evidence suggests that profitable firms do innovate (Capon et al., 1990). However, this does not mean that innovative firms are profitable: as noted, other firms often reap the benefits of NPD. How could the developing firm profit from its own activities? Bayus, Erickson, and Jacobson (2003) lay out four mechanisms. First, new products have new features that could attract buyers. However, this revenue advantage is frequently offset by additional costs of supporting new products. Second, new products might enable the firm to find new markets (or segments) that are price inelastic. Both these arguments involve creating new markets. Third, the firm may be able to move its existing customers to new products that are cheaper to support (e.g., might require less after-sales service). Fourth, the firm might develop new capabilities that are difficult to imitate. Without new products, firms eventually fail to adapt to changing market needs and find themselves unable to meet mounting competitive pressures, an argument congenial to evolutionary reasoning. New product launches galvanize firms, giving them impetus to renew and recombine their competences. In this view, new products are a firm's way to overcome inertia, a force to develop new routines and retire old ones. Viewed this way, the gauge of NPD success is not so much whether new products create new markets as it is how much the firm changes.

Nonetheless, firms do develop new products for the primary purpose of capturing some part of the value they add to society by creating new markets (Schmookler, 1966). How and why does such innovation occur? Classic economic analysis says little about this issue, tending instead to take innovation by the firm as exogenous. Teece (1996) examines why this is so, arguing that innovation has seven properties that defy the conventions of economic analysis (see Table 22.1).

We focus on the governance issues that beset NPD. Below, we examine three ways the firm tries to develop innovations, beginning with vertical integration of new product development activities.

22.3.2 Processes of Internal Development

A very substantial body of descriptive research looks into the "black box" of new product development. Krishnan and Ulrich (2001) review this sprawling literature, seeking to unify it inductively around the literally dozens of decisions firms make to develop and commercialize a new product or service. In the picture that emerges, firms look ahead in a deliberate, goal-directed manner, seeking to develop new products with a minimum of waste and a maximum of market impact. Economic exigencies impose discipline on the process, focusing management on the pursuit of internal process outcomes (such as fast development and rollout) and external market outcomes (such

Table 22.1 Seven properties of innovation that contradict standard economic analysis. (Adapted from Teece, 1996)

Seven fundamental characteristics of technological development make it difficult to explain innovation using the standard economic lens. The lens of NIE is more appropriate. These characteristics are:

1. Uncertainty: innovation is a quest into the unknown. Therefore, serendipity and luck play an important role

2. Path dependence: technology often evolves in path-dependent ways, contoured by a technological paradigm. Within a paradigm, research efforts become channeled along certain trajectories. New product and process developments for a particular organization are likely to lie in the technological neighborhood of previous successes

3. Cumulative nature: technology development, particularly inside a particular paradigm, proceeds cumulatively along the path defined by the paradigm

4. Irreversibilities: technology progress exhibits strong irreversibilities. The evolution of technologies along certain trajectories eliminates the possibility of competition from older technologies, even if their relative prices change significantly

5. Technological interrelatedness: it is infeasible to separate out a technology and specialize exclusively in it. Seemingly unrelated technologies share underlying points of commonality

6. Tacitness: knowledge is difficult to codify, archive, and transmit

7. Inappropriability: firms face serious hurdles to insuring they, and not a rival firm or a customer, will receive a fair share of the value the firm creates via innovation.

as market share, revenue growth, and profitability) rather than on organizational politics or pursuit of private subgoals. In this respect, the picture looks more like economics than sociology or psychology: as Dosi (1997) puts it, abandoning perfect rationality does *not* imply “anything goes.”²

However, complexity, information impactedness, and bounded rationality massively influence how the firm goes about the task. The myriad decisions that must be made (even if “decided” by the default path of inaction) exhibit important interdependencies, not the least of which is that each development project must fit the firm’s competences and strategy: the identity of the players matters. Each project is part of a constellation of other projects, which need to be coordinated. Further, each project team must coordinate across the various functional areas of the organization (R&D, production, marketing, and so forth). Coordination involves friction, which imposes transaction costs. A major finding of this literature is that the transaction costs entailed in new product development *even under unified governance* are high. (This does not contradict TCE, which suggests that firms cannot eliminate transaction costs but can contain them better than do markets.)

²Dosi (1997) reviews a growing body of economic work outside the conventions of rational, identical agents seeking an equilibrium outcome. He argues that an evolutionary lens is particularly well suited to explain technological change. Dosi (1997) concludes that linear models of innovation do not fit the evidence. A more accurate picture is one of feedback loops between innovation, diffusion, and generation of new opportunities.

22.3.2.1 The Bounded Rationality of Potential Buyers

Why are transaction costs high? Developing new products demands a great deal of information, which is not readily available and not equally available to everyone involved in the development effort (information impactedness). Much of the information is tacit: Van den Bulte and Moenaert (1998) demonstrate that sheer physical proximity has much to do with whether tacit information will be shared effectively. Krishnan and Ulrich (2001) conclude that developers must combine many types of information from multiple sources, sifting, weighing, interpreting, and combining in a complex and inherently subjective manner. Combining information strains the cognitive abilities of decision-makers (bounded rationality at individual and group levels). Numerous market research techniques have been developed to assist here (for a review, see Kaul & Rao, 1995), but their usefulness has limits, particularly when developing truly new products and services. One reason is that subjective, holistic aspects of the product (e.g., aesthetic appeal) are difficult to represent and examine, yet are important determinants of success. More fundamentally, the prospect's first response to an innovation is "What is it?" Not surprisingly, potential buyers are usually unable to imagine and report accurately what utility they might derive from an innovation, nor how likely they are to buy it.

A classic example (Nayak & Kettingham, 1986) is the videocassette player: market research undertaken by Sony led management to assume that buyers would use it primarily to record short television programs to play back later. Accordingly, Sony developed a proprietary machine (the Betamax Standard) to play a one-hour tape. Rival developer JVC made no assumptions about how people would use the machine and therefore maximized flexibility (which meant longer tapes). As the market developed, the principal utility that emerged was in playing prerecorded movies. JVC won the subsequent standards race based largely on what turned out to be the superior utility of a long tape. To do so, JVC licensed its VHS standard to competitors, thereby giving away profits from its invention. Sony attempted to do the same with Betamax but could not enlist many other firms—because their managers realized that a short tape would limit the creation of a new market. This is a good example of a firm's efforts to be farsighted: a sociological analysis would have awarded more allies to Sony than to lesser-known JVC.

22.3.2.2 Disparity of Goals

Complicating the development task is disparity of goals. Krishnan and Ulrich (2001) note that the performance of a product development project team is typically indexed by time taken (lead time to market) or by the manufacturing cost, quality, or market attractiveness of the team's output. These goals involve trade-offs (e.g., short lead times frequently drive up manufacturing costs). Further, it is not clear how well such internal project development goals map onto external performance goals, such as revenue growth or incremental profit. For example, a review of research on organizational processes (Brown & Eisenhardt, 1995) concludes that cross-functional communication improves project development outcomes (e.g., speed, productivity). However, Henard and Szymanski (2001), in a review of over 40 empirical studies, find that cross-functional integration has little impact on the next step of market or

firm performance outcomes (e.g., market share, profit margin). Apparently, the advantage of superior project development outcomes can be readily offset by inferior commercialization, a theme we develop in Sect. 22.4 of this review.

22.3.2.3 Development Routines

Firms develop routines to cope with such challenges, routines that have an important influence on the firm's competences (Nelson, 1995; Dosi, 1997). A common routine ("sequential development") is to split development efforts into phases, assign each phase to a functional group, and set deadlines for each group to complete its input and hand the project off to the next group (e.g., R&D might hand off to Production, which hands off to Marketing). Sequential development addresses the problem of bounded rationality by separating tasks and responsibilities, confining them to groups with similar viewpoints. Sequential development reduces internal friction, but at the price of longer lead times and, more importantly, new products that are poorly suited to fast-changing environments. The voice of the potential customer is usually lost as the project moves through internal groups, awaiting its turn for attention in each group. The solution is to overlap the stages ("concurrent development"). Concurrent development requires managing the information flow so as to insure that each stage not only has more information *but understands the information in the same way*. This raises transaction costs but gives firms the flexibility they need to make major changes quickly to adapt to turbulent environments (Krishnan & Ulrich, 2001).

The sequential development issue underscores four points that fit well in the NIE paradigm. First, routines are critical and are costly to alter. Second, bounded rationality means that players will not use information in the same way, even in the unlikely event that all information is available to all players. Third, the uncertainty inherent in a firm's environment has an important impact on the appropriate way of organizing its activities. Fourth, unified governance does not banish transaction costs. Therefore, how a firm structures its internal processes is of importance: firms should not be treated as black boxes nor as interchangeable production functions.

22.3.2.4 Internal Team Processes

Brown and Eisenhardt (1995) review a large body of microlevel research examining the workings of project teams. A central challenge here is overcoming the information impactedness that is endemic to innovation development processes. A major obstacle is that functional groups operate in their own "thought worlds," which condition how they obtain, share, and process information. *Even though they work for the same organization*, individuals do not readily cooperate, share information, or reach agreement, especially across functions. Conflicts and information hoarding abound. Traditional internal incentives alone are not sufficiently high-powered to bring forth best efforts from all employees (Teece, 1996). Thus, group composition matters because internal cohesion creates better project development outcomes. Moorman (1995) extends this finding from project outcomes to market outcomes: new products are better received 12 months after launch when developed in firms

with a strong culture of interpersonal support. In these firms, employees feel a personal loyalty to (and trust in) their employer. Such “clan” structures work by supplementing traditional employment incentives, motivating people to collect, pool, and utilize information candidly. Clan mechanisms, which TCE labels “the economics of atmosphere,” are important to the understanding of economic organization: however, they are difficult to unravel (Williamson, 1996, p. 270). So doing is a promising research direction in the study of in-house NPD.

The speed and productivity of the project development team increase in teams that rely on “gatekeepers,” individuals who go to great lengths to scan for information, bring it into the team, and make sure it is dispersed (Brown & Eisenhardt, 1995). Three findings are striking. First, effective teams engage in external communication, bringing fresh information and viewpoints into the organizations. Information from suppliers is particularly useful: it appears to substantially improve the development team’s results. Second, attempting to reduce friction by burying conflict is counterproductive: better performance comes when team members, who inevitably see the situation differently, discuss freely. Such teams negotiate their way to a solution that mitigates bounded rationality at the individual level: internal cohesion facilitates by keeping the conflict manageable and task focused. Third, the classic approach is for the team to develop one best new project, for which it then garners organizational support. This minimizes transaction costs and is effective in stable, relatively mature environments. But in rapidly changing environments, planning one’s way to a single best project outcome is a counterproductive exercise because such environments strain rationality. Project development teams perform better by engaging in seemingly wasteful trial-and-error processes, such as experimenting, moving quickly without thorough analysis, and testing competing designs simultaneously. Trial and error, in turn, requires a risk-tolerant, improvisational management style that goes against the routines embedded in many firms.

22.3.2.5 Purposefully Designing Creative Products with High Market Appeal

How can teams set about purposefully to design something highly creative that will attract buyers, thereby creating new markets? The evidence is that they often cannot, and frequently, they do not. Many successful innovations are not the deliberate output of a creative process. They are by-products, often accidental, of efforts to create a variation of something known (e.g., the weak, unconventional glue on Post-It Notes was a failed by-product of 3M’s efforts to make conventional glue stronger). Efforts to distinguish what purposeful processes do lead to successful innovative products have not found a dominant answer. However, some generalizations emerge. For example, at a very microlevel, Dahl, Chattopadhyay, and Gorn (1999) uncover thought processes that help designers imagine products that meet the dual (and somewhat conflicting) criteria of being creative, yet acceptable to buyers. Urban and Von Hippel (1988) show that, in business-to-business markets, certain customer profiles correspond to “lead users” who are excellent forecasters of overall market reaction. Enlisting such customers to cooperate with designers early in the development process raises the probability that an innovation will find buyers after launch.

Brown and Eisenhardt (1995) extend the idea: in many markets, project teams that seek information from potential customers tend to develop products that are better accepted in the marketplace. This is particularly the case in uncertain (complex, fast-changing) markets, where, by nature, it is difficult to understand customers. In uncertain markets, an overriding customer orientation improves new product performance by guiding managers to devote resources to decoding customers' needs and options, in spite of the difficulty in so doing (Gatignon & Xuereb, 1997). The difficulty of this challenge is underscored by Moorman (1995), who finds that in turbulent markets, firms tend to introduce new products that are less creative than those in more stable markets. A likely explanation is that such markets strain the decision-makers' ability to discern customer reaction, making it more risky to introduce something novel. Hence, more conventional new products can be seen as a means of coping with bounded rationality in the face of market uncertainty.

22.3.2.6 When Is Experience an Asset?

Firms invest in a market, acquiring experience, much of which is idiosyncratic to the firm. Experience should serve useful purposes for some time, which would make it an asset. Indeed, there is a tendency in TCE to equate idiosyncratic investments and idiosyncratic assets, as though all investments generate durable utility. But an investment is a commitment of resources. This does not always create an asset. Indeed, it could create a lasting disadvantage.

Is a firm's market experience truly an asset in developing new products, particularly innovations? In general, the answer is yes (Moorman & Miner, 1997): organizations that gather experience and are capable of dispersing it to multiple decision-makers create new products that are more successful (though not more creative). However, there are important qualifications here. Experience can be a drawback, particularly in the context of innovations. Experienced, successful firms often fail to recognize shifts in technologies or markets because they are biased toward existing markets (Krishnan & Ulrich, 2001). This focus leads them to overlook or distort information that threatens the status quo.³ Moorman and Miner (1997) show that, in experienced firms in which project team members have a strong consensus about what they know, rapid changes in the technological environment overwhelm team members. Rather than update, they develop *less* creative products than do competitors that have less experience and consensus. Brown and Eisenhardt (1995) show that project teams can become too cohesive by being together too long. After 5 years of working together, team members lose speed and productivity, apparently because they are less likely to cultivate and use external information.

³Two examples, once considered plausible, are ludicrous in hindsight. "It is an ideal dream to imagine that auto trucks and automobiles will take the place of railways in the long-distance movement of freight and passengers," proclaimed a railway trade association in 1913. And in 1977, the president of Digital Equipment Corporation, Ken Olsen, opined that "There is no reason for any individual to have a computer in their home." DEC's product line rested on the mainframe computer: the company ultimately was put out of business by personal computers. See Cerf and Navasky (1984).

Net, it appears that not all idiosyncratic investments that build up over time can really be considered assets. Experience, in particular, can become a liability, or a “core rigidity” (Leonard-Barton, 1992) because the experience is *too* specific. The routines acquired by experience often do not generalize to other circumstances. By attempting to preserve and apply overly specific knowledge, the firm creates path dependence, which in turn makes it difficult to be flexible and creative. Experience prevents the firm from moving to a new, superior path.

If so, a transaction-specific investment is one thing, and a transaction-specific asset is another. An asset generates value, and an idiosyncratic asset does so in a manner uniquely suited to a usage or user. In changing environments, idiosyncratic investments may become idiosyncratic liabilities. Afuah (2001) demonstrates this effect in the computing industry. A massive, competence-destroying technological change (the move to reduced instruction set computing (RISC) technology) threatened all competitors—but not equally. Those whose performance suffered the most were vertically integrated into the old technology and used suppliers to learn the new technology. Via vertical integration, these firms had maximized idiosyncratic assets that were rendered useless, even a liability, by the advent of RISC (which required a great deal of unlearning and then relearning). Further, the firms had no routines for dealing in a relational manner with suppliers. They were unable to transition effectively using arm’s-length market contracting because the newness of RISC required developing new idiosyncratic assets, which in turn demand safeguarding by vertical integration or by relational governance. In contrast, the best performers were not vertically integrated in the old technology but did vertically integrate to acquire RISC capabilities. Thus, they could more easily shed their old capabilities and could more readily develop idiosyncratic new ones.

Afuah (2001) is a first step in a valuable research direction: when is a transaction-specific investment truly an asset, rather than merely being idiosyncratic per se—or worse yet, a liability? TCE has tended to take the decision to create an idiosyncratic asset as given, or exogenous. But TCE has developed to the point that it is useful to endogenize asset specificity, that is, to explain the decision to invest in specific assets. When should firms undertake the transaction costs of customization? When they settle instead for generic investments—or walk away from a transaction altogether? Testing predictions about walking away is an empirical challenge: although TCE predicts circumstances that will not support transactions, it is difficult to find traces of transactions that were considered without being executed.

22.3.3 Development Relying on Third Parties

The last two decades have seen a strong trend toward outsourcing even sensitive functions once considered too central to confide to outsiders, such as new product development. A significant percentage of the sample of firms surveyed in Robertson and Gatignon (1998) reports using partners to develop new products. Shared development invokes the TCE prediction that outsourcing innovating activity creates high exposure to small-numbers bargaining, hence opportunism. The

contractual hazards that arise challenge firms to refine ways to safeguard these transactions, including relational governance, or alliances (see Gulati & Singh, 1998 for a review). By TCE reasoning, *the move to outsource NPD will fail unless firms dramatically increase their capacity to ally by such means as the exchange of credible commitments.*

22.3.3.1 Credible Commitments

The automobile industry is a leading example. Automakers are in the forefront of efforts to delegate new product development upstream. Contractual hazards arise because suppliers are expected to invent components that work with each particular brand and model, which are highly idiosyncratic. Thus, suppliers must invest in idiosyncratic learning to develop highly specific components. This puts suppliers squarely into a position of small-numbers bargaining, exposing them to the buyer's opportunism. TCE predicts that suppliers will hesitate and will demand concrete assurances in order to proceed: promises of good faith will not suffice. The single best assurance is when a buyer creates a reciprocal exposure by itself making investments that are idiosyncratic *to the supplier*. In so doing, the buyer creates a vulnerability to a supplier: mutual vulnerability creates incentives to eschew opportunism, thereby protecting the relationship.

Bensaou and Anderson (1999) examine when buyers take the risk of investing idiosyncratically in a supplier. They find firms trade off between production costs motives to invest and transaction cost disincentives to invest. On the production side, the more technically challenging the necessary invention, the more buyers make idiosyncratic investments in their supplier. Buyers do so because they cannot simply put out a contract to motivate the supplier to make such investments, but must offer safeguards. Offering safeguards to the supplier is particularly necessary under fast technological change: buyers must absorb some of the risk for their suppliers if they are to obtain the close cooperation needed. Notably, the institutional environment matters: Japanese automakers are able to induce considerably more cooperation from their suppliers for a given level of investment, consistent with findings by Sako and Helper (1998).

More generally, Oxley (1999) compares governance choices made by US firms in their collaborations with firms from 110 countries, collaborations intended to develop and commercialize intellectual property. She finds US firms to be sensitive to both the transaction and the environment. Specifically, firms structure their relationships to approximate hierarchies (by using equity joint ventures vs. market contracts) when idiosyncratic assets need safeguarding, when performance is difficult to monitor by observing outputs, and when countries provide weak protection of intellectual property.

TCE argues that idiosyncratic investments serve to generate better performance. In support, Jap (1999) shows that close buyer-supplier relationships built on idiosyncratic investments outperform rivals over time, generating significant competitive advantage. Dyer (1995) finds that automakers which build tightly integrated supplier networks based on high levels of mutual asset specificity reap performance benefits: their new products enjoy higher quality and take less time to develop. These

studies affirm TCE reasoning that offering credible commitments is necessary to forestall opportunism. However, other solutions to the specificity problem exist: for example, Dutta and John (1995) show that buyers are more willing to make supplier-specific investments if the supplier licenses the innovation. Thus, buyers know they can use a second source to safeguard against opportunism.

22.3.3.2 How the Downstream Protects Itself Against Upstream Opportunism

Firms also go downstream to develop innovations and find many ideas among members of their channel of distribution. These are not end customers: channel members move the product or service along the path to the end user. Resellers and sales agencies are prime channel members for purposes of developing new products. Typically, they sell complements, and frequently substitutes, giving them a broad perspective on the market. As noted earlier, project teams perform better when they bring this expertise into their deliberations. Some producers develop close relationships with channel members in order to secure their cooperation in the development effort. Anderson and Weitz (1992) show that they do so by exchanging mutual idiosyncratic investments and investing heavily in communication over time. However, relationship building exhibits considerable path dependence: reseller-producer dyads find it difficult to set aside a contentious history in their efforts to achieve coordination. Sako and Helper (1998) reach the same conclusion upstream, for customer-supplier dyads.

An emerging issue is how franchising systems (for a review, see Coughlan et al., 2001) generate new product ideas. Franchising is a form of relational governance, a hybrid balancing properties of both markets and hierarchies. The institution of franchising is particularly well suited to generating, refining, then transmitting, and commercializing new product ideas. Darr, Argote, and Epple (1995) analyze how they do so, showing that new ideas (e.g., process innovations) spread by interpersonal communication within a social system, facilitated by geographical proximity. Thus, new ideas tend not to spread beyond the inventive franchisee and his/her immediate circle. The franchisor is an institution that collects, culls, refines, and then spreads innovations across franchisees. Schmookler (1966) notes that much of the energy that goes into invention is really diverted into reproducing inventions that have “depreciated” (been forgotten). Darr, Argote, and Epple (1995) show this effect is particularly powerful when personnel turnover is high and production processes are simple: the relevant knowledge is difficult to embed in anything other than an individual’s mind. Getting individuals to share information with each other is therefore critical. Bradach (1997) analyzes how some franchisors do this. He focuses on “plural governance,” which is frequent, structured interaction between company-owned and independently owned outlets. Plural governance facilitates remembering and transmitting innovations. This is a new rationale for a poorly understood phenomenon, the simultaneous use of market *and* hierarchy (dual distribution). Explaining dual distribution, a seemingly wasteful duplication of resources, deserves research priority.

22.3.3.3 Observations on the Hazards of Cooperation

It is highly risky to engage in mutual building of idiosyncratic investments. Giving and taking hostages (exchanging credible commitments) is a difficult affair to calibrate and to execute in practice (Williamson, 1996; Wathne & Heide, 2000). Teece (1992) uses TCE reasoning to build the argument that vertical integration is appropriate for assets that are “co-specialized,” i.e., both complementary and specialized to the innovation. Teece argues that hazards are pronounced with human assets: integration is vital to impede people from leaking knowledge or switching to a competitor. Indeed, Brown and Eisenhardt (1995) underscore that project team gatekeepers enhance performance by performing a “guard” function to protect proprietary information.

There has been a trend (reviewed by Osborn & Hageboorn, 1997) toward increasing cooperation among producers, particularly producers of potentially complementary goods or services. Cooperation among producers is a new way to outsource the development of innovation. One purpose of these programs is what may broadly be termed co-promotion or co-marketing of each other’s brands, including new products. A promising research area is how firms manage the transactional hazards of cooperation. For example, Terpstra and Yu (1990) examine carrier-rider relationships (“piggybacking”), in which firm A (the carrier) uses its sales force to promote the (complementary) new products of firm B (the rider). The obvious danger is that A will behave opportunistically to appropriate B’s innovation, (reverse engineering, misleading B about the product’s market position, and so forth). How do firms govern such relationships, and which methods are more effective? For example, there is some evidence that, in the pharmaceutical industry, firms exchange hostages: piggybacking is reciprocal, with A and B exchanging carrier-rider roles in multiple markets. Such issues deserve research.

22.3.4 Develop by Acquisition or Appropriation

Given the difficulties of successfully developing innovations, including new products, it is not surprising that many producers don’t bother to do so. Instead, they watch the activities of other firms, waiting for a “winner” to emerge, which they then acquire—or appropriate. Kogut and Kulatilaka (2001) point out that where firms can merely observe, they will do so. Frequently, mere observation is inadequate: the firm must have some involvement so as to discern the emerging trends and be able to adapt readily. Many seemingly wasteful or indecisive actions, such as investing in multiple (potentially competing) channels of distribution or multiple technologies, or spending inordinately on sales forces or market research, can be interpreted as investments in “real options” that enable the firm to act as soon as uncertainty falls.

Teece (1996) questions the practicality of buying technology elsewhere (e.g., licensing another firm’s innovation). Teece concludes that, in spite of the myriad obstacles to developing new products internally, vertical integration is usually the most attractive of the feasible alternatives. Our earlier discussion of the transactions costs inherent in new product development supports this approach by suggesting that

complex contingent claims contracting, based on a high level of rationality and information, is simply not practicable. However, management practice indicates that many firms do effectively outsource innovation.

Regarding truly innovative new products (not mere extensions of the firm's existing products), a common strategy is to merely purchase rights to innovations developed elsewhere. For example, as developing new pharmaceuticals has become slower, more expensive, and more prone to failure, more producers let others invest and then buy licenses to market the results (Tapon, 1989). While promising, this approach is fraught with transaction cost perils. To evaluate an innovation developed elsewhere, the firm needs more information than the seller would be wise to reveal. Even if the firm manages to buy the innovation, it may find that idiosyncratic assets are difficult to uproot and relocate, especially if they involve tacit knowledge or go counter to the firm's established routines.

An example is the biotechnology industry. Start-up firms regularly emerge with innovations but lack the co-specialized assets (such as sales forces) to commercialize them. Therefore, large, established players, facing internal barriers to developing innovation themselves, often invest in purchasing start-ups, rather than in R&D.⁴ However, a considerable body of research on acquisition (for reviews, see Capron, 1999; Haspeslagh & Jemison, 1991) concludes that many of the same obstacles to licensing examined by Teece (1996) persist, notably difficulties in valuing the acquisition and in meshing the acquiring and acquired firm. Thus, Pisano (1990) finds, even for biotechnology innovations, firms integrate R&D to mitigate hazards of opportunism. Argyres and Liebeskind (2002) examine the trade-off between acquisition and integration, incorporating path dependence in their analysis.

A promising research direction here is to connect models of what firms do (which is the bulk of this literature) to the performance outcomes they achieve. Such an approach fits squarely within the NIE paradigm, in which it is accepted that, although there is a selection mechanism, uncertainty, bounded rationality, heterogeneous agents, and path dependence conspire to permit some firms to "get it wrong," yet survive alongside firms that "get it right" (Dosi, 1997).

22.4 How Firms Create New Markets by Commercializing New Products

To this point, we have focused on how firms attempt to create new markets by inventing novel products. As noted earlier, developing a promising concept is one step, creating a market is another step, and deriving profits from the exercise is still another step. Here, we focus on step two: how do prospects become buyers, thereby

⁴This is an interesting twist on Schumpeter (1943), who feared that large firms would eventually stifle the entrepreneurial spirit necessary to drive successful internal R&D. Schumpeter did not foresee the solution of buying the entrepreneur's firm, in part because today's institution of venture capital to fund entrepreneurs was not well developed in the 1940s.

calling new markets into existence? Following Griliches (1957), we focus on the adoption (i.e., *first-time purchase*) of the innovation. Markets are born when prospects adopt. But markets grow and become established only when adopters make repeat purchases (replacements, upgrades, multiple units, gifts). As a general rule, if a new product is widely adopted, it has sufficient utility that repeat purchases will occur. Therefore, the creation of new markets rests heavily on the diffusion of the innovation, i.e., the spread of first-time purchases among members of a target population.

Typically, for successful innovations, diffusion starts slowly, then accelerates, and then tapers off to a saturation level that is less than 100% of all the potential buying units in a population. Such a process produces a capped S-shaped curve when cumulative first-time sales are plotted over time. Griliches (1957) pioneered the study of such curves: he examined farmers' adoption of hybrid corn, noting that the same innovation diffused in a different way among different groups of seemingly identical farmers.

Why does one innovation penetrate the market more thoroughly (closer to 100% adoption) than another? Why do some innovations reach their potential more quickly than others? These are central questions. Subsequent to Griliches (1957), a huge literature on diffusion theory has developed in parallel in the fields of psychology, sociology, economics, and marketing. Most of it rests on the S curve. Zvi Griliches offered this observation:

The basic notion was that here was a technical change that was going on, and it was not just happening entirely out of the air, but it was being affected by economics, and to some extent, being created by the economic situation, as well as being affected by it. . . . The funny thing is that diffusion, as such, has never taken off on a large scale in economics. It's a major topic for people in marketing, and they do that. But not very much in mainstream economics, partly because of the way, possibly, I formulated the problem—there is a fundamental disequilibrium. . . . knowledge matters, and the spread of knowledge matters. It's interesting to model how the knowledge spreads. (Krueger & Taylor, 2000, p. 180)

Nerlove (2001) also concludes that Griliches' ideas about diffusion have had their greatest influence outside mainstream economics. In this section, we review briefly the diffusion theory and modeling literatures. We first focus on why some innovations diffuse successfully, a necessary condition for new markets to come into being. Then we discuss the strategic commercialization decisions taken at the time of the introduction of the new product or service, focusing on how these decisions help to create (or foreclose) new markets.

22.4.1 Diffusion Theory and Models

New products are outside a buyer's routines. Radical innovations go further: they are outside a buyer's cognitive space; their attributes are difficult to understand, let alone to value. Therefore, innovations arouse a buyer's sense of risk. If they displace the buyer's current routines (e.g., new software replacing software the buyer has

mastered), innovations also arouse resistance. And innovations can create suspicion. If I become locked in to this brand, will the supplier subsequently exploit my dependence, say by raising price or by failing to deliver on service promises? Uncertainty, fear of small-numbers bargaining, disruption of routines—this scenario fits comfortably in evolutionary economics and in TCE.

22.4.1.1 Cognitive and Social Processes of Adoption of Innovations

Diffusion theory has focused on fundamental cognitive and social processes as an explanation of adoption decisions by individual consumers or by organizational customers (Rogers, 1983). Until recently, this process has typically ignored the marketing of the innovation by the firm commercializing the product or service and the competitive forces at play (Gatignon & Robertson, 1989, 1991; Robertson & Gatignon, 1986) with the exception of studies of mass communication (Katz, 1957) and interpersonal influence through word of mouth (Brown & Reingen, 1987; Herr et al., 1991). We shall return to these influences.

The basic adoption process that underlies most diffusion models considers that individual buyers (who are potential users of the new product or service) can receive information about this new product or service from two sources: inside and outside the “social system” (i.e., the set of individuals whom they can observe and with whom they interact). Sources outside the social system include mass media (buyers can be exposed to these without mediation by the influence of other people). Sources within the social system are a function of the number of prior adopters of the innovation (creating network externalities). Diffusion theory designates two categories of adopters. *Innovators* are those who adopt early because they learn about the innovation from sources independent of prior adopters (as these do not exist early on in the product life cycle). Innovators have a profile (see Rogers, 1983; Gatignon & Robertson, 1985) that suggests Veblen effects are powerful: innovators seek social status by being among the first to own something new and often pay high prices to do so. Fortunately for the firm, innovators like to be seen with their acquisition. Their display spurs *imitators*, who tend to be more risk averse and/or more bound to their habits. Imitators adopt later than the innovators, after they gather information (proactively or passively) from those buyers who have already experienced the innovation (have already adopted). The innovator-imitator process implies that, rather than making broad efforts to gain acceptance, firms need to target potential innovators (e.g., by mass media) and win them over first, so that the imitation process can get underway and gather momentum. Diffusion models represent this dynamic in an as-if fashion, for both consumers (individuals) and businesses.

Typically, it is in the firm’s interest to spur diffusion as much as possible. However, markets are differentiated institutions. Not all of them depend on widespread adoption. Some rest on a very selective approach, in which exclusivity is considered a positive and widespread diffusion is viewed as a negative. Luxury items, such as certain cars, jewelry, and clothing are examples. Their makers search for terms to differentiate them from their widely diffused mass-market counterparts. For example, “*haute couture*,” “designer clothes,” and “ready-to-wear” are terms used to designate three levels of exclusivity in the clothing category, which generates

a wave of new products every season. Makers of *haute couture* seek media mention that they design clothes for specific individuals and occasions (e.g., a named actress at the Academy Awards ceremony). Their customers don't want to see other people wearing a similar outfit: diffusion is a negative and exclusivity is a positive. In contrast, makers of designer clothes seek somewhat wider diffusion: they advertise that their wares are derived from *haute couture* and are sold only in selected stores. Seeking maximal diffusion, makers of ready-to-wear advertise that their clothes are readily available and are appropriate for almost anyone on any occasion.

22.4.1.2 Some Generalizations from Empirical Models

Bass (1969) proposed a robust, generalizable, readily estimable model that integrates both a propensity to innovate (through a coefficient of external influence) and a propensity to imitate (through a coefficient of internal influence). Based on the Bass model, a large number of empirical studies have been conducted over the last 30 years, permitting us to reach generalizable conclusions regarding diffusion research (Mahajan et al., 1993, 2000; Sultan et al., 1990). Chief among these is that imitation forces are much more powerful than innovation forces. In other words, a firm's efforts amount to little until there is a solid base of visible users: factors such as word of mouth matter more than the inherent innovative tendencies of buyers. Further, as noted by Griliches (1957), there are many ways to draw an S curve: variation in diffusion patterns is considerable, even for the same innovation in different populations. Finally, diffusion models are robust: they provide a good overall representation of a very large range of situations, including both process and product innovations, and organizations as well as individuals, across a broad range of environments.

A major contribution of the marketing field to diffusion modeling concerns the incorporation of marketing activities into diffusion models. Based on these observed generalizations of the patterns of diffusion, it is clear that *what firms do (or fail to do) matters*. The addition of factors controlled by the firm that is marketing the innovation enables management to influence the adoption and diffusion by consumers. This issue is developed in a later discussion of marketing decisions.

22.4.1.3 International Diffusion

One particularly recent development in this stream of research concerns the role of multi-market marketing, especially the international introduction of new products and services in multiple countries. It is rare for a firm to enter multiple markets simultaneously. Typically, firms follow a sequence, conquering first one market and then another. To some extent, this is explained by resource constraints.

One of the major ways management affects the creation of new markets is in deciding which countries to enter. The question of knowing which countries to select for entry to commercialize a new product or service, and in which order to enter target countries, has received much attention in international business. However, the focus has been in identifying segments of countries sharing similarities. The idea is to find submarkets that make some groups of countries more attractive and less risky to enter than others. This segmentation by clustering of countries is typically based

on macro-political, demographic, geographic, or economic variables (Sethi, 1971) and is usually rather atheoretical.

Gatignon, Eliashberg, and Robertson (1989) brought a new theoretical focus to this issue by comparing countries in terms of their diffusion parameters and searching for explanations of similar parameters. They explain similarity and differences in patterns of diffusion by invoking sociological explanations concerning the cosmopolitanism of the culture in a country, the mobility of its population, and the role played by working women. Helsen, Jedidi, and DeSarbo (1993) develop clusters of countries based on these similarities.

This segmentation, however, ignores the fact that some countries adopt earlier than others. In fact, the decision to enter one country is not independent of knowing if and when entries in other countries will occur. Usually, the innovation is marketed first in the country that is the home of the firm that developed the innovation (the lead country). Then, other countries are attacked with a lag of different durations. Understanding the reasons for these lags is important: see Dekimpe, Parker, and Sarvary (2000), who investigate the reasons for initial sales and for the diffusion speed across a large group of countries.

Another aspect of the multinational diffusion of an innovation concerns cross-country effects, that is, the role that diffusion in one country plays in determining diffusion in another one. Putsis et al. (1997) show that it is possible to analyze the leading role of some countries in explaining the speed of diffusion in other countries. For example, they find that within European countries, Germany, France, Italy, and Spain demonstrate a high level of influence on the diffusion of several innovations in the other countries. These findings suggest which countries should be entered first, as the (later) diffusion in the lag countries will require smaller investments on the part of the firm. Therefore, this research demonstrates the interrelationships that exist between countries that lead to the necessity of taking a global management approach to world markets.

In short, the creation of one market is not independent of the creation of another. Decisions taken earlier as to which countries to enter first can have a very substantial influence on how successfully the firm creates a later market. In other words, there is path dependence in the creation of new markets.

22.4.1.4 Industry Effects

Initially, the diffusion process was viewed as a monolithic centralized process focused on the innovation itself and on innate features of the adopter. A more recent research thrust focuses on how technological innovations diffuse among firms, as opposed to individuals (Robertson and Gatignon (1986). This stream considers characteristics of the *industry in which the innovation occurs* (e.g., overall competitiveness, reputation, marketing intensity), as well as characteristics of the *target industry* (e.g., demand uncertainty, professionalization). Gatignon and Robertson (1989) and Parker and Gatignon (1994) offer empirical evidence that such factors significantly impact the diffusion pattern.

Taking into account features of the industry of the innovator and the adopter follows naturally from evolutionary economics (which presumes that actors are not

homogeneous) and from TCE (which stresses differences in a firm's stock of idiosyncratic assets). This stream of research directly addresses the issue that motivated Griliches (1957): why do seemingly identical farmers react differently to the same innovation? This research stream unpacks the "seemingly identical" aspect of the target population, as well as differences among the brands competing to be the most successful prototype.

22.4.2 Marketing Entry Strategy Effects

Kuester, Gatignon, and Robertson (2000) identify five strategic actions of the firm at the time of entry that affects the speed of diffusion of an innovation: (1) the choice of the market segment(s) targeted, (2) the order of the firm's entry into the market, (3) preannouncing activities, (4) market-entry commitment, and (5) the distribution strategy. We cover these areas below. In addition, the mode of entry into foreign markets has been a favorite subject of research in the international business field and will be reviewed.

22.4.2.1 Choice of the Market Segment(s) Targeted

It follows from diffusion theory that the innovators (who adopt without expecting or awaiting information from prior adopters) should be the first group of consumers to be targeted. Characteristics of innovators have been studied extensively, and generalizations emerge (reviewed in Gatignon & Robertson, 1985). For example, innovators spread information about the innovations more than do others and tend to be more exposed to, and more receptive to, mass media communications. Innovators have also been shown to tend to be younger, richer, and less price sensitive than late adopters.

Becker (1970) argues that this profile of innovators may not be the same for all products. For innovations that are inconsistent with the norms in place in the social system, marginal individuals (i.e., nonconformists at the margin of their community) can actually have more influence on other prospective adopters. In contrast, usual groups of innovators typically include individuals who are very well socially integrated with the majority in their community.

22.4.2.2 Order of Entry

That there is an advantage to being the first entrant in what becomes a product category has been demonstrated in a significant number of empirical studies (Bowman & Gatignon, 1996; Kalyanaram & Urban, 1992; Mascarenhas, 1992; Urban et al., 1984). Gielens and Dekimpe (2001) find that order of entry is the most critical factor of those they study in analyzing the entries of retail firms in foreign markets. However, this first-mover advantage is not always maintained, depending on the later entrants and on the marketing decisions of the first entrant, especially its reactions to the subsequent entries (Bowman & Gatignon, 1995; Gatignon et al., 1997; Shankar et al., 1998).

There are many explanations for the first-mover advantage. One is switching costs: early buyers become locked in, for example, by investments in learning how to use the innovative brand. Other explanations for the first-mover advantage include becoming the reference brand for preferences in the consumers' cognitions, distributional advantage, and cost advantages through economies of scale. Regardless of the source of the advantage, first-mover benefits tend to last beyond the entry period and throughout the diffusion process in terms of market potential (saturation or cumulative penetration) and market share (Kalyanaram & Urban, 1992). However, late movers tend to exhibit a faster speed of diffusion than the pioneers, who must face resistance to trial, due to the complexity of innovations as well as the new patterns of behaviors characteristic of discontinuous innovations (Robertson, 1971).

22.4.2.3 Preannouncements

The marketing of a product may actually precede its availability in the market. This is certainly the case with the distribution system: distribution channel members face routine requests that they pre-commit to carry products and services that are said to be under development and are promised at a future date. Preannouncements (vigorous promotion of nonexistent products) has become common, even the prevailing practice in some industries (such as the software industry—vaporware—or the movie industry—the endlessly forthcoming latest film from director X or star Y). Eliashberg and Robertson (1988) discuss two reasons for the practice, one being to gain an advantage with the consumer (who will hopefully wait for the product) and the other as a preemptive move vis-à-vis the competitors (in hopes they will withdraw resources from the market). While the benefits of preannouncements are clear in theory, empirical evidence of the extent of their effects is limited. In addition, there are also clear risks associated with the firm's potential inability to bring the new product to market on time (or at all). The firms may be suspected of incompetence—or worse yet, opportunism.

22.4.2.4 Entry Commitment

The commitment of the firm to a new product in a given market is demonstrated by different actions. Commitment is important: it reassures buyers worried about the firm's potential opportunism (sell and then leave) and discourages competition. Three actions to signal commitment are covered below.

Scale Scale is a classical deterrent to entry. Earlier work has concentrated on barriers that form naturally (Bain, 1956; Scherer & Ross, 1990). The more recent focus has been on decisions that have the express intent to deter entry, especially decisions on manufacturing capacity (Spence, 1977; Dixit, 1979; Demsetz, 1982). Gielens and Dekimpe (2001) provide empirical support for the long-term impact of the scale of entry in the analysis of retail entries into foreign markets. *Product Adaptation* This factor concerns the adaptation of the product to the local environment (as opposed to product standardization—the same item in all markets). While standardization creates economies of scale, adaptation exhibits sensitivity to local market needs. The positive impact of adaptation on the long-term sales of an entry is studied by Gielens and Dekimpe (2001). While adaptation may be beneficial (as it raises demand), some standardization may also be necessary. Standardization is

particularly appropriate when externalities are involved. In this case, compatibility is a critical factor in the innovation's development of the market. As global externalities become prevalent, the global standardization of products may be inevitable. Path dependence looms here: the standard is likely to be decided in a country quite different from the one entered—and may be rather poorly suited to countries entered after the standard has been set.

Price Achieving a high level of market penetration through low prices at initial stages of the introduction of a product serves to achieve rapid diffusion, as buyers outside the innovator segment are price sensitive. Further, low initial prices show both consumers and competitors that the firm is committed to the market. The opposite strategy (initial high prices, or skimming) leads to slow, shallow diffusion and implies the firm may readily leave the market as prices drop.

Promotion A penetration strategy is also obtained through intensive advertising and communication expenditures. These fuel the rate of diffusion and enlarge the market, in part by converting early prospects (innovators) into adopters. Heavy promotion may also be interpreted as a sign of the firm's commitment to the innovation and hence may function like a hostage: prospects may reason that a firm would not knowingly invest heavily in a dubious innovation. In this respect, promotion reinforces a firm's reputation, which in turn reassures potential adopters about the (unknown) quality of the innovation (Kirmani & Rao, 2000).

Sales Force The firm's sales force is a critical factor in gaining acceptance among business buyers. In pharmaceuticals, Aitken et al. (2000) argue that the key to obtaining good licenses to new products (a crucial factor in this industry) is to have a strong sales force. McGrath (1997) notes that high-technology firms field extremely expensive sales forces. The logic is that salespeople build relationships with prospects and redeploy these relationships over successive innovations to reduce a prospect's uncertainty over the latest generation of product. Given the high rate of change in these industries, the sales force's ability to speed up adoption justifies its high cost.

Here, too, transaction costs arise. Salespeople frequently resist selling new products, particularly innovative ones. Overcoming this resistance is not easily done and requires resources and active management intervention (Anderson & Robertson, 1995). Being vertically integrated forward into sales is a major advantage in so doing. Anderson and Schmittlein (1984) compare producers that have their own sales forces to those that contract with third-party sales organizations (a classic choice between market and hierarchy).⁵ As per TCE predictions, in-house sales

⁵The difference between these two institutions—employee sales force and third-party sales force—is not always understood. In particular, independent sales forces are sometimes confused with franchisees. They are also sometimes assimilated to employee salespeople who are paid on commission. This is incorrect. The third-party organization is paid on commission, but it is a company, and therefore makes its own decisions how to pay its salespeople. Frequently, the organization pays its salespeople on salary, not commission. An independent sales force is akin to an independent advertising agency, law firm, accounting firm, or consulting firm. For each function—selling, advertising, legal advice, accounting, consulting—the firm's choice is whether to perform the function with its own employees or contract with another organization. How the individual's compensation is determined is a separate issue.

forces possessed significantly higher levels of transaction-specific assets, in more volatile environments, and operated under conditions of greater performance ambiguity. These are precisely the circumstances surrounding the introduction of new products.

A distinguishing feature of personal selling is that the salespeople are better informed about their customers than are their superiors. This information impactedness underlies why independent insurance agents thrive, in the face of myriad institutional and transaction cost factors that suggest they should not (Regan & Tennyson, 1996). Accordingly, a major role of salespeople is not to sell anything. Instead, they act as market researchers and as partners to marketers in new product development. Such non-sales behavior is essential for an innovation, which invariably has difficulty finding its markets/applications.

22.4.2.5 Distribution

The firm's *channels of distribution* have substantial influence on how well the innovation connects with buyers. Resellers and agents cultivate a reputation among their clientele and put this reputation behind what they sell. Producers effectively rent the selling firm's reputation and find it particularly valuable to overcome resistance to new products.

As noted earlier, producers may seek to build close, committed relationships with resellers and agents and exchange credible commitments to do so. One object of such relational governance is to secure quality of effort: producers want their channels to present the innovation a certain way to a targeted segment. Selective distribution is important to this effort. Fein and Anderson (1997) show that resellers and producers employ selectivity in an elaborate exchange of hostages: producers concede market exclusivity in return for category exclusivity (non-representation of competing brands), as well as other safeguards. One reason selective distribution is effective for selling innovations, particularly in final goods markets, is that representation by the "right" channels sends a quality signal to consumers (Wathne and Heide (2000).

An issue resellers and agents face is how to cope with opportunism by suppliers. For example, suppliers have an incentive to be opportunistic by telling resellers that all their new products are highly saleable, including those they know to be otherwise. Channel members need to safeguard: one way to do so is to screen out such false information. Here, Chu (1992) offers a novel interpretation of what has become an institution in grocery retailing. Slotting allowances are fixed fees producers pay to "rent" shelf space to introduce new products and are in addition to margins. Resellers can use them to oblige producers to signal which new products they truly consider to be most likely to sell. Slotting fees can be viewed as a means of negating supplier opportunism (self-interest seeking with guile) concerning its new products. This is particularly important in fast-moving consumer goods, in which literally thousands of supposedly "new" products come out each year.

More generally, channel members need to safeguard against supplier opportunism whenever they sell new products. By the time channel members discover the

product's shortcomings, they have already invested. Further, the producer may fail to live up to its promises, including the promise of a fair return to commercialization efforts. Not surprisingly, Heide and John (1988) show that agents who safeguard their idiosyncratic investments are more profitable than those who do not.

22.4.2.6 Modes of Entry

How a firm enters a market impacts how well its new products perform there. *Modes of entry* are institutional arrangements (e.g., minority joint ventures) that firms use to govern their activities when launching operations in foreign markets. There are many entry modes: these arrangements can be understood by arraying them along Williamson's (1996) market-to-hierarchy continuum. Gatignon and Anderson (1988) study over a thousand entry mode decisions and show that they tend to follow transaction cost reasoning about the trade-off between the benefits of control and the costs of risk-taking. In particular, when firms invest heavily in R&D or in advertising (and for innovations, they tend to do both), they safeguard these investments by such high-control entry modes as owning their entry vehicle outright or holding enough equity to dominate their partners.

In short, the diffusion of an innovation rests heavily on the governance structure the firm uses to enter a new market and on the marketing strategy the firm employs. These are decisions made by managers operating under bounded rationality in uncertain markets. The identity of the firm matters because new products are sufficiently unique that they cannot be readily compared to existing products. Traditional economic analysis is difficult to apply in its entirety to these situations.

22.5 Reprise

How do firms create new markets? We have learned a good deal by positive study of what firms do and how well it works, especially in the marketplaces that arise in response to market offerings. These patterns fit well in the paradigm of New Institutional Economics, particularly with TCE and evolutionary reasoning. The patterns do not fit an optimizing logic, nor an equilibrium logic, nor a logic of full rationality. Instead, they fit a comparative logic, in which only concrete alternatives are considered by boundedly rational actors, and equilibrium may never be achieved.

The processes uncovered in this research reflect several critical themes. Firms strive to follow an economic logic and exercise their capacity for conscious foresight to the fullest. But the nature of innovation strains this capacity severely. Bounded rationality, information impactedness, and the tacit nature of knowledge shape what firms do and how well their practices work. Forecasting abilities are highly limited. Calculations are problematic, valuation is error-prone, and risks and uncertainty are irreducibly high. Opportunities for self-interest seeking with guile abound: the institutional environment cannot offer enough protection of property rights to make market contracting the best approach in many circumstances. Thus, many

patterns of business practice are motivated heavily by the need to erect governance mechanisms that safeguard innovation activities and their outputs against opportunism. Opportunism cuts multiple ways. Suppliers fear being shortchanged by owners of co-specialized assets. Business partners fear being drawn into small-numbers bargaining and then exploited by their suppliers. Prospective buyers fear being misled by firms that oversell their innovations or exit the market, leaving the buyer with an obsolete or unsupported product. Short of outright vertical integration, hostages appear to be effective safeguards. Mechanisms for creating markets merit a broader analysis than any single perspective provides: in particular, a transaction cost approach needs to be informed by considerations of path dependency. Research directions here are offered by Argyres and Liebeskind (1999) and Williamson (1999).

For the most part, contracts are an unsatisfactory way to safeguard, leading to high levels of vertical integration (or failing that, posting of mutual credible commitments). *Ex post* adaptation to unfolding events is the order of the day, aided by the low-powered nature of incentives in firms: because employees are *not* rewarded as entrepreneurs, it is easier to coordinate the massive efforts needed to develop and commercialize innovations. The identity of the players matters. In particular, the internal structure, culture, and routines of organizations are critical. Path dependence also plays a substantial role because firms cannot easily acquire what they have failed to build painstakingly over time.

Patterns on the buyer side reflect that prospects are not perfectly informed about how well an innovation meets their needs. Indeed, they may not even sense their needs unless the innovation emerges as a solution. Fearful of making an error and unable to resolve their uncertainty, most prospects wait to adopt the innovation until they see others do so or know the product has been successful in another country. Hence, the firm must find, target, and convince a small set of innovative souls in a lead country. This requires intensive marketing effort and opens the possibility that a superb innovation will not create a market because the firm mishandled its introduction. Conversely, a relatively mediocre innovation may be skillfully commercialized, create a market, and even become a standard because the firm astutely priced, promoted, distributed, and targeted the innovation. Further, the diffusion of innovations exhibits path dependence. How the S curve unfolds depends on the firm's order of entry into a given market, order of entry into serial markets, and early success (or lack thereof) with the innovators.

A key theme of this work is that organizations matter, not only in the development of an innovation but in the creation of new markets. Buyers are not homogeneous in their needs. They do not simply emerge, because they do not easily appreciate the utility an innovation offers to them. How a firm takes its innovation to market has a great deal to do with whether markets come into existence and how those markets function. These patterns merit further study. The paradigm of NIE is a fruitful way to frame them.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Afuah, A. (2001). Dynamic boundaries of the firm: Are firms better off being vertically integrated in the face of a technological change? *Academy of Management Journal*, 44(6), 1211–1228.
- Aitken, M., Baskaran, S., Lamarre, E., Silber, M., & Walters, S. (2000). A license to cure. *McKinsey Quarterly*, 10(1), 80–89.
- Anderson, E., & Robertson, T. S. (1995). Inducing multi-line salespeople to adopt house brands. *Journal of Marketing*, 59, 16–31.
- Anderson, E., & Schmittlein, D. C. (1984). Integration of the sales force: An empirical examination. *RAND Journal of Economics*, 15(3), 385–395.
- Anderson, E., & Weitz, B. (1992). The use of pledges to build and sustain commitment in distribution channels. *Journal of Marketing Research*, 24, 18–34.
- Argyres, N. S., & Liebeskind, J. P. (1999). Contractual commitments, bargaining power, and governance inseparability: Incorporating history into transaction cost theory. *Academy of Marketing Review*, 24(1), 49–63.
- Argyres, N. S., & Liebeskind, J. P. (2002). Governance inseparability and the evolution of the US biotechnology industry. *Journal of Economic Behavior and Organization*, 47(2), 197–219.
- Bain, J. (1956). *Barriers to competition*. Harvard University Press.
- Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science*, 15(5), 215–227.
- Bayus, B. L., Erickson, G., & Jacobson, R. (2003). The financial rewards of new product introductions in the personal computer industry. *Management Science*, 49(2), 197–210.
- Becker, M. H. (1970). Sociometric location and innovativeness: Reformulation and extension of the diffusion model. *American Sociological Review*, 35, 267–282.
- Bensaou, B. M., & Anderson, E. (1999). Buyer-supplier relations in industrial markets: When do buyers enter the trap of making idiosyncratic investments? *Organization Science*, 10, 460–481.
- Bowman, D., & Gatignon, H. (1995). Determinants of competitor response time to a new product introduction. *Journal of Marketing Research*, 32(1), 42.
- Bowman, D., & Gatignon, H. (1996). Order of entry as a moderator of the effect of the marketing mix on market share. *Marketing Science*, 15(3), 222–242.
- Bradach, J. L. (1997). Using the plural form in the management of restaurant chains. *Administrative Science Quarterly*, 42(2), 276–303.
- Breshnahan, T. F., & Gordon, R. J. (1996). *The economics of new goods*. University of Chicago Press.
- Brown, S. L., & Eisenhardt, K. M. (1995). Product development: Past research, present findings, and future directions. *Academy of Management Review*, 20(2), 343–378.
- Brown, J., & Reingen, P. H. (1987). Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14, 350–362.
- Capon, N., Farley, J. U., & Hoenig, S. (1990). Determinants of financial performance: A meta-analysis. *Management Science*, 36(10), 1143–1159.
- Capron, L. (1999). The long term performance of horizontal acquisitions. *Strategic Management Journal*, 20(11), 987–1018.
- Cerf, C., & Navasky, V. (1984). *The experts speak*. Pantheon Books.
- Chandler, A. D. (1977). *The visible hand: The managerial revolution in American business*. Belknap Press.
- Chu, W. (1992). Demand signalling and screening in channels of distribution. *Marketing Science*, 11(3), 327–347.
- Coughlan, A. T., Anderson, E., Stern, L. W., & El-Ansary, A. I. (2001). *Marketing channels* (6th ed.). Prentice-Hall.
- Dahl, D. W., Chattopadhyay, A., & Gorn, G. J. (1999). The use of visual mental imagery in new product design. *Journal of Marketing Research*, 36(1), 18–28.
- Darr, E. D., Argote, L., & Eppler, D. (1995). The acquisition, transfer, and depreciation of knowledge in service organizations: Productivity in franchises. *Management Science*, 41(11), 1750–1762.

- Dekimpe, M. G., Parker, P. M., & Sarvary, M. (2000). Global diffusion of technological innovations: A coupled-hazard approach. *Journal of Marketing Research*, 37(1), 47–59.
- Demsetz, H. (1982). Barriers to entry. *American Economic Review*, 72(1), 47–57.
- Dixit, A. (1979). A model of duopoly suggesting a theory of entry barriers. *Bell Journal of Economics*, 10(1), 20–32.
- Dosi, G. (1997). Opportunities, incentives and the collective patterns of technological change. *The Economic Journal*, 107(3), 1530–1547.
- Dutta, S., & John, G. (1995). Combining lab experiments and industry data in transaction cost analysis: The case of competition as a safeguard. *Journal of Law, Economics, and Organization*, 11, 87–111.
- Dyer, J. H. (1995). Specialized supplier networks as a source of competitive advantage: Evidence from the auto industry. *Strategic Management Journal*, 11(3), 42–55.
- Eliashberg, J., & Robertson, T. S. (1988). New product preannouncing behavior: A market signaling study. *Journal of Marketing Research*, 25(3), 282–292.
- Fein, A. J., & Anderson, E. (1997). Patterns of credible commitments: Territory and category selectivity in industrial distribution channels. *Journal of Marketing*, 61, 19–34.
- Gatignon, H., & Anderson, E. (1988). The multinational corporation's degree of control over foreign subsidiaries: An empirical test of a transaction cost explanation. *Journal of Law, Economics, and Organization*, 4(2), 305–336.
- Gatignon, H., & Robertson, T. S. (1985). A propositional inventory for new diffusion research. *Journal of Consumer Research*, 11(4), 859–867.
- Gatignon, H., & Robertson, T. S. (1989). Technology diffusion: An empirical test of competitive effects. *Journal of Marketing*, 53(1), 35–49.
- Gatignon, H., & Robertson, T. S. (1991). Innovative decision processes. In H. Kassarian & T. S. Robertson (Eds.), *Handbook of consumer behavior theory and research* (pp. 316–348). Prentice-Hall.
- Gatignon, H., & Xuereb, J.-M. (1997). Strategic orientation of the firm and new product performance. *Journal of Marketing Research*, 34(1), 77–90.
- Gatignon, H., Eliashberg, J., & Robertson, T. S. (1989). Modeling multinational diffusion patterns: An efficient methodology. *Marketing Science*, 8(3), 231.
- Gatignon, H., Robertson, T. S., & Fein, A. J. (1997). Incumbent defense strategies against new product entry. *International Journal of Research in Marketing*, 14(2), 163–176.
- Gielen, K., & Dekimpe, M. G. (2001). Do International entry decisions of retail chains matter in the long run? *International Journal of Research in Marketing*, 18(3), 235–259.
- Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technical change. *Econometrica*, 25(4), 501–522.
- Griliches, Z., Hall, B., & Pakes, A. (1991). R&D, patents, and market value revisited: Is there a second (technological opportunity) factor? *Economic Innovation and New Technology*, 1(1), 183–201.
- Gulati, R., & Singh, H. (1998). The architecture of cooperation: Managing coordination costs and appropriation concerns in strategic alliances. *Administrative Science Quarterly*, 43(4), 781–814.
- Haspeslagh, P., & Jemison, D. (1991). *Managing acquisitions*. Free Press.
- Heide, J. B., & John, G. (1988). The role of dependence balancing in safeguarding transaction-specific assets in conventional channels. *Journal of Marketing*, 52(1), 20–35.
- Helsen, K., Jedidi, K., & DeSarbo, W. S. (1993). A new approach to country segmentation utilizing multinational diffusion patterns. *Journal of Marketing*, 57, 60–71.
- Henard, D. H., & Szymanski, D. M. (2001). Why some new products are more successful than others. *Journal of Marketing Research*, 38(3), 362–375.
- Herr, P. M., Kardes, F. R., & Kim, J. (1991). Effects of word-of-mouth and product attribute information on persuasion: An accessibility-diagnostics perspective. *Journal of Consumer Research*, 17, 454–462.
- Jap, S. D. (1999). 'Pie-expansion' efforts: Collaboration processes in buyer-supplier relationships. *Journal of Marketing Research*, 36, 461–475.
- Kalyanaram, G., & Urban, G. L. (1992). Dynamic effects of the order of entry on market share, trial penetration, and repeat purchases for frequently purchased consumer goods. *Marketing Science*, 11, 235–250.

- Katz, E. (1957). The two-step flow of communication: An up-to-date report on an hypothesis. *Public Opinion Quarterly*, 21, 61–78.
- Kaul, A., & Rao, V. R. (1995). Research for product positioning and design decisions: An integrative review. *International Journal of Research in Marketing*, 12(4), 293–320.
- Kirmani, A., & Rao, A. R. (2000). No pain, no gain: A critical review of the literature on signaling unobservable quality. *Journal of Marketing*, 64, 66–79.
- Kogut, B., & Kulatilaka, N. (2001). Capabilities as real options. *Organization Science*, 12(6), 744–758.
- Krishnan, V., & Ulrich, K. T. (2001). Product development decisions: A review of the literature. *Management Science*, 47(1), 1–21.
- Krueger, A. B., & Taylor, T. (2000). An interview with Zvi Griliches. *Journal of Economic Perspectives*, 14(2), 171–189.
- Kuester, S., Gatignon, H., & Robertson, T. S. (2000). Firm strategy and speed of diffusion. In V. Mahajan, E. Muller, & Y. Wind (Eds.), *New-product diffusion models*. Kluwer Academic Publishers.
- Leonard-Barton, D. (1992). Core capabilities and core rigidities: A paradox in managing new products. *Strategic Management Journal*, 13(1), 11–125.
- Mahajan, V., Muller, E., & Bass, F. (1993). New product diffusion models. In J. Eliashberg & G. L. Lilien (Eds.), *Handbooks in operations research and management science* (pp. 349–408). Elsevier Science.
- Mahajan, V., Muller, E., & Wind, Y. (Eds.). (2000). *New-product diffusion models*. Kluwer Academic Publishers.
- Mascarenhas, B. (1992). Order of entry and performance in international markets. *Strategic Management Journal*, 13, 499–510.
- McGrath, R. G. (1997). A real options logic for initiating technology positioning investments. *Academy of Management Review*, 22(4), 974–996.
- Moorman, C. (1995). Organizational market information processes: Cultural antecedents and new product outcomes. *Journal of Marketing Research*, 32(3), 318–335.
- Moorman, C., & Miner, A. S. (1997). The impact of organizational memory on new product performance and creativity. *Journal of Marketing Research*, 34(1), 91–106.
- Nayak, P. R., & Kettingham, J. M. (1986). *Breakthroughs!* Rawson Associates.
- Nelson, R. R. (1995). Recent evolutionary theorizing about economic change. *Journal of Economic Literature*, 32, 48–90.
- Nerlove, M. (2001). Zvi Griliches, 1930–1999: A critical appreciation. *The Economic Journal*, 111(3), F422–F448.
- Osborn, R. N., & Hageboom, J. (1997). The institutionalization and evolutionary dynamics of interorganizational alliances and networks. *Academy of Management Journal*, 40(2), 261–278.
- Oxley, J. E. (1999). Institutional environment and the mechanisms of governance: The impact of intellectual property protection on the structure of inter-firm alliances. *Journal of Economic Behavior and Organization*, 38, 283–309.
- Parker, P., & Gatignon, H. (1994). Specifying competitive effects in diffusion models: An empirical analysis. *International Journal of Research in Marketing*, 11(1), 17–23.
- Pawakapan, N. (2000). Trade and traders: Local becomes national. *Journal of Southeast Asian Studies*, 31(2), 374–389.
- Pisano, G. P. (1990). The R&D boundaries of the firm: An empirical analysis. *Administrative Science Quarterly*, 35(3), 153–176.
- Putsis, W. P., Balasubramanian, S., Kaplan, E. H., & Sen, S. K. (1997). Mixing behavior in cross-country diffusion. *Marketing Science*, 16(4), 354–369.
- Regan, L., & Tennyson, S. (1996). Agent discretion and the choice of insurance marketing systems. *Journal of Law and Economics*, 39, 637–666.
- Robertson, T. S. (1971). *Innovative behavior and communication*. The Free Press.
- Robertson, T. S., & Gatignon, H. (1986). Competitive effects on technology diffusion. *Journal of Marketing*, 50(3), 1.
- Robertson, T. S., & Gatignon, H. (1998). Technology development mode: A transaction cost conceptualization. *Strategic Management Journal*, 19(6), 515–531.
- Rogers, E. M. (1983). *Diffusion of innovations*. The Free Press.

- Sako, M., & Helper, S. (1998). Determinants of trust in supplier relations: Evidence from the automotive industry in Japan and the United States. *Journal of Economic Behavior and Organization*, 34(3), 387–417.
- Scherer, F. M., & Ross, D. (1990). *Industrial market structure and economic performance* (3rd ed.). Houghton Mifflin.
- Schmookler, J. (1966). *Invention and economic growth*. Harvard University Press.
- Schumpeter, J. A. (1943). *Capitalism, socialism, and democracy*. George Allen & Unwin.
- Sethi, S. P. (1971). Comparative cluster analysis for world markets. *Journal of Marketing Research*, 8(3), 348–354.
- Shankar, V., Carpenter, G. S., & Krishnamurthi, L. (1998). Late mover advantage: How innovative late entrants outsell pioneers. *Journal of Marketing Research*, 35(1), 54–70.
- Spence, A. M. (1977). Entry, capacity, investment and oligopolistic pricing. *Bell Journal of Economics*, 8(2), 534–544.
- Spulber, D. F. (1999). *Market microstructure: Intermediaries and the theory of the firm*. Cambridge University Press.
- Sultan, F., Farley, J. U., & Lehmann, D. R. (1990). A meta-analysis of applications of diffusion models. *Journal of Marketing Research*, 27(1), 70–77.
- Tapon, F. (1989). A transaction costs analysis of innovations in the organization of pharmaceutical R&D. *Journal of Economic Behavior & Organization*, 12(2), 197–213.
- Teece, D. J. (1992). Competition, cooperation, and innovation. *Journal of Economic Behavior and Organization*, 18, 1–25.
- Teece, D. (1996). Firm organization, industrial structure, and technological innovation. *Journal of Economic Behavior and Organization*, 31, 193–224.
- Terpstra, V., & Yu, C.-M. J. (1990). Piggybacking: A quick road to internatlisation. *International Marketing Review*, 7(4), 52–63.
- Urban, G. L., & Von Hippel, E. (1988). Lead user analyses for the development of new industrial products. *Management Science*, 34(5), 569–582.
- Urban, G. L., Carter, T., Gaskin, S., & Mucha, Z. (1984). Market share rewards to pioneering brands: An empirical analysis and strategic implications. *Management Science*, 32(6), 645–659.
- Van den Bulte, C., & Moenaert, R. K. (1998). The effects of R&D team co-location on communication patterns among R&D, marketing, and manufacturing. *Management Science*, 44(11), S1–S18.
- Wathne, K. H., & Heide, J. B. (2000). Opportunism in interfirm relationships: Forms, outcomes, and solutions. *Journal of Marketing*, 64(4), 36–51.
- Williamson, O. E. (1979). Assessing vertical market restrictions. *University of Pennsylvania Law Review*, 127, 953–993.
- Williamson, O. E. (1996). *The mechanisms of governance*. Oxford University Press.
- Williamson, O. E. (1999). Strategy research: Governance and competence perspectives. *Strategic Management Journal*, 20(3), 1087–1108.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Martin Peitz

Contents

23.1	Introduction: Platforms and the Management of Complex Ecosystems	566
23.2	Governance Decisions by Platforms	568
23.2.1	Pricing	568
23.2.2	Endogenous Strength of Network Effects on the Platform	572
23.2.3	The Platform Within Its Broader Ecosystem	574
23.2.4	Consumer Steering and Self-Preferencing	576
23.3	Platform Decisions with Cross-Platform Spillovers	578
23.3.1	Platform Exclusivity	578
23.3.2	Price-Parity Clauses	580
23.3.3	Bundling and Tying	582
23.4	Public Regulation	583
23.4.1	Merger Control	584
23.4.2	“Regulating” Conduct Under Competition Law or Specific Regulation	585
23.4.3	Public Governance, Tort Law, and Other Regulations	586
23.5	Conclusion	588
	Bibliography	589

Abstract

In this chapter, we discuss how platforms manage the interaction between various users. First, we discuss and exemplify governance decisions by platforms that affect access and interactions of users regarding a platform service. Here, we investigate the choice of price structure and the choice of non-price strategies. We also address the horizontal and vertical scope of these platforms. Second, we

The author thanks two reviewers, the Editors Claude Ménard and Mary Shirley, Paul Belleflamme, and Markus Reisinger for the helpful comments. He discloses that he frequently advises competition authorities on issues related to this chapter and gratefully acknowledges support from the Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224 (Project B05).

M. Peitz (✉)

Department of Economics, University of Mannheim and MaCCI, Mannheim, Germany

consider platform decisions that generate spillovers to other platforms or channels, and we explore private incentives and welfare effects. Third, we discuss the role of government regulation in a broad sense, that is, the laws and regulations that constrain platforms and shape their incentives regarding their governance decisions. Emphasis is given to interventions against anticompetitive conduct and practices that may lead to consumer harm.

Keywords

Platform governance · Platform regulation · Digital ecosystems · Digital markets · Competition policy · Network effects

23.1 Introduction: Platforms and the Management of Complex Ecosystems

Platforms can be broadly defined as managed marketplaces in which interactions between platform users take place and which typically feature network effects; that is, the user benefits depend on the decisions of other users (Belleflamme & Peitz, 2018a, 2021). Many market interactions have always relied on the services provided by platforms. Traditionally, secondhand items changed hands, thanks to announcements made in the classified ads section of a newspaper or flea markets providing space for sellers to display their wares. Manufacturer products and some services also, in part, relied on shopping malls as physical platforms that allowed brand manufacturers to offer their products in their own or franchised shops, or shops operated by multi-brand retailers.

With more and more interactions being facilitated through digital intermediaries and other activities moving into the digital sphere (e.g., music and video streaming as a substitute for music and video made available on physical devices), most of the recent public interest and research activities concern digital platforms. Platforms contribute to solving market failures when they make it possible for users to easily find good matches and when they generate trust among users. In a broad sense, platforms can be seen as reducing transaction costs. Moreover, the value that platforms create for their users tends to grow organically, thanks to positive network effects: As more users join a given platform, interacting on this platform becomes more valuable, which contributes to attracting even more users, serving as a self-reinforcing mechanism. For instance, more participation may increase the likelihood of encountering a good match or allow the platform to put mechanisms in place that foster trust (like reputation mechanisms). In a nutshell, platforms create value by bringing economic agents together and may engage in complex strategies to manage network effects to generate economic value.

Network effects can be of different types, as illustrated by the example of software platforms. These platforms bring together application developers and end users. End users may benefit from the increased presence of other users, leading to *direct* or *within-group network effects* (for instance, through the exchange of tips or

the fixing of bugs). Users may also enjoy a larger number and quality of application developers, leading to positive *cross-group network effects* from app developers to end users. On the other side of the platform, developers may take advantage of a larger number and a more intensive usage of end users, leading to positive cross-group network effects from end users to platform developers. Mutual positive cross-group network effects then generate positive *indirect network effects*. The term “indirect” refers to the fact that end users care indirectly about the participation and usage of other end users because more end users attract more developers, which is beneficial for every end user. Correspondingly, developers also experience positive indirect network effects. The platform manages the interaction among users, taking into account the various network effects that exist among them. If the platform addresses the two different user groups differently, we call the platform two-sided. Network effects may arise because a user cares about the presence and engagement of other users or because other users leave a footprint that matters. The latter may be the result of data in which case network effects are data-enabled (Hagiu & Wright, 2023). For example, recommender systems lead to such data-enabled network effects when they provide valuable information to users when using data collected from other users (Belleflamme & Peitz, 2021, Chap. 2).

Some platforms allow for the interaction of buyers and sellers. Non-digital platforms of this type include trade fairs, flea markets, auction houses, and yellow pages. Shopping malls are another example, as they offer retail space to sellers and invite buyers to go shopping. All else being equal, sellers prefer a shopping mall that attracts more buyers, and buyers prefer a shopping mall that hosts more sellers. Shopping malls are actively managed with different rental contracts applied to different types of shops: for instance, anchor stores are used to attract traffic to a shopping mall and therefore receive more favorable rates (or are vertically integrated).¹ While some of these platforms have been around for a long time, platforms as “orchestrators” of market activities have arguably gained more prominence with the rise of the Internet. To enable consumers to choose among a myriad of offerings, horizontal and vertical search engines as well as price search engines, booking portals, online auctions, and e-commerce platforms have become commonplace. As many of these digital platforms are not subject to physical capacity constraints and can quickly scale up their activities, they can swiftly conquer whole countries and thus become or at least appear to be dominant.

A for-profit platform that is successful in generating value can use monetization strategies to appropriate part of the generated value and, possibly, make even higher profits if it can impair users’ outside options. Platforms, and in particular digital platforms, can set rules on how transactions can be enabled and possibly also how contracts are enforced. Thus, *a platform manages or orchestrates interactions between different users*. Part of its activity is to police the digital ecosystem that

¹ Thus, buyers and sellers may also care about the composition of users on the other side of the platform (and consumers may also care about the composition on the same side). Pashigian and Gould (1998) analyzed how shopping malls internalize externalities.

has formed around the platform and includes offers and connected services by third parties. For example, in an e-commerce setting, it may have developed rules on how to deal with sellers that offer counterfeit products or that sell products in a way that does not comply with certain quality standards. This belongs to what can be called the governance of the platform. The platform may also set contractual terms that constrain user behavior outside the platform. Price-parity clauses (often called platform MFNs in the USA) are an example of this in a buyer-seller context. Such clauses prevent sellers from offering better terms on a competing platform and/or when dealing directly with buyers. Platforms will have to respect the laws enacted by parliaments and decisions by regulators. Thus, a platform's governance decisions are possibly constrained by regulators' decisions.

Complementary works: This chapter complements several other overviews and surveys. In particular, we draw on the monograph by Belleflamme and Peitz (2021). The following are complementary surveys: Belleflamme and Peitz (2018a) give an introduction on platform economics with a particular focus on monopoly pricing; Jullien, Pavan, and Rysman (2022) focus their exposition on a platform's pricing decision covering monopoly and oligopoly settings; Peitz and Reisinger (2016) provide an overview on ad-funded content platforms; Belleflamme and Peitz (2018b) provide a survey on rating and recommender systems, which are integral components of many digital platforms; Jullien and Sand-Zantman (2021) review academic research on competition policy issues around multi-sided platforms. In addition to the economics literature on multi-sided platforms, there exists a related literature in strategic management on platforms (see, e.g., McIntyre et al., 2021).

Organization of the chapter: This chapter contains three sections. In Sect. 23.2, we address governance decisions that are aimed at managing interactions on platforms. Here, we first look at a platform's pricing decision and then turn to several non-price decisions. In Sect. 23.3, we consider several platform strategies that directly affect competition between platforms or competition with trading outside the platform. In Sect. 23.4, we address public policies that impose restrictions on what a platform can do (or even mandate certain behaviors).

23.2 Governance Decisions by Platforms

Prices set by the platform allow the platform to monetize its service; at the same time, the choice of the price structure can be seen as an instance of managing participation and interaction of the platform. We will first take a look at the price structure of a two-sided platform before turning to non-price strategies.

23.2.1 Pricing

To obtain a first understanding of the choice of price structure, we look at one platform in isolation. The platform owner offers a service that may have stand-alone value r_i for users of a certain group i and may offer benefits that depend on the usage

and participation decisions of other users. To fix ideas, suppose that the platform caters to two user groups a and b and that each user cares about the number of users from the same and/or the other group in a linear fashion. For example, one may set the outside option to zero and write a user's valuation on the platform in an additively separable form with linear network effects; that is

$$v_a = r_a + \alpha_a n_a + \beta_a n_b - A_a \quad (23.1)$$

for users of group a and

$$v_b = r_b + \alpha_b n_b + \beta_b n_a - A_b \quad (23.2)$$

for users of group b , where α_i is the strength of the within-group network effects, β_i is the strength of the cross-group network effect, n_i is the size of the participating user group i , and A_i is the access or participation fee charged to users of group i .

For the given network effect parameters α_a , α_b , β_a , and β_b and stand-alone parameters r_a and r_b , the platform can manage participation through its participation fees. The work following Armstrong (2006) focuses on the pricing of access to a platform in the presence of cross-group network effects only (i.e., $\alpha_b = \alpha_a = 0$).

For any given access fees, there may exist multiple consumer participation equilibria including two stable equilibria: one with zero participation in both user groups and the other with positive participation by both user groups. If consumers tend to coordinate on the outside option, the platform owner may then choose an asymmetric price structure, even in a symmetric environment, to make sure that users in the group with the lower price (say group a) will participate. If all users observe the full price structure, users in group b then infer that (many) users in group a participate. This induces many users in group b to join even when they face a higher price. This is an instance in which the platform owner uses an asymmetric price structure to solve the chicken-and-egg problem (also sometimes referred to as the mutual baiting problem). The asymmetric price strategy in response to this problem is called a divide-and-conquer strategy.

If the two user groups are different, a profit-maximizing platform owner is not indifferent as to which group to use as bait. When applying a divide-and-conquer strategy, a monopoly platform tends to subsidize the group that exerts the largest cross-group network effect on the other group and monetizes users in the other group (irrespective of the relative size of the two groups).² The fear of user miscoordination may also make it more attractive to monetize through transaction fees rather than access fees. With access fees, users may not be confident to find counterparts in the other group. Hence, they may be reluctant to pay a membership fee upfront, as they fear not being able to conduct any transaction once subscribed to the platform.

²For a formal investigation, see Belleflamme and Toulemonde (2009); for a summary, see Belleflamme and Peitz (2021, Chap. 4). For the use of divide-and-conquer strategies under Bertrand competition between two platforms, see Caillaud and Jullien (2003). Here, the result will be market tipping, and the only active firm may not make any profit.

If instead the platform resorts to transaction fees, the fear is unwarranted as transaction fees are only paid if an effective transaction takes place.

Suppose instead that the platform does not have to deal with the coordination problem of users; that is, consumers are assumed to coordinate on the participation equilibrium that is most favorable to the platform. To understand the monopolist's pricing incentives, we compare how the price structure chosen by the monopolist differs from the one chosen by a social planner who maximizes total surplus. In the presence of positive cross-group network effects, the welfare-maximizing solution features access fees below the marginal cost of serving an additional user. For several reasons, a profit-maximizing platform chooses a price structure that differs from the one that would maximize total surplus, but has some resemblance: the monopoly platform restricts output (market-power distortion), cares about marginal users rather than about average users (Spence distortion), and induces different interaction benefits (displacement distortion) and different participation rates (scale distortion) —see Weyl (2010) and Tan and Wright (2018, 2021). Depending on the specifics of the environment (strengths of the cross-group network effects, costs, and stand-alone benefits), the combined impact of these distortions may lead the monopoly platform to set prices that are below or above the efficient level on either side. However, the monopoly platform will never set prices below the efficient level on both sides.

To analyze how the monopoly platform sets its participation fees, we return to the specific setting developed above and derive the “demands for participation” for each user group.³ Suppose for simplicity that, in each group, the value of the outside option is uniformly distributed over some sufficiently large interval. Then, the number of users who decide to participate in group i , n_i , is simply equal to v_i .⁴ Setting $v_a = n_a$ and $v_b = n_b$ in Eqs. (23.1) and (23.2) above and rearranging terms, we find:

$$(1 - \alpha_a)n_a = r_a + \beta_a n_b - A_a \text{ and } (1 - \alpha_b)n_b = r_b + \beta_b n_a - A_b.$$

We see that participation in one group depends on participation in the other group, and vice versa. The next step consists of solving this system of two equations in n_a and n_b . A useful shorthand notation for the solutions is $n_a(A_a, A_b)$ and $n_b(A_a, A_b)$. It is indeed important to emphasize that participation on each side depends on *both* participation fees. In particular, if the network effect parameters respect some conditions, participation decreases as any fee increases.⁵ The intuition is simple. Fewer users decide to participate if they or users in the other group are charged a larger fee. The first effect is common (this is the expression of the Law of Demand). The second effect is peculiar to two-sided platforms, as it follows from the presence

³Because of network effects, participation decisions are interdependent, and demands are derived by solving for the Nash equilibrium of the “participation game” that users play.

⁴A user of group i participates if and only if $v_i \geq X$, where X denotes the user's outside option. Given the uniform distribution of X with density 1, the number of users such that $X \leq v_i$ is just equal to v_i .

⁵See Belleflamme and Peitz (forthcoming) for explicit expressions of the demands and the conditions to be imposed on the parameters.

of positive cross-group network effects: if a fee increase leads to fewer users participating in, say, group b , then users in group a are less keen to participate as interacting on the platform becomes less valuable.

When choosing its fees, the platform internalizes these effects. Typically, as explained by Armstrong (2006), if users in group a exert a positive cross-group network effect on users in group b , then the platform has an extra incentive to lower the price on side a because attracting more users on side a also allows the platform to raise more revenues on side b (and not just on side a as would be the case in the absence of network effects). Note that, if the two groups are not symmetric, this logic may drive the platform to subsidize the participation of one group of users (i.e., to set the price below the marginal cost, which can be implemented, e.g., through cash-backs or in-kind payments). This is likely to be the case when one group exerts a positive cross-group network effect on the other while the other does the opposite, as is often the case when one of the two groups contains advertisers and advertising is considered a nuisance by consumers, who form the second user group.

The previous reasoning relies on two important conditions. First, the platform must be able to freely set its fees. This is not always the case as platforms may face constraints that prevent them from setting their optimal prices. For instance, below-cost prices may be prohibited or infeasible. Then, a platform may be limited in its ability to internalize cross-group network effects.⁶

The second condition is that users on each side can observe variations in prices on the other side. The previous reasoning indeed relies on the fact that users in group b react to a change in A_a , which supposes that they can observe such a change. This may not be obvious in the case, for instance, of software platforms, which act as intermediates between end users and application developers (e.g., smartphone operating systems or video game consoles). Although the two groups are linked by mutual positive cross-group network effects, they do not interact directly with one another. This prevents, in particular, end users from having a clear view of what and how much the platform charges app developers. In such a context, end users must base their participation decision on some predictions of what the participation level of app developers will be. Since users do not observe the price charged to the other group, the platform is tempted to raise this price too much for its own good—this is an instance of the classic opportunism problem (Hart & Tirole, 1990). The platform may then want to devote resources to make prices charged to the other group observable, for instance, through dedicated advertising. Gross of the associated costs, this yields higher profits but also higher surpluses for all users.⁷

⁶See Belleflamme and Peitz (2021, Sect. 5.3.1) for a formal treatment. The zero-price constraint is of high relevance in the case of ad-funded media platforms (including social networks) because often the cross-group network effect exerted by advertisers on consumers is negative (i.e., consumers consider advertising to be a nuisance). For formal models of platform competition with one-sided pricing, see Anderson and Coate (2005) and Anderson and Peitz (2020).

⁷Hagiu and Halaburda (2014) and Belleflamme and Peitz (2019c) study this issue. They also show that competition among platforms may attenuate or reverse the platforms' incentives to advertise non-observed prices.

Finally, it is also important to stress that the previous reasoning (according to which a platform adjusts its prices to internalize network effects) does not rest on the presence of easily distinguishable groups of users but on the platform's ability to target groups with different prices. Many platforms cater mostly to a single group of users (think, e.g., of social networks and messaging applications) and manage the direct network effects that exist within this group. They may, nevertheless, be able to segment their single audience into subgroups that differ along some characteristic and condition prices on these user characteristics. If so, Belleflamme and Peitz ([forthcoming](#)) show that a platform catering to a single—but segmented—audience chooses prices very much like a multi-sided platform.

Our discussion focused on monopoly platforms; we confine ourselves to pointing to some works that address competition between two-sided platforms. Assuming that users are either singlehomers (i.e., they join at most one platform) or (potential) multihomers (i.e., they consider joining more than one platform), one can distinguish between three oligopoly settings of two-sided platforms: singlehoming by both groups (as analyzed by Armstrong, [2006](#); Tan & Zhou, [2021](#); Peitz & Sato, [2023](#)), singlehoming by one group and multihoming by the other (Armstrong, [2006](#); Sect. 6 in Anderson & Peitz, [2020](#)) including purely ad-funded platforms (Anderson & Coate, [2005](#); Anderson & Peitz, [2020](#)), and multihoming by both groups (Bakos & Halaburda, [2020](#)). Rochet and Tirole ([2003](#), [2006](#)) characterize the outcome when platforms charge transaction fees and users have different inclinations to use a platform. Teh et al. ([2023](#)) consider the setting in which users in both groups first make their participation decision (e.g., merchants decide which payment cards to accept and consumers decide which cards to carry) and then users in one group decide which available option to pick (e.g., each consumer chooses which of the cards to pick that they carry and are accepted by the merchant).

23.2.2 Endogenous Strength of Network Effects on the Platform

A platform may become active in multiple ways to affect stand-alone benefits and the strength of the network effects that users on its platform experience. The way users interact with each other determines the strength of the network effects, and the platform can often affect this strength through price and non-price instruments.

Platforms operating e-commerce marketplaces, which enable trade between sellers and buyers, are a case in point. The strength of network effects is endogenous and depends on the degree of seller competition (Nocke et al., [2007](#); Hagiu, [2009](#); Belleflamme & Peitz, [2019a](#)), which in turn can be affected by the platform. A monopoly platform may want to increase the horizontal differentiation among the products that it lists (by selecting the appropriate sellers, removing the visibility of some sellers, steering consumers to a subset of sellers, or by influencing how buyers perceive the differentiation).

An extreme version of reducing competition between sellers is to grant improved visibility or category exclusivity to one seller. An example is exclusivity for certain category sellers in shopping malls; see the empirical study of exclusivity for burger

restaurants in Israeli shopping malls by Ater (2015). Another example is the agreement between Amazon and the brand manufacturer Apple, according to which only Amazon and selected sellers are allowed to sell Apple and Beats products. According to the Italian competition authority, which investigated the case, other sellers were excluded (this finding was contested by Amazon and Apple); such a practice is also investigated by the German Cartel Office. Limiting seller competition by a platform can be a response to platform competition and explain the coexistence of profitable non-differentiated platforms (Karle et al., 2020).

A more general approach to analyzing a platform's non-price strategy and its effects on the strength of network effects is provided by Teh (2022) as well as Choi and Jeon (2023). The platform may take some, possibly costly, action that affects α_a , α_b , β_a , and/or β_b in Eqs. (23.1) and (23.2) as well as stand-alone benefits. For example, in an e-commerce setting in which a platform charges a mix of participation fees and ad valorem transaction fees and each seller is a monopolist in its product category, with linear demand $1 - p$ and zero marginal costs of production, the profit-maximizing seller makes a per-buyer profit of $\beta_a = 1/4$, and each consumer obtains a per-seller surplus of $\beta_b = 1/8$ under uniform pricing, while the corresponding values under perfect price discrimination would be $\beta_a = 1/2$ and $\beta_b = 0$ (gross of any fees). Thus, by disclosing consumer valuations to sellers, the platform affects the strength of network effects.

Price competition between sellers can be affected through other means by the platform. For instance, in de Cornière (2016), a monopoly platform with a fixed advertising fee can reduce the accuracy of targeting. Doing so induces buyers to search less, which deteriorates the match quality and relaxes competition between sellers. Another instance is Karle and Peitz (2017) in which a platform taxes seller profits and can enlarge consumers' consideration set. With expectation-based loss-averse consumers, this manipulates consumers' reference points and thereby relaxes competition.

Several works have looked at other environments in which a platform (or competing platforms) makes decisions that manage the interaction between the different user groups and thereby affect the strength of network effects.⁸ In these environments, one can study whether and to which extent platform incentives are aligned with buyer and/or seller incentives (and possibly distinguish between

⁸Dinerstein et al. (2018) theoretically and empirically analyze a platform's decision on how much to steer consumers to their most desired product taking into account the sellers' response in their pricing decision. Johnson, Rhodes, and Wildenbeest (2023) consider a platform's demand-steering rules that reward sellers when they cut prices. A platform may decide on delisting or demoting low-quality sellers or to delist sellers of counterfeits (Casner, 2020; Hua & Spier, 2023; Jeon et al., 2021). Short of delisting, a platform may design its rating and recommendation systems such that inferior sellers are more easily identified or become less visible (Belleflamme & Peitz, 2018b). It may introduce deceptive features (Johnen & Somogyi, 2022) or engage in content moderation (Liu et al., 2021; Madio & Quinn, 2023). Instead of disclosing consumer valuations to sellers, it may give consumers the possibility to voluntarily disclose some information on their valuation to sellers. Here, the platform chooses a disclosure technology that affects the strength of network effects (Gambato & Peitz, 2023, building on Ali et al., 2023).

different types of sellers or buyers), possibly depending on the price instruments available to the platform. Some of the non-price strategies require investments by the platform, and the platform will invest if the marginal cost is less than the extra benefit that is extracted by the platform.

Depending on the type of non-price strategy and the environment in which the platform operates, either one or both groups benefit. In the example about the platform providing consumer information to sellers that enables them to perfectly price discriminate, such a non-price strategy benefits sellers and harms consumers. By contrast, if sellers offer an experience good and, thus, are subject to a moral hazard problem resulting in low product quality, a non-manipulated rating system can be an important source of information for subsequent users (e.g., they learn about the quality of the rated hotels on a hotel booking platform). If sellers choose high quality in response to the rating system, the expected gains from trade rise and cross-group network effects are likely to be stronger for both user groups (consumers and sellers).⁹

23.2.3 The Platform Within Its Broader Ecosystem

A platform may take a central position in an ecosystem and decide on how much to extend its reach, both in terms of its horizontal scope (e.g., which product categories to cover and which consumer segments to address) as well as its vertical scope (which added service to integrate or offer through complementors).

Digital platforms can often (but not always) easily scale their business to other products and consumer segments. Scope economies are favorable for increasing the horizontal scope. For example, Amazon's investments in its logistics network allow it to easily add new product categories with FBA offers (fulfilled by Amazon) to its marketplace. In general, certain assets may be used broadly (e.g., the brand and the associated consumer trust, certain software components, or AI capabilities). Also favorable are data-enabled network effects across different services or consumer segments (de Cornière & Taylor, 2020) and consumer benefits from one-stop shopping. Finally, a platform may strategically use bundling and tying to increase its scope; we provide some more detail on this issue in the next section.

Regarding the vertical scope, an e-commerce platform may provide warranties, insurance, and integrated payments. It may offer warehousing services and provide delivery services. It may fully vertically integrate some activities. For example, in ad tech, Google and Facebook have (partially) vertically integrated and are active in multiple layers of the value chain.

A platform (e.g., a hotel booking platform) may face the threat that some users meet on its marketplace but complete the transaction off the platform. For a platform that raises its revenues through transactions, this constitutes a revenue loss—this

⁹For further examples of the impact of a platform's non-price strategy on network effects, see Sect. 4.3 in Belleflamme and Peitz (2021).

phenomenon is called platform leakage. Here, the platform provides a showrooming service. Platforms can combat leakage through a number of measures: they may make it more difficult for users to transact off the platform (e.g., on AirBnB by hiding the identity and contact information of the transaction partner), by delisting or demoting sellers that use the platform as a showrooming service, by removing consumers' incentives to transact off the platform through price-parity clauses (e.g., on hotel booking platforms such as Booking; see Hunold et al., 2020),¹⁰ and/or by offering additional benefits for completing a transaction on the platform (e.g., on Amazon Marketplace through superior logistics or payment options). The platform may also adjust its monetization model and rely less on transaction fees and more on advertising or referral fees.¹¹

A platform may not fully vertically integrate certain products or services but operate in dual mode; that is, the platform admits third-party providers but also offers services or products itself.¹² Consider a setting in which a platform charges sellers for the transactions on a platform. A possible defense for the practice of introducing first-party offers is that a platform may want to provide an anchor for retail prices of third-party sellers. This is relevant in markets with little competition between third-party sellers.¹³ In this case, the platform as a guardian of the ecosystem may be worried about consumers receiving a bad deal and therefore introduce a first-party product to stimulate competition.¹⁴

In Anderson and Bedre-Defolie (forthcoming), a monopoly firm can operate as a pure retailer, as a platform running a marketplace with third-party sellers, or as a platform in dual mode running a marketplace on which it also sells products as a retailer itself. A platform in dual mode sets the retail price of its own product and a percentage transaction fee; third-party sellers observe these prices and decide whether to enter and, if so, set their retail prices; finally, buyers make purchasing decisions. In that setting, prohibiting the dual mode increases consumer surplus if and only if the prohibition leads to a pure marketplace.¹⁵

¹⁰The competitive effects of price-parity clauses are discussed in Sect. 23.3.

¹¹For a formal analysis of a monopoly platform's responses to the leakage problem, see Hagiu and Wright (2024).

¹²The exposition on the dual mode is taken partly verbatim from Peitz (2022b).

¹³Take as an extreme case a situation of full seller collusion and step demand, which implies that sellers will charge the monopoly price that is independent of the level of the fee charged by the platform.

¹⁴This may be a more attractive option for the platform than lowering fees charged to sellers; in particular, if such fee reductions are not fully passed through to consumers. In such a case, a platform is particularly inclined to introduce those first-party offers for which it has a cost or quality advantage over third-party sellers.

¹⁵For further theoretical work, see Etro (2023a). Crawford et al. (2022) empirically assess the effect of Amazon's retail entry competing against third parties offering the same product. They find that entry is correlated with high growth and a low degree of competition. Overall, they read their findings as Amazon internalizing externalities, which makes the platform more attractive to consumers. A different market expansion effect can arise if a platform invites the entry of successful offline brands (Jin et al., 2021).

If the marketplace includes product categories in which innovative sellers may appear, the marketplace helps consumers in the discovery process and limits the market power of an innovative seller. According to Hagiu, Teh, and Wright (2022), this implies that the dual mode always gives higher consumer welfare than the pure marketplace. Furthermore, a ban on the dual mode never increases consumer welfare.¹⁶

When operating in the dual mode, the platform may use information on the success of third-party sellers to decide which product category to enter.¹⁷ Some research looks at the dynamic effects this might have. First, a third party may anticipate the platform's imitation decision in the case of high demand and hide information related to demand (Jiang et al., 2011). Alternatively, third-party sellers may reduce investment¹⁸ or opt for product categories for which it is known that demand is low so that the risk of the platform entering with a first-party product is also low. To address the concern of underinvestment and distorted entry by third-party sellers because of the imitation threat, a possible remedy is to ban the platform (or at least its first-party division) from having access to any private information generated by the third-party seller (Hagiu et al., 2022). However, a platform with access to this information may operate more efficiently, and just banning the first-party division from accessing this information may be difficult to enforce. Another possible remedy is to prohibit the platform from entering new product categories with first-party products for a certain amount of time (Madsen & Vellodi, forthcoming).

23.2.4 Consumer Steering and Self-Preferencing

Product recommendations and rating information play an important role in the consumer experience, and when platforms compete, the platform with the recommendation or ratings system that better serves consumer interests may have the edge over its competitors.¹⁹ However, platforms with market power may make product recommendations or modify the rating system to serve their own interests, which may well be different from consumers' best interests. Several contributions provide formal arguments that platforms as pure intermediaries may make recommendations

¹⁶Other contributions include Hagiu and Spulber (2013) and Etro (2021a). Etro (2021b) and Jeon and Rey (2021) investigate how the platform's monetization model affects its incentives to enter with first-party content and the incentives of third-party developers.

¹⁷Platforms like Amazon Marketplace obtain information on which products or product categories are particularly successful. Zhu and Liu (2018) provide empirical evidence that Amazon is more likely to enter as a first-party seller into more-successful product spaces.

¹⁸For some evidence in the mobile app market, see Wen and Zhu (2019).

¹⁹The exposition is based on Peitz (2022b) and partly uses material verbatim. For surveys on self-preferencing, see Kittaka, Sato, and Zennyo (2023) as well as Etro (2023b).

that are not in the best interest of consumers.²⁰ This is most easily seen when the platform does not charge users directly and only extracts some of the surplus generated by sellers. In the context of search engines, this was noted by Brin and Page in 1998: “we expect that advertising funded search engines will be inherently biased towards the advertisers and away from the needs of the consumers” (Brin & Page, 2012, p. 3832). Furthermore, sellers may differ in their ability to extract rents from consumers (who will be active on the platform in any case), and therefore a platform may favor those sellers that are better at extracting such a surplus. Similarly, some sellers may operate on different terms than others, which provides incentives to the platform to engage in biased recommendations. For example, Spotify is a platform that strongly affects consumers’ streaming behavior through its popular playlists, some of them algorithmic, others curated (Aguiar & Waldfogel, 2021). Aguiar, Waldfogel, and Waldfogel (2021) provide evidence that Spotify biases recommendations against major labels, which may be the response to the fact that major labels ask for higher royalties.

If a platform operates in dual mode, the platform internalizes the profits it makes from its vertically integrated activities and may engage in self-preferencing; that is, it steers consumers to first-party products or services when this is not in consumers’ best interest. For example, if a consumer could get a lower price for the same service quality from a third-party seller, then steering consumers toward a first-party product constitutes self-preferencing. The issue gets more complicated when products or services are differentiated, and consumers have different tastes about those products or services. For example, if some consumers have a strong taste for quick delivery, while others do not, it becomes difficult to assess when actual recommendations qualify as self-preferencing. Theoretical work has identified instances in which a platform with market power decides to engage in self-preferencing. For instance, a platform operating in dual mode may engage in self-preferencing to address the problem of bypass that otherwise limits the fees it can charge third parties. Hagiu et al. (2022) show that self-preferencing can then result in higher fees and consumer prices.²¹ Recent empirical work has identified instances of self-preferencing (broadly defined) and explored counterfactuals.²²

²⁰For work in the context of search engines, see Hagiu and Jullien (2011, 2014) as well as de Cornière and Taylor (2014). More broadly, see Heidhues et al. (2023), Lee (2013), and Peitz and Sobolev (forthcoming). For overviews that address the incentives of a platform regarding which recommendations to give, see Belleflamme and Peitz (2018b, 2021).

²¹Other theory contributions on self-preferencing include Bourreau and Gaudin (2022), de Cornière and Taylor (2014, 2019), Padilla et al. (2022), and Zennyo (2022).

²²Chen and Tsai (forthcoming) investigate Amazon’s recommendations through its “Frequently Bought Together” algorithm. Products are sold by Amazon as a retailer, by sellers as part of the “Fulfillment by Amazon” (FBA) program, and non-FBA sellers. The authors conclude that the steering via Amazon’s FBT algorithm is driven by seller identity rather than consumer preference. Lee and Musolff (2023) evaluate the effect of Amazon’s use of the buy box on consumer welfare using high-frequency data with the help of a structural model and find that the way Amazon preferentially treats first-party products increases consumer welfare because everything else given, consumers appear to prefer the product sold by Amazon instead of a third-party seller.

23.3 Platform Decisions with Cross-Platform Spillovers

Platforms may impose contractual obligations on some of the users or make decisions that directly affect competition between platforms or competition with an outside option. We look at three such practices—platform exclusivity, price-parity clauses, bundling and tying—and assess the competitive and consumer surplus effects of such practices.

23.3.1 Platform Exclusivity

A platform may want to make some sellers (or their products and services) exclusive.²³ In some environments, this may serve as a facilitating device and lead to higher prices. Yet, in other environments, there may be efficiencies associated with granting platform exclusivity. What is more, exclusivity may affect incumbent and entrant platforms differentially—it may serve as an entry deterrent or, conversely, benefit an entrant platform.

Platform exclusivity can be addressed in standard models of platform competition (such as the one by Armstrong, 2006). By imposing exclusivity agreements upon sellers, a platform can force them to singlehome.²⁴ A case in point is the so-called “radius clause,” whereby shopping malls prevent retail chains from opening another outlet in a competing shopping mall located within an agreed radius. Also, the ride-hailing companies Uber and Lyft have designed their application to make it difficult, if not impossible, for drivers to multihome (i.e., to compare ride offers from the two companies). However, third-party applications now exist that present offers from Uber and Lyft to drivers on a single screen, which facilitates multihoming.

If at least a fraction of sellers are exclusives, this has the potential to increase the differentiation between platforms in the eyes of the buyers, and this may reduce the pressure on prices for buyers, but it also affects the platform’s pricing incentives

With endogenous seller entry and exit, they find that the impact on consumer welfare is negligible but remains positive. Lam (2023) considers consumer searches that are guided by the platform’s decision on how to position different products in a product category and proposes a setting with heterogeneous consumers who sequentially search for differentiated products. Using Amazon data in the “Home & Kitchen” category, he estimates his model under the assumption that the ad valorem fee does not change. If Amazon’s position advantage is removed, profits are shifted from Amazon to third-party sellers. Such neutral positioning is shown to reduce the value of consumer search, and as a result, consumers are worse off after such an intervention. This means that Amazon’s steering incentives are aligned with consumers’ interests. Self-preferencing is also an issue in mobile app stores. Teng (2022) finds evidence of self-preferencing in Apple’s App Store. Self-preferencing affects consumer search and developers’ investment in app quality. In the counterfactual that removes self-preferencing, independent apps would increase investments. Overall, according to her estimates, such an intervention increases consumer and developer welfare.

²³ Exclusive content may serve as a substitute to first-party content, and the incentives to sign exclusivity contracts may depend on the presence of first-party content.

²⁴ To be precise, this holds under platform duopoly. If more than two platforms are active and one platform imposes exclusivity, this would not restrict users from multihoming on the other platforms.

regarding the sellers. Even if users care only about the number of users in the other group but not its composition, there are equilibrium effects to consider. Consider a setting with exogenously differentiated duopoly platforms in which buyers always singlehome, whereas sellers are forced to singlehome under exclusivity but can multihome otherwise. Without exclusivity, the environment has been called a competitive bottleneck (Armstrong, 2006) because platforms compete for buyers, whereas they operate as a monopolist on the seller side as they provide exclusive access to buyers. Under some conditions (but not always), this leads to a price structure that is favorable to buyers but unfavorable to sellers compared to the setting in which sellers must sign exclusivity contracts to be admitted to the platform.²⁵

Exclusive content may be offered by “large” content providers such that platforms bid for such exclusive content. Such a strategic content provider partly internalizes the impact of its own price on platform demand, and depending on the characteristics of the content, the content provider signs an exclusive agreement or multihomes (Hagiü & Lee, 2011). Focusing on a single strategic content provider, this content tends to be exclusive if platform competition is intense, as this allows the platform with the exclusive content to attract a large number of consumers (which implies that exclusivity does not sacrifice much of the network size) and the strategic content provider extracts surplus through an auction with a reserve price (Carroni et al., 2023).

To address the entry deterrence argument, consider an environment in which an incumbent platform can sign up some users upfront and, thus, deprive a more-efficient entrant platform of getting access to these users (Doganoglu & Wright, 2010). In the case of two-sided platforms, the incumbent platform can sign exclusivity contracts prior to entry. Here, the incumbent can divide the interests of sellers and consumers by offering attractive conditions to sellers such that they never have an incentive to reject the offer. Knowing this, consumers will join the incumbent subsequently. With homogeneous consumers, the incumbent platform extracts the full consumer surplus, and using exclusivity preserves the incumbent platform’s position. If the incumbent could not require exclusivity in the contract, sellers would be able to multihome, and the incumbent platform would not be in a position to profitably deter the entrant platform. Thus, exclusivity contracts deter a more-efficient potential entrant from entering.

On the contrary, exclusive content (which may be vertically integrated) may also work as an effective entry strategy; think of the decision by Disney to remove content on video streaming platforms such as Netflix and launch its own streaming platform, Disney+. Less recently, Lee (2013) analyzed exclusive content in the US video game industry (2000–2005). According to the estimates of his structural model, consumers would have benefitted if game platforms (console makers) had not been allowed to own or exclusively contract content, but entrant platforms would

²⁵Endogenizing the choice of exclusivity (in a setting with linear demand), whenever platforms benefit from imposing exclusivity, doing so may benefit or hurt sellers depending on the model parameters, but always hurts buyers (Belleflamme & Peitz, 2019b). Another important observation is that the use of exclusivity contracts in one group changes the incentives of users in the other group to become multihomers (Armstrong & Wright, 2007).

have been worse off because more high-quality content would have been available on the incumbent platform due to its larger installed base.

In contrast to platforms striving for exclusive content or services, platforms may make their use compatible and thereby remove platform-specific network effects. Compatibility is often costly and leads to higher prices. Symmetric firms then have a socially excessive interest in providing two-way compatibility (Doganoglu & Wright, 2006).²⁶

23.3.2 Price-Parity Clauses

Price-parity clauses stipulate that sellers on a platform cannot set higher retail prices on this platform than in a certain set of alternative sales channels.²⁷ This may include certain direct sales channels or other indirect sales channels provided by competing platforms. So-called wide price-parity clauses stipulate that sellers must not offer a lower price through any other channel (including direct and indirect channels), while narrow price-parity clauses stipulate that sellers must not offer a lower price in the direct sales channel but are allowed to set lower prices on other platforms. Wide price-parity clauses are often seen as anti-competitive, while there is substantial disagreement about the likely effects of narrow price-parity clauses.²⁸

Price-parity clauses have been imposed by several large digital platforms in the past. This includes hotel booking platforms such as Booking, which led to abuse cases in several jurisdictions in the 2010s. It also includes Amazon with its general pricing rule. Amazon addressed the sellers on its platform as follows: “you must always ensure that the item price and total price of an item you list on [Amazon.com](https://www.amazon.com) are at or below the item price and total price at which you offer and/or sell the item via any other online sales channel.” After the competition authorities had initiated investigations, Amazon removed price-parity clauses in Europe in 2013²⁹ but continued to impose the clause in the USA. In 2019, it then appeared to also remove the clause in the USA; however, the clause was replaced by a similar “fair pricing

²⁶Doganoglu and Wright (2006) also study the interaction between multihoming and compatibility and find that the former may be a poor substitute for the latter. For a recent contribution on the potential pitfalls of mandated interoperability, see Bourreau and Krämer (2023). Note that sometimes a single firm may decide to facilitate content becoming available on other platforms (one-way compatibility). A platform may thus decide to be horizontally open (e.g., by publishing its own interface specifications) and let its base of content or services be accessed from users attached to a competing platform (through converters); for a discussion, see Farrell and Simcoe (2012).

²⁷The exposition follows (mostly verbatim) Peitz (2022a).

²⁸Practitioners and academics often call price-parity clauses most-favored-customer clauses or “MFNs” (standing for most-favored-nation clauses), which can be seen as unfortunate and is possibly misleading. Most-favored-customer clauses traditionally stipulate that a seller cannot set different prices to different consumers or different prices over time. Price-parity clauses do not contain such restrictions but impose restrictions concerning prices faced by a given consumer across different distribution channels.

²⁹See press release of the Bundeskartellamt of November 26, 2013, “Amazon abandons price parity clauses for good” https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Meldungen%20News%20Karussell/26_11_2013_Amazon.html

policy.”³⁰ Yet another example is that Apple obliged publishers to set e-book prices in Apple’s iBookstore at the lowest retail price available in the market.

The basic argument as to why price-parity clauses are anticompetitive goes as follows. Consider a single platform that charges fees on the seller side and competes against the direct sales channel. If the platform obliges sellers on its platforms to not offer a lower price in the direct channel, consumers are not inclined to use the direct channel if the platform offers some convenience benefit. The platform will then set a high fee and extract a large fraction of seller profits. If price-parity clauses were prohibited, the platform’s fee setting would be constrained because the sellers would serve consumers at a low price in the direct channel if the fee were too high. This is a powerful argument against any price-parity clauses.

If there are competing platforms, the argument applies to wide price-parity clauses. Since sellers’ retail prices must be the same across the competing platforms under wide price parity, a seller cannot serve more consumers on a platform that lowers its fee. This reduces the incentive of a platform to offer a reduced fee. This means that wide price-parity clauses can be used as a facilitating device to soften platform competition. At the same time, consumers have little reason to try out new look-alike platforms, and thus, barriers to entry are higher with such clauses being in place.

One possible limitation of the above reasoning is that platform quality has been treated as exogenous. With price parity in place, platforms may have a strong incentive to increase the service quality offered to consumers to attract them to their platform. However, economic theory predicts that costly investments in service quality will be socially excessive. The net effect of price parity clauses on consumers is negative because the consumer surplus gain from higher service quality is more than offset by higher retail prices (Edelman & Wright, 2015).

Another qualification is that the above reasoning abstracted from the possibility that, absent price parity, consumers may use the platform to obtain valuable information, but with lower retail prices elsewhere, they will leave the platform and finalize the transaction elsewhere. Platforms would then receive no compensation for such showrooming services, which weakens their incentive to provide such a useful service to consumers. Price-parity clauses make seller free-riding unlikely since consumers cannot find lower prices elsewhere.³¹

³⁰In May 2021, the District of Columbia filed a complaint against Amazon at the Superior Court of the District of Columbia that contains more details on the contractual clauses imposed by Amazon.

³¹Absent price parity, consumers search on the platform and will not transact via the platform if the price differential between the price on the platform and the price on the direct distribution channel exceeds the convenience benefit from transacting on the platform. Sellers may want to set low prices in the direct channel that induce consumers to switch. This constrains the platform’s fee setting since the platform will want to avoid free-riding. As shown by Wang and Wright (2020), when price parity clauses are prohibited, consumers are better off if the platform remains viable. With competing platforms and showrooming, wide price parity clauses continue to decrease consumer welfare, while results regarding narrow price parity clauses are less clear-cut: if narrow price parity is needed for the viability of platforms and platform competition is sufficiently intense, narrow price parity clauses are in the interest of consumers (Wang & Wright, 2020). Even in the case of a monopoly platform, price parity can be profitable and, at the same time, increase consumer welfare (see Liu et al., 2021; Peitz & Sobolev, 2023).

23.3.3 Bundling and Tying

Examples of bundled offers are cable TV bundles as well as subscription services by streaming platforms (such as Netflix). Amazon with its Prime membership is another example of a bundling strategy. Bundling is particularly attractive for a firm offering digital products since the marginal cost is typically negligible. Thus, bundling can be used for price discrimination purposes but also as a way to offer multiple products in a more convenient way to consumers. However, bundling may not be in the interest of consumers when it is used as a facilitating device or as a deterrence device. What is more, it may be used to leverage market power. While bundling is a common practice, it has some distinguishing features in the context of platforms.

Amelio and Jullien (2012) point to the fact that bundling can relax the zero-price constraint that applies if the platform cannot subsidize a user group. Suppose that there is a monopoly platform that caters to two user groups that are connected through cross-group network effects. Furthermore, suppose that in this setting the zero-price constraint is binding such that the monopoly platform would find it profitable, but is not allowed, to subsidize one of the user groups. Consider the option of the platform to sell another product that generates positive gains from trade for one of the two user groups and users in this group have the same willingness to pay for this product. Through bundling, the platform can then relax the zero-price constraint and implicitly make a subsidy. When the platform sells the bundle and the second product separately, in the profit-maximizing outcome, the platform makes strictly higher profits than it would absent bundling and both user groups are better off (Amelio & Jullien, 2012).

Network effects can lead to anticompetitive bundling or tying. In Choi, Jeon, and Whinston (2023), a firm is a monopolist in the primary market (where consumers have heterogeneous valuations for this product) and competes against a competitor in a second market in which consumers experience positive direct network effects. Under independent pricing, the firm would set the monopoly price in the primary market, and consumers would receive the consumer surplus associated with monopoly pricing. As Choi, Jeon, and Whinston (2023) explain, when the firm bundles its two products, consumers with high valuations in the primary market may continue to purchase the bundle even if other consumers were to buy from the competitor in the second market. The existence of such high-valuation consumers guarantees a minimum market share for the firm in the second market. Because of network effects in the second market, this installed-base advantage may induce low-valuation consumers to buy the bundle. This may lead to tipping in the second market in favor of the firm offering the bundle even though the competitor is more efficient in the second market.³²

³² Fumagalli and Motta (2020) also consider tying between a primary market in which an incumbent firm starts as a monopolist and a complementary market. The incumbent firm is willing to sacrifice current profits when tying in order to exclude a more efficient rival from a complementary market by depriving it of the critical user size that it needs to be successful. This leads to a favorable position for the incumbent when a more-efficient rival enters the primary market and allows it to extract part of the rival's efficiency rents. In this argument, the presence of non-negative price constraints is crucial for exclusion.

Choi and Jeon (2021) show that bundling can be anticompetitive when firms operate as platforms and cannot set negative prices to consumers. They consider the interplay between two markets, one monopoly product market and another competitive product market that is ad-funded and served by the firm operating as a monopolist in the first market as well as a more-efficient competitor.³³ Since firms cannot subsidize consumers, the more-efficient competitor is constrained in its ability to offer a better deal to consumers in response to the offer by the firm that is less efficient in providing the second service.³⁴ The point is that because of advertising opportunities, there is a positive surplus on the table in the market for the second service. If the more-efficient competitor lacks instruments to offer a higher surplus to consumers, it is vulnerable to losing out to the less-efficient firm that can offer a bundle.

23.4 Public Regulation

Legislators may endow regulators with powers to restrict the behavior of (certain types of) digital platforms or intervene directly. In addition, general competition law may provide the basis for intervention by competition authorities if digital platforms engage in anticompetitive conduct or if platforms engage in anticompetitive merger activities. Also on the table may be the possibility for competition authorities or regulators to impose structural measures such as forced divestitures. Apart from the regulation of digital platforms and general competition law, other areas of laws may be applicable as well: tort law, consumer protection law (including privacy), intellectual property law (counterfeit, copyright violations), laws against unfair trade practices, and telecom and media regulation. The EU legislator not only grants intervention possibilities under general competition law (articles 101 and 102 TFEU and competencies in merger control) but also introduced specific regulations that address several concerns (that are to be enforced by the EU member states or EU institutions). Most notable are the platform-to-business regulation (P2B regulation), the package of Digital Services Act (DSA), and Digital Markets Act (DMA) that came into force in 2022. The P2B regulation establishes rules to be

³³ For simplicity, assume that consumers do not mind advertising combined with content from this second category. The competitor is more efficient in the sense that it offers higher service quality to buyers and incurs the same cost.

³⁴ The firm that operates as a monopolist in one market can offer a better deal to consumers for the second type of content because it can offer the bundle of both types of content at a lower price. It also has the incentive to do so if it is not too much at a disadvantage compared to its more-efficient competitor. Since consumers would like to have both services, they choose the bundle if the bundled price is not too high (assuming that consumers singlehome). In return, the firm offering the bundle attracts all consumers and can monopolize the advertising market that comes with the second service. If access to the two types of content were sold separately, this would not be an issue, and the monopoly intermediary would sell the first service at the monopoly price, while the more-efficient competitor for the second service would sell in the other market at a price that reflects consumers' willingness to pay for higher content quality and, in addition, make positive ad revenues.

followed by digital platforms in their dealings with smaller businesses and sellers; the DSA, which updates the e-commerce Directive from the year 2000, imposes differential obligations on platforms regarding illegal content, transparent advertising, and disinformation; the DMA imposes obligations and prohibitions on a few “gatekeeper platforms” regarding their “core platform services” (e.g., Google Search and Google Maps).

Possible interventions by legislators, competition authorities, and regulatory authorities include forced divestitures and the prohibition of a proposed merger, prohibition of certain price or non-price strategies (or the obligation imposed on platforms to take certain actions), fines for violations, and liability rules.

23.4.1 Merger Control

From an economic point of view, forced divestitures are the mirror image of blocking a merger, as they dissolve an integrated firm, while a merger would generate an integrated firm. Nevertheless, competition law treats these two types of interventions quite differently since a forced divestiture is an intervention that happens *ex post* and is more drastic, insofar as it may force the firm to make costly adjustments. Digital platforms have been very active in merger activities—this applies to Google, Amazon, Facebook, Apple, and Microsoft (GAFAM) but also to other digital platforms.³⁵

In the presence of network effects, an increase in market concentration may be beneficial for consumers. At the same time, a merger makes market tipping more likely, and monopolization is often not beneficial for consumers. Horizontal mergers between multi-sided platforms tend to be difficult to evaluate, which makes it harder to obtain clear-cut theoretical predictions and empirical results.³⁶

Regarding the role of network effects, if the consumer base is important up to a certain level, the most profitable merger may lead to consumer harm according to the following argument: When firms are able to combine the installed consumer base through a horizontal merger and there are several firms willing to bid for a takeover target, the acquisition by the highest bidder may lead to a worse outcome from a consumer welfare perspective than if the acquisition were made by a firm with a lower bid. The reason is that the firm with the lower bid would achieve critical mass with the merger, while the firm with the higher bid already has it, but deprives the competitor of it (Motta & Peitz, 2021).³⁷

³⁵The following exposition draws partly verbatim from Peitz (2023).

³⁶The Cournot approach developed by Correia-da-Silva et al. (2019) may be useful in this respect. It is more tractable than models with price setting because, by setting participation or usage levels, platforms directly control the network effects.

³⁷The argument also applies when merging firms hoard certain assets or capabilities and these assets or capabilities are scarce overall.

Horizontal mergers may be anticompetitive if more-efficient competitors are acquired quickly and there is competition for the market in the sense that the entering more-efficient competitor would attract all unaffiliated consumers. Following Katz (2021), one way to think about this is that loyal consumers stay with the incumbent firm until they retire from the market, while flexible consumers go for the better offer. A potential entrant must then be confident that it would make the better offer to have an incentive to enter. After entry, it will first attract flexible consumers outcompeting the incumbent, while the incumbent stays on for a while as long as it can profitably sell to the loyal consumers. Eventually, the incumbent leaves the market, and the former entrant becomes the new incumbent, which will enjoy monopoly profits as long as no other firm enters. When the incumbent is still around, it constrains the entrant's pricing power. When the innovation process is exogenous, a merger between the incumbent and the firm that is about to enter removes the competitive constraint, which is harmful to consumers and society.

The expansion of an ecosystem through vertical or conglomerate mergers constitutes an envelopment strategy (Eisenmann et al., 2011). One concern may be that a vertical or conglomerate merger might diminish competition or raise other competition concerns within an ecosystem. However, if this ecosystem is relatively small and engaged in competition with a more popular ecosystem, the merger could potentially bolster the position of this ecosystem relative to its competitors, thus fostering competition.³⁸

This argument can be reversed when the ecosystem under consideration already holds a strong and possibly entrenched position. In such cases, further strengthening the ecosystem might reduce overall competition between ecosystems. This may be of particular relevance in cases (in which the merger expands an ecosystem or the platform's control thereof) such that the platform gains a data advantage. This data advantage could also make it more challenging for outsiders to challenge the ecosystem. Relatedly, a multi-sided platform that monetizes user data may be able and have an incentive to envelop another activity by tying its privacy policies.³⁹

23.4.2 "Regulating" Conduct Under Competition Law or Specific Regulation

A competition authority may investigate a particular practice by a firm to be an abuse of its market power.⁴⁰ The practice may be prohibited if it is anticompetitive. Over the years, several digital platforms have been subject to antitrust scrutiny, and the

³⁸ This argument was made by Compass Lexecon consulting for the acquiring party in the NVIDIA/Arm merger and led to the formal analysis in Bisceglia et al. (2022).

³⁹ Such a strategy relies on overlapping users and monetization of user data also for the activity that is subject to envelopment (Condorelli & Padilla, 2020, 2024).

⁴⁰ Competition law often requires market definition and a market power assessment as the starting point to evaluate a certain practice. These can be particularly challenging in a platform context (Katz & Sallet, 2018; Franck & Peitz, 2021a, 2023).

underlying theory of harm put forward by the authority has in some cases been based on certain features of the digital platform. Also, as explained in the previous two sections, academic economists have developed novel theories of harm in platform settings.

An alternative to the application of general competition law is sector-specific regulation targeted toward certain platforms. This may take the form of a separate regulation (as, in the EU, for telco operators, which provide communication platforms, and, with the DMA, for certain digital platforms). Instead, general competition law may be supplemented by specific provisions that apply only to certain digital platforms.⁴¹

Telco regulation may contain elements that restrict the practices of internet service providers (ISP) as platforms. For example, net neutrality obligations restrict the price and non-price strategy of ISPs regarding digital content. Regarding pricing, the ISP may want to monetize the consumer and content provider side by charging a subscription fee to consumers and a termination fee to content providers for the delivery of their content. Net neutrality regulation can rule out such two-sided pricing and require that (i) the ISP provide the same service to all content providers and users and (ii) only consumers but not content providers be allowed to be charged—Greenstein, Peitz, and Valletti (2016) provide a guide to the net neutrality debate.

The DMA contains a list of prohibitions and obligations (in Articles 5 to 7) applied to designated “gatekeeper platforms” in relation to their “core platform services.” In addition, it contains an anticircumvention prohibition (Article 13) and the possibility to add further obligations and prohibitions in the future. The DMA is enforced by the European Commission. Among the prohibitions and obligations, figure the prohibition of the use of any price-parity clauses,⁴² self-preferencing, and bundling. The DMA also contains interoperability requirements and provisions that address data imbalances between gatekeeper platform and sellers of third-party services. Regarding the former, Article 7 is noteworthy as it provides very detailed regulation with respect to interoperability of messaging services (“number-independent interpersonal communication services”) at the request of competitors to those companies for which messaging is a core platform service.

23.4.3 Public Governance, Tort Law, and Other Regulations

Digital platforms may be subject to regulations that are motivated by concerns outside the competition realm. Social media platforms may be subjected to content moderation and be required to remove hate speech; this may include active

⁴¹ As in Germany with Section 19a of the German Competition Act; see Franck and Peitz (2021b).

⁴² Art. 5(3) DMA states: “The gatekeeper shall not prevent business users from offering the same products or services to end users through third-party online intermediation services or through their own direct online sales channel at prices or conditions that are different from those offered through the online intermediation services of the gatekeeper.”

measures. Certain platforms may be required to take particular measures to protect consumers.

Platforms may be held liable for illegal material. This includes child abuse content, terrorism content, and hate speech on social media platforms such as Facebook and Twitter/X, copyright violations on media platforms such as YouTube, and defective, dangerous, counterfeit or otherwise illegal products on e-commerce platforms such as Amazon. Liability may also apply to news aggregators, search engines, and app stores for providing links to illegal content or websites that pose security risks. Ad-based platforms may be held liable for including misleading advertising. Platforms holding personal data may be held liable for the misuse of data by third parties. Regarding their own operations, platforms such as Uber and TaskRabbit may also be held liable for discrimination on the platform.

Platform liability has traditionally been very limited. For instance, according to the EU's eCommerce Directive that was adopted in 2000, hosting platforms are exempted from liability for hosting illegal material in the European Union (EU) provided that they remove illegal material expeditiously upon obtaining knowledge of it.⁴³

To conceptually address the issue, it is useful to consider “good” and “bad” sellers as well as consumers who may be negatively affected by “bad” sellers. A platform may engage in active measures to detect and remove bad sellers from the platform; its incentives are affected by the liability regime. Beneficiaries of such active measures may be “good” sellers (e.g., brand manufacturers who will be better protected from counterfeit products or copyright holders who will be protected from illegal copies), consumers (e.g., in the case of the removal of defective or dangerous products), and society at large (e.g., after the removal of terrorist content or malware or after removing sellers of exotic animals whose trade is forbidden). A profit-maximizing platform will partly internalize the benefits of those users whose behavior affects revenues. Thus, the platform may have some incentive to combat bad sellers even in the absence of liability, as consumers may not join the platform when anticipating a bad experience. Platform liability strengthens those incentives.⁴⁴

Regulation on data privacy and data security applies to firms holding personal data; in the EU, the General Data Protection Regulation (GDPR) sets out how firms can store, process, and use personal data. While this regulation is not targeted toward digital platforms, they must comply. Digital platforms may also impose rules on third

⁴³ Articles 12 to 15 of Directive 2000/31/EC. For details on the directive, see, e.g., Buiten, de Streef, and Peitz (2020). The key rules and principles continue to apply under the Digital Services Act (DSA).

⁴⁴ De Chiara et al. (2022) consider harm to good sellers (copyright holders) and the incentives of the hosting platform to remove infringers; Jeon, Lefouili, and Madio (2021) endogenize the incentives of good sellers to invest in quality (brand manufacturers suffering from counterfeits). Zennyo (2023) endogenizes the incentives of sellers to maintain quality; see also Yasui (2022). Hua and Spier (2023) investigate a platform's monitoring effort and pricing strategy in response to changes of the liability regime and uncover which market environments call for stricter platform liability and which call for weaker platform liability.

parties regarding the use of personal data and information disclosure thereof to individuals who provide those data. For example, Apple requires information about the app's privacy practice when a developer submits a new app or an app update to the App Store.⁴⁵ This suggests that digital platforms may react to regulation (or anticipate regulatory intervention) or use existing or imminent regulation as an excuse for their updated platform design, and such updates may be socially harmful.

23.5 Conclusion

In digital ecosystems, some firms (typically dominant platforms) operate as private regulators (Boudreau & Hagiu, 2009), which are possibly constrained by public regulation. Public interventions may try to avoid exploitative and exclusionary abuse by these firms. They may also aim at making or keeping the market contestable such that dominant platforms can be successfully challenged. What is more, regulation may aim at addressing externalities to bystanders including society at large.

Existing or proposed changes in competition law and regulations may or may not be well-suited to deal with the dynamic consequences in platform markets characterized by scale economies, network effects, multi-sidedness, and other features. Moreover, the technologies that platforms use and develop evolve very rapidly. This means that questions need to be answered based on a good understanding of the prevailing market characteristics and the institutional features and limitations of regulatory authorities. Economic analysis can provide valuable input to regulatory proposals and their implementation and, as this chapter has documented, has done so with the use of theory and empirical analysis.

As explained in this chapter, digital platforms engage in many activities that aim to solve or mitigate market failures. Regulation must be careful not to reduce the ability of current or future platforms to perform this task. However, firms with market power may not perform the task in the socially optimal way. For example, participation by one or multiple user groups may be socially insufficient because of high prices or low quality. Also, a platform may underinvest in the screening of faulty products. What is more, platform governance rules aimed at facilitating interaction may have negative knock-on effects on users (unfair treatment, invasion of privacy, etc.).

With these insights in mind, one can, for instance, ask to which extent divestiture obligations imposed on a platform would reduce the associated network benefits that platform users enjoy. Would future platforms invest in bringing economic agents together if they anticipate that, once they have succeeded, they will be forced to share the benefits with competing platforms through some form of interoperability? Such a question can be asked in the case of remedies as foreseen in the EU Digital

⁴⁵ See <https://developer.apple.com/app-store/app-privacy-details/>, last accessed November 13, 2023.

Markets Act, such as interoperability or data-sharing obligations. There is a need for more research addressing these questions, and such efforts are needed urgently because regulators have the mandate to act and cannot just sit back and wait for relevant research to emerge.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

Bibliography

- Aguiar, L., & Waldfogel, J. (2021). Platforms, power, and promotion: Evidence from Spotify playlists. *Journal of Industrial Economics*, 69, 653–691.
- Aguiar, L., Waldfogel, J., & Waldfogel, S. (2021). Playlisting favorites: Measuring platform bias in the music industry. *International Journal of Industrial Organization*, 78, 102765.
- Ali, S. N., Lewis, G., & Vasserman, S. (2023). Voluntary disclosure and personalized pricing. *Review of Economic Studies*, 90, 538–571.
- Amelio, A., & Jullien, B. (2012). Tying and freebies in two-sided markets. *International Journal of Industrial Organization*, 30, 436–446.
- Anderson, S. P., & Bedre-Defolie, Ö. (forthcoming). Hybrid platform model: Monopolistic competition and a dominant firm. *RAND Journal of Economics*.
- Anderson, S. P., & Coate, S. (2005). Market provision of broadcasting: A welfare analysis. *Review of Economic Studies*, 72, 947–972.
- Anderson, S. P., & Peitz, M. (2020). Media see-saws: Winners and losers in platform markets. *Journal of Economic Theory*, 186, 104990.
- Armstrong, M. (2006). Competition in two-sided markets. *RAND Journal of Economics*, 37, 668–691.
- Armstrong, M., & Wright, J. (2007). Two-sided markets, competitive bottlenecks and exclusive contracts. *Economic Theory*, 32, 353–380.
- Ater, I. (2015). Vertical foreclosure using exclusivity clauses: Evidence from shopping malls. *Journal of Economics and Management Strategy*, 24, 620–642.
- Bakos, Y., & Halaburda, H. (2020). Platform competition with multihoming on both sides: Subsidize or not? *Management Science*, 66, 5599–5607.
- Belleflamme, P., & Peitz, M. (2018a). Platforms and network effects. In L. C. Corchón & M. A. Marini (Eds.), *Handbook of game theory and industrial organization* (Vol. II). Edward Elgar Publisher.
- Belleflamme, P., & Peitz, M. (2018b). Inside the engine room of platforms: Reviews, ratings, and recommendations. In J.-J. Ganuza & G. Llobet (Eds.), *Economic analysis of the digital revolution* (Funcas Social and Economic Studies, Vol. 4). Funcas.
- Belleflamme, P., & Peitz, M. (2019a). Managing competition on a platform. *Journal of Economics and Management Strategy*, 28, 5–22.
- Belleflamme, P., & Peitz, M. (2019b). Platform competition: Who benefits from multihoming? *International Journal of Industrial Organization*, 64, 1–26.
- Belleflamme, P., & Peitz, M. (2019c). Price disclosure by two-sided platforms. *International Journal of Industrial Organization*, 67, 102529.
- Belleflamme, P., & Peitz, M. (2021). *The economics of platforms: Concepts and strategy*. Cambridge University Press.
- Belleflamme, P., & Peitz, M. (forthcoming). Network goods, price discrimination, and two-sided platforms. *Journal of Institutional and Theoretical Economics*.
- Belleflamme, P., & Toulemonde, E. (2009). Negative intra-group externalities in two-sided markets. *International Economic Review*, 50, 245–272.

- Biscegla, M., Padilla, J., Piccolo, S., & Shekhar, S. (2022). Vertical integration, innovation and foreclosure with competing ecosystems. *Information Economics and Policy*, 60, 100981.
- Boudreau, K., & Hagiu, A. (2009). Platforms rules: Multi-sided platforms as regulators. In A. Gawer (Ed.), *Platforms, markets and innovation*. Edward Elgar Publisher.
- Bourreau, M., & Gaudin, G. (2022). Streaming platform and strategic recommendation bias. *Journal of Economics and Management Strategy*, 31, 25–47.
- Bourreau, M., & Krämer, J. (2023). *Interoperability in digital markets: Boon or bane for market contestability?* Unpublished manuscript.
- Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 56, 3825–3833.
- Buiten, M. C., de Streel, A., & Peitz, M. (2020). Rethinking liability rules for online hosting platforms. *International Journal of Law and Information Technology*, 28, 139–166.
- Caillaud, B., & Jullien, B. (2003). Chicken & egg: Competition among intermediation service providers. *RAND Journal of Economics*, 34, 309–328.
- Carroni, E., Madio, L., & Shekhar, S. (2023). Superstar exclusivity in two-sided markets. *Management Science*, 70, 991.
- Casner, B. (2020). Seller curation in platforms. *International Journal of Industrial Organization*, 72, 102659.
- Chen, N., & Tsai, H.-T. (forthcoming). Steering via algorithmic recommendations. *Rand Journal of Economics*.
- Choi, J. P., & Jeon, D.-S. (2021). A leverage theory of tying in two-sided markets with non-negative price constraints. *American Economic Journal: Microeconomics*, 13, 283–337.
- Choi, J. P., & Jeon, D.-S. (2023). Platform design biases in ad-funded two-sided markets. *RAND Journal of Economics*, 54, 240–267.
- Choi, J. P., Jeon, D.-S., & Whinston, M. D.. (2023). *Tying with network effects*. Unpublished manuscript.
- Condorelli, D., & Padilla, J. (2020). Harnessing platform envelopment in the digital world. *Journal of Competition Law & Economics*, 16, 143–187.
- Condorelli, D., & Padilla, J. (2024). Data-driven envelopment with privacy-policy tying. *Economic Journal*, 134, 515–537.
- Correia-da-Silva, J., Jullien, B., Lefouili, Y., & Pinho, J. (2019). Horizontal mergers between multisided platforms: Insights from Cournot competition. *Journal of Economics and Management Strategy*, 28, 109–124.
- Crawford, G., Courthood, M., Seibel, R., & Zuzek, S. (2022). *Amazon entry on Amazon Marketplace* (CEPR Discussion Paper DP17531). CEPR.
- De Chiara, A., Manna, E., Rubi-Puig, A., & Segura-Moreiras, A. (2022). *Efficient copyright filters for online hosting platforms* (UB Economics Working Paper 433). University of Barcelona School of Economics.
- de Cornière, A. (2016). Search advertising. *American Economic Journal: Microeconomics*, 8, 156–188.
- de Cornière, A., & Taylor, G. (2014). Integration and search engine bias. *RAND Journal of Economics*, 45, 576–597.
- de Cornière, A., & Taylor, G. (2019). A model of biased intermediation. *RAND Journal of Economics*, 50, 854–882.
- de Cornière, A., & Taylor, G. (2020). *Data and competition: A general framework with applications to mergers, market structure, and privacy policy* (TSE Working Paper 20-1076). Toulouse School of Economics.
- Dinerstein, M., Einav, L., Levin, J., & Sundaresan, N. (2018). Consumer price search and platform design in internet commerce. *American Economic Review*, 108, 1820–1859.
- Doganoglu, T., & Wright, J. (2006). Multihoming and compatibility. *International Journal of Industrial Organization*, 24, 45–67.
- Doganoglu, T., & Wright, J. (2010). Exclusive dealing with network effects. *International Journal of Industrial Organization*, 28, 145–154.
- Edelman, B., & Wright, J. (2015). Price coherence and excessive intermediation. *Quarterly Journal of Economics*, 130, 1283–1328.

- Eisenmann, T., Parker, G., & Van Alstyne, M. (2011). Platform envelopment. *Strategic Management Journal*, 32, 1270–1285.
- Etro, F. (2021a). Product selection in online marketplaces. *Journal of Economics and Management Strategy*, 30, 1–25.
- Etro, F. (2021b). Device-funded vs ad-funded platforms. *International Journal of Industrial Organization*, 75, 102711.
- Etro, F. (2023a). Platform competition with free entry of sellers. *International Journal of Industrial Organization*, 89, 102903.
- Etro, F. (2023b). E-commerce platforms and self-preferencing. *Journal of Economic Surveys*, 38, 1516.
- Farrell, J., & Simcoe, T. (2012). Four paths to compatibility. In M. Peitz & J. Waldfogel (Eds.), *Oxford handbook of the digital economy*. Oxford University Press.
- Franck, J.-U., & Peitz, M. (2021a). Market definition in the platform economy. *Cambridge Yearbook of European Legal Studies (CYELS)*, 23, 91–127.
- Franck, J.-U., & Peitz, M. (2021b). Digital platforms and the new 19a tool in the German Competition Act. *Journal of European Competition Law & Practice*, 12, 513–528.
- Franck, J.-U., & Peitz, M. (2023). Market power of digital platforms. *Oxford Review of Economic Policy*, 39, 34–46.
- Fumagalli, C., & Motta, M. (2020). Tying in evolving industries when future entry cannot be deterred. *International Journal of Industrial Organization*, 73, 102567.
- Gambato, J., & Peitz, M. (2023). *Platform-enabled information disclosure* (CRC TR 224 Discussion Paper 468). University of Bonn and University of Mannheim.
- Greenstein, S., Peitz, M., & Valletti, T. (2016). Net neutrality: A fast lane to understanding the trade-offs. *Journal of Economic Perspectives*, 30, 127–149.
- Hagiu, A. (2009). Two-sided platforms: Product variety and pricing structures. *Journal of Economics and Management Strategy*, 18, 1011–1043.
- Hagiu, A., & Halaburda, H. (2014). Information and two-sided platform profits. *International Journal of Industrial Organization*, 34, 25–35.
- Hagiu, A., & Jullien, B. (2011). Why do intermediaries divert search? *RAND Journal of Economics*, 42, 337–362.
- Hagiu, A., & Jullien, B. (2014). Search diversion and platform competition. *International Journal of Industrial Organization*, 33, 48–60.
- Hagiu, A., & Lee, R. S. (2011). Exclusivity and control. *Journal of Economics and Management Strategy*, 20, 679–708.
- Hagiu, A., & Spulber, D. (2013). First-party content and coordination in two-sided markets. *Management Science*, 59, 933–949.
- Hagiu, A., & Wright, J. (2023). Data-enabled learning, network effects and competitive advantage. *RAND Journal of Economics*, 54, 638–667.
- Hagiu, A., & Wright, J. (2024). Marketplace leakage. *Management Science*, 70, 1529–1553.
- Hagiu, A., Teh, T.-H., & Wright, J. (2022). Should platforms be allowed to sell on their own marketplaces? *RAND Journal of Economics*, 53, 297–327.
- Hart, O., & Tirole, J. (1990). Vertical integration and market foreclosure. *Brookings Papers on Economic Activity: Microeconomics*, 205–286.
- Heidhues, P., Köster, M., & Köszegi, B. (2023). Steering fallible consumers. *Economic Journal*, 133, 1420–1465.
- Hua, X., & Spier, K. (2023). *Holding platforms liable*. Unpublished manuscript.
- Hunold, M., Kesler, R., & Laitenberger, U. (2020). Rankings of online travel agents, channel pricing, and consumer protection. *Marketing Science*, 39, 92–116.
- Jeon, D.-S., & Rey, P. (2021). *Platform competition, ad valorem commissions and app development*. Unpublished manuscript, Toulouse School of Economics.
- Jeon, D.-S., Lefouili, Y., & Madio, L. (2021). *Platform liability and innovation*. Unpublished manuscript.
- Jiang, B., Jerath, K., & Srinivasan, K. (2011). Firm strategies in the “mid tail” of platform-based retailing. *Marketing Science*, 30, 757–775.

- Jin, G. Z., Lu, Z., Zhou, X., & Lu, F. (2021). *Flagship entry in online marketplaces* (NBER Working Paper w29239). NBER.
- Johnen, J., & Somogyi, R. (2022). *Deceptive features on platforms* (LIDAM Discussion Paper CORE 2022/19). Université Catholique de Louvain, Center for Operations Research and Econometrics.
- Johnson, J., Rhodes, A., & Wildenbeest, M. (2023). Platform design when sellers use pricing algorithms. *Econometrica*, 91, 1841–1879.
- Jullien, B., & Sand-Zantman, W. (2021). The economics of platforms: A theory guide for competition policy. *Information Economics and Policy*, 54, 100880.
- Jullien, B., Pavan, A., & Rysman, M. (2022). Two-sided markets, pricing, and network effects. In K. Ho, A. Hortaçsu, & A. Lizzeri (Eds.), *Handbook of industrial organization* (Vol. 4). Elsevier.
- Karle, H., & Peitz, M. (2017). De-targeting: Advertising an assortment of products to loss-averse consumers. *European Economic Review*, 95, 103–124.
- Karle, H., Peitz, M., & Reisinger, M. (2020). Segmentation versus agglomeration: Competition between platforms with competitive sellers. *Journal of Political Economy*, 128, 2329–2374.
- Katz, M. L. (2021). Big tech mergers: Innovation, competition for the market, and the acquisition of emerging competitors. *Information Economics and Policy*, 54, 100883.
- Katz, M. L., & Sallet, J. (2018). Multisided platforms and antitrust enforcement. *Yale Law Journal*, 127, 2142–2175.
- Kittaka, Y., Sato, S., & Zenryo, Y. (2023). Self-preferencing by platforms: A literature review. *Japan and the World Economy*, 66, 101191.
- Lam, H. T. (2023). *Platform search design and market power*. Unpublished manuscript.
- Lee, R. S. (2013). Vertical integration and exclusivity in platform and two-sided markets. *American Economic Review*, 103, 2960–3000.
- Lee, K. H., & Musolf, L. (2023). *Entry into two-sided markets shaped by platform-guided search*. Unpublished manuscript.
- Liu, C., Niu, F., & White, A. (2021). *Optional intermediaries and pricing restraints*. Unpublished manuscript.
- Madio, L., & Quinn, M. (2023). *Content moderation and advertising in social media platforms*. Unpublished manuscript.
- Madsen, E., & Vellodi, N. (forthcoming). Insider imitation. *Journal of Political Economy*.
- McIntyre, D., Srinivasan, A., Afuah, A., Gawer, A., & Kretschmer, T. (2021). Multisided platforms as new organizational forms. *Academy of Management Perspectives*, 35, 566–583.
- Motta, M., & Peitz, M. (2021). Big tech mergers. *Information Economics and Policy*, 54, 100868.
- Nocke, V., Peitz, M., & Stahl, K. (2007). Platform ownership. *Journal of the European Economic Association*, 5, 1130–1160.
- Padilla, J., Perkins, J., & Piccolo, S. (2022). Self-preferencing in markets with vertically integrated gatekeeper platforms. *Journal of Industrial Economics*, 70, 371–395.
- Pashigian, B. P., & Gould, E. D. (1998). Internalizing externalities: The pricing of space in shopping malls. *Journal of Law and Economics*, 41, 115–142.
- Peitz, M. (2022a). The prohibition of price-parity clauses and the digital markets act. *TechREG Chronicle*, Competition Policy International.
- Peitz, M. (2022b). *The prohibition of self-preferencing in the DMA*. CERRE Issue Paper. https://cerre.eu/wp-content/uploads/2022/11/DMA_SelfPreferencing.pdf
- Peitz, M. (2023). Mergers in big tech: Recent developments in EU and national case law. *Concurrentes*. e-Competitions Special Issue Mergers in Big Tech.
- Peitz, M., & Reisinger, M. (2016). Media economics of the internet. In S. P. Anderson, D. Stromberg, & J. Waldfogel (Eds.), *Handbook of media economics* (Vol. 1A, pp. 445–530). Elsevier.
- Peitz, M., & Sato, S. (2023). *Asymmetric platform oligopoly* (CRC TR 224 Discussion Paper 428). University of Bonn and University of Mannheim.
- Peitz, M., & Sobolev, A. (2023). *Product recommendations and price parity clauses*. Unpublished manuscript.

- Peitz, M., & Sobolev, A. (forthcoming) Inflated recommendations. *Rand Journal of Economics*.
- Rochet, J.-C., & Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1, 990–1029.
- Rochet, J.-C., & Tirole, J. (2006). Two-sided markets: A progress report. *RAND Journal of Economics*, 37, 645–667.
- Tan, H., & Wright, J. (2018). A price theory of multi-sided platforms: Comment. *American Economic Review*, 108, 2758–2760.
- Tan, H., & Wright, J. (2021). Pricing distortions in multi-sided platforms. *International Journal of Industrial Organization*, 79, 102732.
- Tan, G., & Zhou, J. (2021). The effects of competition and entry in multi-sided markets. *Review of Economic Studies*, 88, 1002–1030.
- Teh, T.-H. (2022). Platform governance. *American Economic Journal: Microeconomics*, 14, 213–254.
- Teh, T.-H., Liu, C., Wright, J., & Zhou, J. (2023). Multihoming and oligopolistic platform competition. *American Economic Journal: Microeconomics*, 15, 68–113.
- Teng, X. (2022). *Self-preferencing, quality provision, and welfare in mobile application markets* (CESifo Working Paper 10042). CESifo.
- Wang, C., & Wright, J. (2020). Search platforms: Showrooming and price parity clauses. *RAND Journal of Economics*, 51, 32–58.
- Wen, W., & Zhu, F. (2019). Threat of platform-owner entry and complementor responses: Evidence from the mobile app market. *Strategic Management Journal*, 40, 1336–1367.
- Weyl, E. G. (2010). A price theory of multi-sided platforms. *American Economic Review*, 100, 1642–1672.
- Yasui, Y. (2022). *Platform liability for third-party defective products*. Unpublished manuscript.
- Zenno, Y. (2022). Platform encroachment and own-content bias. *Journal of Industrial Economics*, 70, 684–710.
- Zenno, Y. (2023). *Should platforms be held liable for defective third-party goods?* Unpublished manuscript.
- Zhu, F., & Liu, Q. (2018). Competing with complementors: An empirical look at Amazon.com. *Strategic Management Journal*, 39, 2618–2642.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part V

Contracts

The Empirics of Relational Contracts: The Good, the Bad, and the Ugly

24

Rocco Macchiavello

Contents

24.1	Introduction	598
24.2	Methodological Considerations	600
24.3	The Good	602
24.4	The Bad	610
24.5	The Ugly	613
24.6	Conclusions	617
	References	618

Abstract

This chapter reviews selected contributions within a recent body of empirical studies of relational contracts. Existing reviews have almost exclusively emphasized how relational contracts benefit participating parties: given parties' inability to write complete contracts and courts' inability to enforce those, relational contracting allows for the realization of gains from trade that would otherwise be foregone. This chapter argues that the empirical study of relational contracts should be extended to consider contexts in which such arrangements support undesirable transactions. After a brief discussion of approaches to empirically study relational contracts, this review thus considers three broad contexts of applications of relational contracts: *the good*, where relational contracts between firms provide a possibly constrained efficient governance form between the market and the firm; *the bad*, where relational contracts are used to sustain

I am grateful to Luigi Guiso for many inspiring conversations and to Oliver Seager provided outstanding research assistance. Support from ERC Consolidator Grant *Sharing Gains* is gratefully acknowledged.

R. Macchiavello (✉)

London School of Economics and Political Science, London, UK

e-mail: r.macchiavello@lse.ac.uk

uncompetitive practices such as collusion; and finally, we sketch potential connections between studies of relational contracting and organized crime—the *ugly*.

Keywords

Relational contracts · Collusion · Organized crime

24.1 Introduction

Virtually every commercial transaction has within itself an element of trust, certainly any transaction conducted over a period of time. It can be plausibly argued that much of the economic backwardness in the world can be explained by the lack of mutual confidence. (Arrow, 1972)

Relational contracts are “informal agreements sustained by the value of future relationships” (Baker et al., 2002). Given this definition, it is highly likely, if not outright evident, that in our economies the vast majority of resources are allocated through transactions governed by relational contracts. Such an argument would unfold as follows. In the private sector of a “market” economy like the United States, the ratio of value added to sales is approximately half. This ratio gives us a proxy for vertical integration (Adelman, 1955) and thus an idea of the share of private sector transactions that occur *within firms*. Most of these transactions, therefore, happen within the context of employment relationships which are (almost by definition) highly incomplete (Simon, 1951) and most likely governed by relational contracts (Gibbons & Henderson, 2012). What about the rest of the transactions undertaken by the private sector? Here, data are harder to come by, but it likely wouldn’t be completely off the mark to conjecture that a very significant share—if not the vast majority—of transactions *between firms* occurs within the context of long-term supply arrangements that are also governed by relational contracts. For example, approximately 80% of US imports occur in preexisting relationships. Although the majority of relationships are newly formed, in any given year, preexisting relationships that are 3 years or older account for roughly half of imports. It could be argued that transactions across borders are particularly difficult, e.g., because of distance, different jurisdictions to enforce contracts, and so on, and that, therefore, these statistics overstate the importance of long-term relationships between firms. But there is now a sufficiently large body of evidence from industries as diverse as fishing, auto parts, airlines, trucking, pulp paper, and coffee—just to mention a few—that suggests that even domestic transactions are mostly governed through long-term supply relationships. Perhaps surprisingly, this appears to be the case also in industries in which parties trade relatively simple and undifferentiated products.

Relational contracts are thus ubiquitous both within and between firms. They are thus perhaps best understood not as an alternative mode of governance to spot markets and hierarchies, which is one of the potentially many hybrid modes of

governance supported by the prospect of an ongoing relationship (Williamson, 2005). Well-functioning relational contracts rather are the foundation upon which most—if not all—alternative modes of governance *should* operate. I emphasize “*should*” to highlight how, in practice, most relational contracts will be messy, and parties will struggle to achieve all the potential gains from cooperation. The reason is that relational contracts are built upon trust. Trust—the “firm belief in the reliability, truth, ability, or strength of someone or something,” according to the Oxford Dictionary—in turn requires parties to be able to exchange clear and credible promises.¹ Relational contracts will thus be hard to *build*, they will be hard to *manage*, and they will be hard to *adjust* when circumstances require change (Gibbons & Henderson, 2013). Many valuable relational contracts will thus never materialize or will perform well below the “Pareto frontier” available—at least in principle—to the parties. A firm, or hierarchy, in which managers and workers have failed to build a web of well-functioning relational contracts will not achieve its full potential. Similarly, a marketplace in which participants cannot rely on well-functioning relationships is unlikely to allocate resources close to the efficient, first-best, benchmark.

It is essential to gather a detailed understanding of how relational contracts work in practice. Although this understanding should be gained from a variety of sources and methodologies, quantitative empirical studies of relational contracts have a role to play in that process. Happily, relational contracting is a well-trodden area of study in economics, and many excellent reviews of the literature predate this chapter. On the theory side, Mailath and Samuelson (2006) and Watson (2021) review repeated games of relationships; Malcomson (2012) provides a comprehensive exposition of the theory of incentives in relational contracts. Fahn et al. (2023) introduce a recent special issue dedicated to relational contracts, with several chapters focusing on the theory. On the empirical side, Lafontaine and Slade (2012) review studies of interfirm contracting, both formal and informal (see also Gil & Zanarone, 2016, 2017). More recently, Macchiavello (2022) reviews empirical studies of relational contracts focusing on the importance of taking dynamic enforcement constraints—the building block of any repeated game theory of relationships—to the data. Empirical studies of relational contracts have also been reviewed through the lens of their applications in different fields. For example, Macchiavello (2022) focuses on studies of relational contracts in developing economics. Antras and Chor (2022) argue that relational contracts are particularly relevant in global value chains. This idea is well supported by detailed evidence from the coffee and garments value chains reviewed in Boudreau et al. (2023) and by Cajal-Grossi et al.’s (2023a) review of how relationships foster supply chain resilience. Michler and Wu (2020) and Macchiavello et al. (2022) review studies of relational contracts in agricultural

¹Trust, of course, can also arise in one-shot interactions through internalized social norms or heuristics (Boyd & Richerson, 1988). Trust in another party’s adherence to social norms however is not always optimal (see, e.g., Butler et al., 2016), and, therefore, while these norms might well be *necessary* to underpin, or facilitate, trust within ongoing relationships, they are probably not *sufficient* to achieve the complex cooperation that is required by most economic exchanges.

economics. Closer to this chapter, Macchiavello and Morjaria (2022) and Macchiavello and Morjaria (2023) review recent empirical work on relational contracting, focusing on their connection to alternative notions of trust.

This chapter argues for establishing stronger connections between empirical studies of relational contracts and other areas of work, thus complementing the existing reviews. In particular, existing reviews have almost exclusively emphasized how relational contracts benefit participating parties: given parties' inability to write complete contracts and courts' inability to enforce those, relational contracting allows for the realization of gains from trade that would otherwise be foregone. There is no reason to believe, however, that relational contracts will only be used to achieve socially desirable gains. Almost by definition, transactions that support illegal activities will have to be arranged through relational contracts. Indeed, many socially undesirable economic arrangements *must* be sustained informally. After a brief discussion of approaches to empirically study relational contracts, this review considers three broad contexts of applications of relational contracts: *the good*, where relational contracts between firms provide a possibly constrained efficient governance form in between the market and the firm; *the bad*, where relational contracts are used to sustain uncompetitive practices such as collusion; and, finally, we sketch potential connections between studies of relational contracting and organized crime—*the ugly*.²

24.2 Methodological Considerations

To fix ideas, this section provides a brief discussion of different approaches to the empirical study of relational contracts. The reader is referred to Macchiavello (2022) for a more detailed discussion. Back to our opening definition, relational contracts are “informal agreements sustained by the value of future relationships.” What do we mean by sustained? According to Baker et al. (2002), “the value of the future relationship must be sufficiently large that neither party wishes to renege.” The simplest framework to capture this intuition is the infinitely repeated prisoner's dilemma. Consider the familiar grim-trigger strategy subgame perfect Nash Equilibrium in which the two players cooperate provided both have done so in the past and otherwise permanently revert to the one-shot Nash Equilibrium in which both defect. The dynamic enforcement constraint gives a necessary condition for cooperation:

$$V \equiv \underbrace{\frac{\delta}{1-\delta}(C-P)}_{\text{Value of Relationship}} \geq \underbrace{D-C}_{\text{Temptation to Deviate}} \quad (24.1)$$

²We thus focus on empirical studies of relational contracts *between firms* and omit those *within* firms. However, we need to stress how relational contracts are, if anything, especially central to the employment relationship and to within firms organization. Besides space constraints, our choice is dictated by certain methodological considerations, further discussed in Sect. 24.2.

where δ is the (common) discount factor across periods, C is the payoff when both players cooperate, D is the payoff when the player deviates and the opponent cooperates, and P is the payoff when both deviate. Informal cooperation can be sustained if the opportunity cost of deviation—the “value of the relationship”—exceeds the opportunity cost of cooperation, the “temptation to deviate.” Ideally, therefore, empirical studies of relational contracts would *directly* measure and test predictions derived from the dynamic enforcement constraint. Doing so, however, presents several challenges to the applied researcher. Chief among these is that the value of the relationship V is not directly observed: it depends, *inter alia*, on discount rates and each player’s beliefs concerning the payoffs of its counterpart both on- and off-the-equilibrium path. The temptation to deviate is also difficult to observe: for example, unless the researcher has direct knowledge of the promises exchanged by parties in the relationship, she will not be able to know whether a party has deviated or not from the pre-agreed (contingent) plan of action. Some recent studies, however, have used conditions in the spot market—when it coexists alongside relationships—to proxy for temptations to deviate. For example, a supplier might be tempted to default on a promised shipment to the buyer if spot market conditions suddenly become more favorable. These proxies can then be used to bound the value of the relationship V and to test predictions regarding behavior when different types of shocks alter temptations. Note, however, that while this approach can be fruitfully applied when studying relationships *between firms*, it is significantly harder to proxy temptations to deviate in relational contracts *within firms*. This justifies this review’s focus on the former rather than the latter.

In the absence of direct proxies for temptations to deviate, empirical studies have often proxied relational contracts with measures of the age of the relationship, measured either in terms of calendar time elapsed from the earliest interaction or in terms of past volumes of trade. While this approach is common and intuitive, it is not ideal because—on top of various measurement concerns, including both left and right censoring in this type of data—our definition of relational contracts emphasizes the forward-looking nature of these arrangements: repeated trade doesn’t necessarily imply relational trade (see Macchiavello, 2022; Cajal-Grossi et al., 2023b for a discussion).

Finally, since relational contracts are imbued with exchanges of promises that are so rooted in parties’ circumstances that it is difficult for an outsider to adjudicate any dispute, a natural way to identify and measure relational contracts is through surveys. Survey instruments that are widely applicable across industries have been used in studies of formal management practices (see the influential body of work on management practices by Bloom & VanReenen, (2007, 2010) and in studies of relational contracts between firms (see, e.g., McMillan & Woodruff, 1999). A more tailored approach in which the survey tool is expressly designed to elicit the exchange of promises that are likely to be most relevant in a particular industry and context, however, has perhaps a greater potential to uncover the mechanics of, and constraints to, relational contracting (Macchiavello & Morjaria (2021, 2022) Acquisitions, management and efficiency in Rwanda’s coffee industry. Unpublished manuscript).

24.3 The Good

Among the legal privileges of corporations, two that are mentioned in textbooks are the right to sue and the “right” to be sued. Who wants to be sued! But the right to be sued is the power to make a promise: to borrow money, to enter a contract, to do business with someone who might be damaged. (Schelling, 1956, p. 299)

This section reviews selected empirical studies of relational contracts between firms, emphasizing the benefits that relational contracting brings to the parties involved. There are several themes the section seeks to highlight. First, we will document a variety of examples in which relational contracting helps parties overcome a variety of contracting problems. Somewhat ironically, empirical studies that demonstrate the usefulness of relational contracts often paint them in an unflattering light. On the one hand—as already mentioned—temptations to deviate, and deviations when they occur, are particularly useful to understand the value of relational contracts. Evidence of what triggers parties to cheat in the relationship, and the consequences of such behavior, is particularly helpful in understanding the inner functioning of relationships. On the other hand, many studies that attempt to quantify the welfare consequences of relational contracting use a “zero contracting frictions” benchmark as a counterfactual. Thus, the “welfare consequences” of relational contracts are presented as a shortfall, rather than, say, as a gain vis-à-vis an environment where agents can only interact anonymously and repeated interactions are altogether unfeasible. We ask that the reader interpret the evidence reviewed below bearing in mind that relational contracts are often a *solution* (though imperfect) to the underlying frictions of which they are *symptoms*. The quote above from Schelling (1956) would suggest that in weak institutional environments, where firms are not credibly vulnerable to lawsuits, relational contracting affords firms a “right to be punished,” and, therefore, the ability to make a credible “promise: to borrow money, to enter a contract, to do business with someone who might be damaged.” Extending this argument only slightly, this would suggest that relational contracting should be particularly widespread in developing economies with ineffectual courts and in international trade, where third-party arbitration and enforcement can be costly. In other words, in such environments, there will be a higher *demand* for relational contracting between firms.

Indeed, many of the studies reviewed in this section are taken from such environments.

And yet, I would disagree with such a generalization. On theoretical grounds, there are several arguments for why environments characterized by weaker institutions might have a lower *supply* of relational contracting (see Macchiavello, 2022). On the empirical front, I am not aware of systematic evidence that has rigorously tested the hypothesis. The section will indeed also review many studies that have demonstrated the importance of relational contracting between firms in the domestic portion of supply chains in developed countries. I will also argue that relational contracting might be leveraged to generate benefits that accrue to parties beyond those directly engaged in the relationship. This observation is particularly relevant in

the context of sustainable sourcing practices. Finally, it is important to understand how relational contracting between pairs of a buyer and a seller interact with each other, with other governance forms (most notably, spot markets and firms), with community structures, and with broader societal norms around trust and cooperation. Evidence is beginning to emerge in these areas, but more work is certainly needed.

Macchiavello and Morjaria (2015) offer perhaps the first empirical study of relational contracts between firms that takes the dynamic enforcement constraint to the data. The paper focuses on relationships between Kenya exporters and foreign importers of roses. The fact roses cross a border implies that transaction-level data between firms are observed from the customs records. More importantly, alongside these direct relationships between buyers and sellers, roses are internationally traded in a spot market, the Dutch Auctions in the Netherlands. Since most exporters also keep accounts open at the auctions, the spot market price in a particular week provides a direct proxy for the temptations to deviate: an exporter can always side-sell flowers to the auctions if this happens to be particularly profitable. The paper establishes three facts. First, it develops a test for whether the dynamic enforcement constraint (the equivalent of Eq. 24.1 above) is binding. The test reveals that at Valentine's Day—when roses are in highest demand, but also when the temptation to deviate is stronger since spot market prices are highest—the (unobserved) future value of the relationship constrains the amount of roses a buyer and a seller can trade. The temptation to deviate on Valentine's Day, thus, reveals the value of the relationship. Second, the value of relationships increases over time. This does not appear to be entirely driven by selection (more valuable relationships are more likely to survive). Finally, the paper asks why that might be the case. To answer the question, it exploits unanticipated supply shortages due to an intense episode of electoral violence (see Ksoll et al. (2021) for further details). The pattern of deliveries during the shock is most consistent with the hypothesis that a seller's reputation for reliability—which is accumulated over time—is an important source of value in these relationships.

Blouin and Macchiavello (2019) document how relationships mitigate strategic default on forward contracts in the international coffee market. The main idea of the test is that when the world market price unexpectedly surges relative to what parties had anticipated at the time of contracting, a seller who has committed to deliver coffee at a future date for a fixed price will have a strong incentive to default. Using an extremely detailed dataset of such contracts, the paper finds evidence for such behavior. A relationship between the buyer and the seller, however, mitigates this behavior. Parties can, of course, prevent such default by entering price-indexed contracts in which the price to be paid upon delivery tracks the world market price. Indeed, defaults do not occur under such contracts. There thus is a trade-off: insuring against price risk generates counter-party risk. This risk is sizeable: the estimates reveal that the threat of strategic default reduces output traded in the average relationship by 26%. Buyers and sellers start with price-indexed contracts and only when they have accumulated sufficient trust, i.e., relationship value, switch to fixed price contracts. Using reference prices in the relevant spot market, both Macchiavello and Morjaria (2015) and Blouin and

Macchiavello (2019) quantify the value of relationships. In both cases, relationships appear to be very valuable (see Fig. 24.1).

Antràs and Foley (2015) demonstrate how long-term relationships between buyers and sellers affect the terms of payment in international trade. The authors study the payment terms specified by a large US firm exporting to more than 140 countries. They find that new buyers in locations with weak contract enforcement are required to pay up-front for goods. As the relationship ages—and as the sellers' belief about the value of the relationship increases—trade credit is more commonly extended to buyers in destinations where institutions are weak. In a macroeconomic note, the authors highlight the liquidity problems firms faced during the 2007–2008 financial crisis. They then show that credit *within* relational contracts can make supply chains more resilient to such crises.

Brugués (2023, Take the goods and run: Contracting frictions and market power in supply chains. Unpublished manuscript) also studies how long-term relationships are used to support trade credit and discipline buyers' opportunistic behavior. The paper leverages transaction-level VAT data across 49 Ecuadorian firms across the cement, pharmaceutical, and textiles industries. If extended credit, at any point in a relationship, the buyers can “take the goods and run,” reneging on payment. Estimating a structural model, Brugués compares welfare gains in each trade relationship to the first-best case of perfect contract enforcement. He finds that, in its first year, the median relationship achieves less than 25% of first-best surplus; after 5 years, the median relationship achieves over 80% of first-best surplus. Crucially, the seller's ability to price discriminate is essential to provide the dynamic incentives in the relationship—there is thus a tension between the extent of competition and the potential for relational contracting.³

Several studies have also demonstrated the importance of relational contracting between firms in institutionally more advanced economies. Taken together, these studies highlight how relational contracting helps parties adapt to circumstances in a wide variety of contexts. Barron et al. (2020) show how relational contracts between movie distributors and theater operators sustain mutually beneficial adaptation to changing circumstances, even when other aspects of the interfirm relationship can be formally contracted upon. In their context, parties sign a formal contract to share box office revenues for movie showings, leaving open when or how frequently a movie will be shown. Theaters are free to change their schedules as they update their beliefs about how popular a movie will be. The split of box-office revenue—which is included in the contract—is often renegotiated *after* a movie's run. This ex post

³ Also on trade credit, and also from a developing country setup, Ghani and Reed (2022) provide a fascinating account of what happened to relationships between ice retailers and informal fishermen when the market for ice in Freetown, Sierra Leone, was disrupted by the entry of a new ice manufacturer. Initially, there was a marked increase in relationship instability, with fishermen switching suppliers. In turn, retailers began offering a new relational contract in which interest-free credit was provided to fishermen. This change in the relationship progressively restored the stability of relationships.

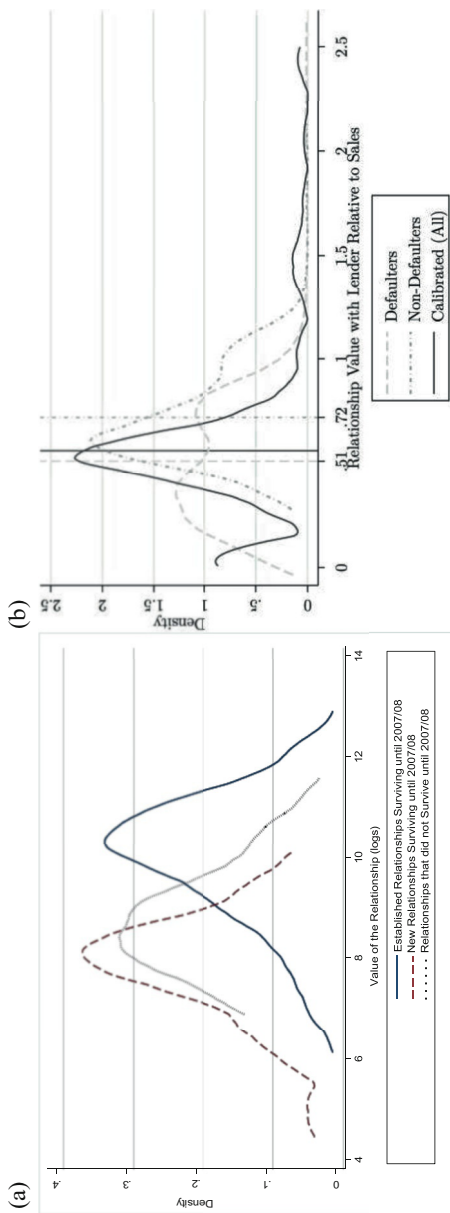


Fig. 24.1 Value of relationships. (a) Relationship value—roses (The left panel reproduces Fig. 7 from Macchiavello and Morjaria (2015); the right panel reproduces Fig. 11 from Blouin and Macchiavello (2019). The combined figures also appear in Macchiavello and Morjaria (2022). Both panels illustrate the distribution of the estimated value of relationships. The left panel distinguishes established relationships that survive into the following season (solid line); newly formed relationships that do not survive onto the next season (gray line). The right panel distinguishes a higher and a lower bound for the value of the relationship estimated for defaulters and non-defaulters, respectively, from the distribution estimated from the structural model. Vertical lines report average estimated relationship values for the three different estimates)

adaptation signals the theater's goodwill in the ongoing relational contract and helps in negotiations over revenue sharing for future movies.

Long-term relationships governed with highly incomplete formal contracts are common in many other industries as well. For example, Darmouni et al. (2023, *Pulp friction: The value of quantity contracts in decentralized markets*. Unpublished manuscript) show how, despite its highly commodity-like nature, the vast majority of pulp is traded within long-term relationships in which parties agree on highly incomplete quantity contracts that are then "filled-in" over time. Besides coffee (see Blouin and Macchiavello (2019) above and Macchiavello and Miquel-Florensa (2018, *Vertical integration and inter-firm relationship in the Costa Rica coffee chain*. Unpublished manuscript)), similar arrangements have been documented in coal (Joskow, 1988), bulk shipping (Brancaccio et al., 2023), and liquefied natural gas (Zahur, 2022, *Long-term contracts and efficiency in the liquefied natural gas industry*. Unpublished manuscript).

Gil et al. (2022) study relational contracts between major airlines and the regional airlines to whom they outsource short-haul routes. In the event of negative weather conditions, airport "landing spots" become scarce, and adaptation is required (see Forbes & Lederman, 2009). A major airline may then ask that a regional carrier to whom it outsources postpones a flight (which is costly to the regional) to free up a landing spot (which is beneficial to the major). Despite software's increased sophistication in the industry, this practice remains hard to formally contract upon and must be sustained relationally. In the wake of the 2007–2008 financial crisis and the concomitant drop in airline profitability, relationships that historically have had to endure worse weather shocks—and that, therefore, have had more opportunities to learn about the value of their relationship—are more likely to survive. Harris and Nguyen (2023a, *Long-term relationships in the US truckload freight industry*. Unpublished manuscript) study relational contracting between shippers and preferred freight carriers in the US trucking industry. A shipper has a preference ranking over carriers for each different shipping route. A centralized software then sequentially offers a shipper's consignment moving down the ranking until a carrier willing to accept the order is found. They show how carriers are more likely to reject the offer when the spot market rates on a particular route are high and that, when this happens, the carrier is more likely to be subsequently demoted from the list. Since shippers and carriers often interact over multiple shipping routes, the authors test whether the relational contract applies at the firm-to-firm level or whether they are limited to specific lines. Somewhat surprisingly, they find that most often shippers limit punishment to the specific route on which the carrier defected, while multi-lane punishment is rarer.

So far, we have reviewed studies that focus on how a bilateral relationship between a buyer and a seller functions. Several recent studies focus on how these relational contracting is affected by, or affects, other parties. Space constraints prevent us from a detailed coverage of this emerging body of work. The first strand of work studies how competition—i.e., the availability of alternative parties to trade with—affects relational contracting. If the availability of alternative partners increases temptations to deviate and/or, through competitive forces, reduces the

value of the relationship, it will make it harder to satisfy the dynamic enforcement constraint, potentially leading to relationship breakdown. Consistent with this hypothesis, Boudreau et al. (2023) document how the prevalence of long-term relationships at multiple stages of both the coffee and garment value chains negatively correlates with the intensity of competition. Fig. 24.2, taken from Boudreau

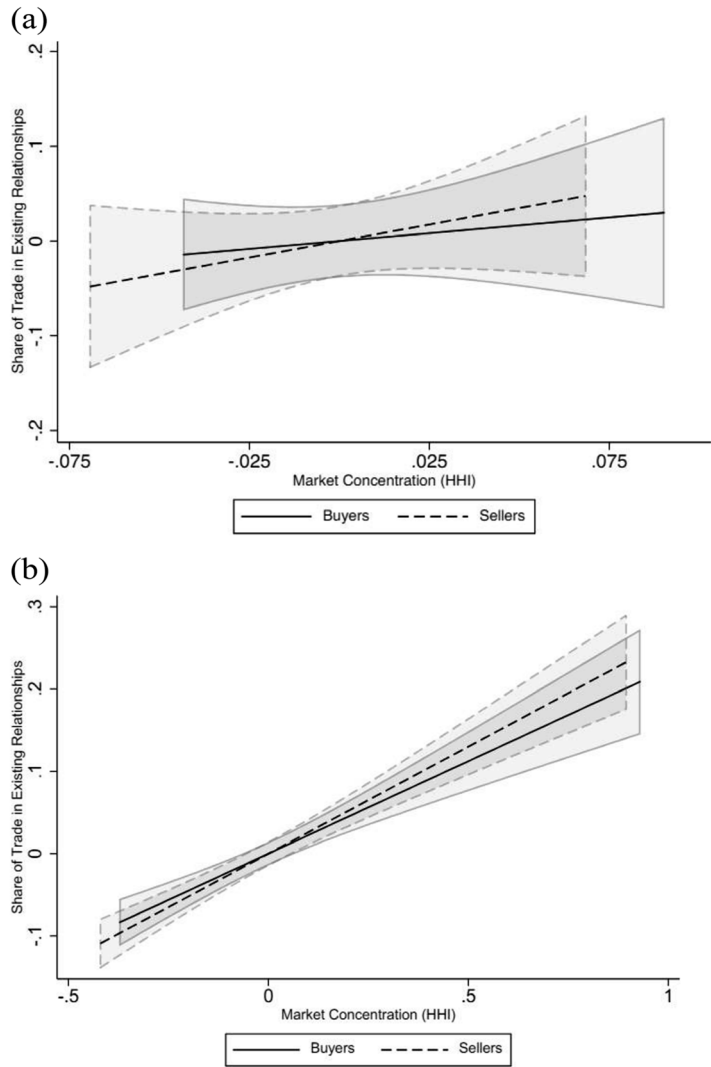


Fig. 24.2 Market concentration and relationships at the export gate (a) Coffee (b) Garments. (This figure is reproduced from Boudreau et al. (2023)). The figure shows the correlation between market concentration and the share of exports traded in relationships in coffee and garments. Horizontal and vertical axis' variables are residualized against the size of the market, in terms of exported values. The linear fits over 14 observations in coffee and 1113 observations in garments are presented alongside 95% confidence intervals)

et al. (2023), illustrates the point. At the export gate, the share of a given type of garments sourced from a particular origin that is transacted in long-term relationships is positively correlated with the degree of concentration in that particular market segment. The same happens to be the case in coffee, although data limitations reduce the available sample size. Further upstream, the quality of relational contracts is also inversely related to competition. Macchiavello and Morjaria (2021) more rigorously test the hypothesis in the Rwanda coffee sector. The paper combines original surveys to measure relational contracting between coffee mills and farmers and, through an ingenious instrumental variable strategy, demonstrates how higher competition between mills lowers relational contracts with farmers (Fig. 24.3a). Boudreau et al. (2023) also find that the quality of industrial relationships between workers and employers—i.e., an index of the quality of the relational contracts *within* large garment exporters—is negatively correlated with the density of factories in the area (Fig. 24.3b).

More recently, Harris and Nguyen (2023b, Long-term relationships and the spot market: Evidence from US trucking. Unpublished manuscript) extend the analysis in Harris and Nguyen (2023a, Long-term relationships in the US truckload freight industry. Unpublished manuscript) to model the two-way interaction between long-term relationships and the spot market. On the one hand, and as noted above, the spot market introduces temptations to deviate in long-term relationships. On the other hand, trade in long-term relationships potentially exerts a negative externality reducing the thickness of the spot market. Although they also find evidence for this second channel, estimates from the structural model confirms that parties' abilities to sustain relational contracting is essential for the overall efficiency in the industry. In particular, the benefits of long-term relationships—which the paper “backs out” from carriers' initial bidding and subsequent order acceptance behavior—outweigh the negative effects on the spot market.

A second strand of work focuses on sourcing systems (see Taylor and Wiggins (1997) for an early model). Cajal-Grossi et al. (2023b) show how international buyers of garments differ in their sourcing strategies. They develop a proxy for relational sourcing and find that buyers greatly differ in their approach to procure garments, even when sourcing from the same origin, within narrowly defined product lines. These differences correlate across both countries and products. In other words, holding constant the buyer's overall procurement volumes, some buyers are “relational,” i.e., tend to source through long-term relationships with relatively few suppliers, while others are “spot”—i.e., frequently switch between a larger number of suppliers. The authors also document that suppliers earn higher margins on otherwise identical orders produced for relational as opposed to spot buyers. This is consistent with the theoretical prediction that relational sourcing provides rents to the supplier. These rents are likely needed to discipline the supplier to take certain hard-to-contract upon actions (e.g., be reliable or spare capacity in the event of an unexpected surge in demand). The suppliers' capabilities to deliver on these important margins, in turn, increase domestic value addition and constitute an important dimension of “upgrading,” particularly for exporters in low-income

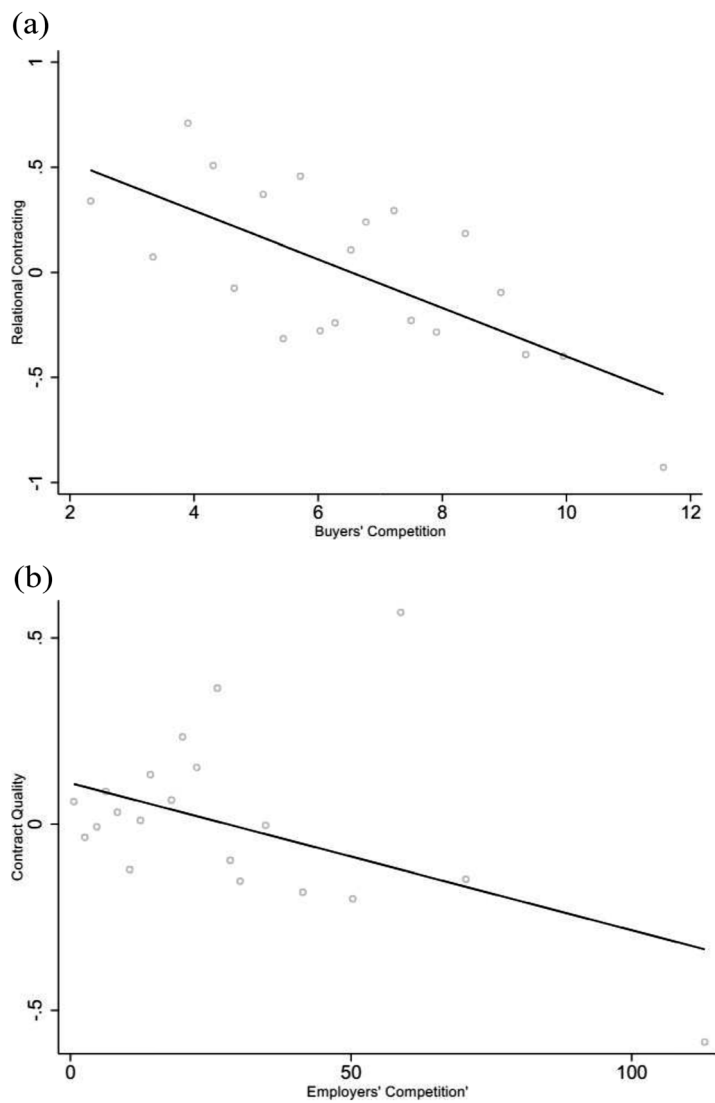


Fig. 24.3 Competition and relationships in the domestic stage of the supply chain **(a)** Coffee **(b)** Garments. (This figure is reproduced from Boudreau et al. (2023). Panel A shows the correlation between spatial competition and relational contracting between mills and farmers in the Rwanda coffee chain (data from Macchiavello & Morjaria, 2021). The variables in both axes are residualized against a set of geographic controls. Panel B shows the correlation between spatial competition and relational contracting between garment factories and workers in the Bangladeshi garment chain. (Data from Boudreau et al., 2024, Relationships and responsibility. Unpublished manuscript). The relational contracting index is residualized against plant size (as measured by its own exports on the year of assessment)

countries. Alongside traditional interventions targeting specific products or destination markets, export promotion agencies might want to particularly promote exports to “relational” buyers.

Reviewing evidence from garments (Amengual and Distelhorst, 2020, Cooperation and punishment in regulating labor standards: Evidence from the Gap Inc supply chain. Unpublished manuscript; Boudreau, 2024, Multinational enforcement of labor law: Experimental evidence on strengthening occupational safety and health (OSH) committees. Unpublished manuscript) and coffee (Macchiavello & Miquel-Florensa, 2019, Improving export quality: The case of the sustainable quality program in Colombian coffee. Unpublished manuscript), Boudreau et al. (2023) argue that buyers’ ability to develop and maintain relational contracts with suppliers might be a *necessary* precondition for responsible sourcing. Boudreau et al. (2024, Relationships and responsibility. Unpublished manuscript) more rigorously test this hypothesis and find evidence that relational sourcing is strongly associated with better working conditions in the garments apparel chain.

24.4 The Bad

People of the same trade seldom meet together, even for merriment and diversion, but the conversation ends in a conspiracy against the public, or in some contrivance to raise prices. (Smith, 1776)

Parties therefore rely on informal arrangements sustained by the value of future interactions—relational contracts—to cooperate when contracts are unenforceable. While the studies reviewed above demonstrate how these arrangements benefit their participants, there is no guarantee that they would benefit the market as a whole. Horizontally competing firms colluding to raise prices—cartels—offer perhaps the most prominent example. As with relational contracting, one might imagine that cartels might be particularly relevant in developing countries due to higher entry barriers (Djankov et al., 2002), weaker competition authorities (World Bank, 2016; Besley et al., 2020), and more concentrated markets (Mitton, 2008; Leone et al., 2024, The high and falling price of cement in Africa. Unpublished manuscript). However, as with relational contracting, I am not aware of rigorous tests of this hypothesis. The Private International Cartel database reveals that only 5% of proven cartels are in Africa (72% of those were proven in South Africa alone), 7% in Latin America, and 11% in Asia. These figures however must be taken with a pinch of salt—weaker competition institutions in lower income countries imply that fewer cartels will end up being detected. At any rate, according to Adam Smith, we should expect collusion to be a pervasive phenomenon of economic life.

And yet, it has for long been well-known to industrial organization economists that cartels appear to be not quite as widespread. There are two main reasons for that. First, free-riding: when competitors collude, a firm has the incentive to stay out of the cartel enjoying the higher prices generated by the cartel without having to restrict its behavior (see D’Aspremont et al., 1983) for an early formalization of this

argument). Second, as eloquently articulated by Stigler (1964), there is “the problem of policing a collusive agreement, which proves to be a problem in the theory of information.”

This section highlights a few selected recent contributions that best illustrate the fruitful connection between empirical studies of cartels and relational contracts. From a theoretical perspective, the connection is very strong: often advances in the repeated-game theory of cartels (see, e.g., Green & Porter, 1984; Athey & Bagwell, 2001, 2008) have contributed to develop the “machinery” required to formally study relational contracts in other applications. On the empirical front, the enormous literature on cartels can be divided into two branches. One strand focuses on *detecting* collusion in the data, designing empirical tests to identify conduct in the tradition of Bresnahan (1982). A second strand studies the behavior of *known* cartels, shedding light on the mechanisms through which agreements are sustained and enforced. Our goal is to highlight how connections with the empirical literature on relational contracts can be fruitful for both strands. We thus focus on papers that have studied collusion deriving tests from, or explicitly discussing, the dynamic enforcement constraints associated with collusive behavior.

Chassang and Ortner (2019) derive a test for collusion from dynamic enforcement constraints and apply it in the context of repeated procurement auctions in Japan. In the presence of collusion, they argue that an auctioneer’s stipulating higher *minimum* bids for procurement projects can cause the average bid to *decrease*. This prediction—which is strongly supported by a “differences-in-differences” analysis of detailed procurement data from 14 municipalities in Japan—naturally arises from the dynamic enforcement constraint. A higher minimum bid leaves more rents to auction participants when they compete. If, upon deviating from the collusive arrangement, cartel members revert to a competitive equilibrium, then a higher minimum bid reduces the future value of the collusive agreement, potentially leading to its demise and to lower, more competitive, bids.

Igami and Sugaya (2022), instead, study a well-documented case in which 21 firms, led by Hoffman-La Roche, cartelized several vitamin and mineral products in the 1990s. The authors develop a structural model to calibrate the dynamic enforcement constraint for cartel members. This requires estimating each firm’s “incentive to collude,” defined as the discrepancy between the value of the relationship under collusion and the temptation to deviate. In turn, this requires knowledge of key primitives to compute counterfactual conduct outcomes, since cartel members are either observed colluding or not. This methodology does a wonderful job at matching the timing and behavioral patterns of the cartel and its demise and provides a valuable tool for future studies of cartels.

Bernasconi et al. (2022, Market transparency, relational contracts and collusion in the Colombian electricity market. Unpublished manuscript) uncover collusion in the Colombian energy sector deploying a novel test that is derived from dynamic enforcement constraints considerations. Collusive behavior is notoriously difficult to identify (Chassang & Ortner, 2022). The reason is that when colluding, firms deviate from current profit maximization in anticipation of future rewards. Profit maximization, however, places few restrictions on firms’ behavior. Tests for

collusion, then, have to pay a hefty price in terms of required assumptions. The paper takes advantage of a regulatory change that, by reducing market transparency, would have potentially made collusion harder to sustain. The paper finds that a subset of the firms in the market lowered bids by between 47% and 30% immediately after the *announcement* — and well before the actual *implementation* — of the regulatory change. Consistent with the relational contracting logic (at least some), members of the cartel reacted to the announcement in an anticipatory way, leading to its instantaneous unraveling. Exploiting future changes in relationship values, in this case driven by the *announcement*, provides a robust test that could be applied in other contexts. The test doesn't rely on cartel members perfectly foreseeing the actual change in transparency enacted by regulators and how it would make the cartel unsustainable. It simply requires some of them to become sufficiently pessimistic about their *future* ability to sustain a cartel.⁴

The empirical study of known cartels sometimes can rely on access to investigative records that provide precious information to understand exactly how the relational collusive arrangement worked. Genesove and Mullin (2001) consider candid notes covering meetings of the Sugar Institute from January 1929 through July 1930 and provides perhaps the most notable example of such studies. The Sugar Institute comprised 14 firms and nearly all US sugar refining capacity. The authors note that the Sugar Institute never explicitly fixed the prices or output quantities of its members. Instead, the cartel forbade price and non-price discrimination among a firm's buyers (e.g., discounted freight rates, favorable credit terms, and favorable delivery schedules). Additionally, the Institute mandated the public posting of sugar prices. This facilitated tacit collusion by making price cuts by any one firm easier to monitor. While the general policies of the cartel accord with insights from repeated-game logic, the authors identify several aspects of the cartel's operations that are not well-captured by standard models of collusion. For example, subgame-perfect Nash equilibrium requires that parties form a complete contingent plan. At a minimum, we would expect the cartel to agree on a known punishment for each eventuality. Instead, the Sugar Institute met weekly to review the observed behavior and reported plans of cartel members: the cartel is committed only to making democratic resolutions in the light of each circumstance, not to a fully mapped "punishment function." Furthermore, integral to the Sugar Institute's enforcement mechanism was *ex ante* communication (as opposed to *ex post* verification) of a firm's plans. Members were expected to notify—and indeed did so—the Institute any time they changed their terms of trade, or sold damaged sugar at a discount, or launched a new brand or advertised. Finally, even though several instances of cheating through creative accounting were detected within the Sugar Institute, none of those were met with a reversion to competitive conditions. Deviations were either ignored or matched without further repercussions.

⁴In particular, the evidence doesn't pin down a particular equilibrium concept of collusive conduct. For example, while subgame-perfect equilibrium (SPE) relies on players reacting to information instantaneously, it also assumes correct beliefs about future payoffs and plays both on- and off-equilibrium paths, which are not confirmed by the evidence.

24.5 The Ugly

In every transaction in which at least one party does not trust the other to comply with the rules, protection becomes desirable, even if it is a poor and costly substitute for trust. This book shows that the mafia protection fulfills this role, albeit in an erratic and limited fashion. (Gambetta, 1996)

Tens of millions of people worldwide live in areas in which powerful criminal organizations (OCGs) rule or exercise significant influence on many facets of daily life (Lessing, 2021). In contexts as diverse as Latin America, Italy, the United Kingdom, India, South Africa, and the American prison system, criminal organizations play an important role in organizing complex illegal activities, such as international drug trafficking, arms smuggling, racketeering, and many others. The influence of these organizations is not confined to countries with weaker institutional environments; indeed organized crime increasingly poses serious threats in advanced economies as well (see, e.g., Transcrime, 2017). Beyond the obvious illegal activities, however, these organizations often also provide basic public goods and social order, enforce property rights, provide mechanisms for dispute resolution, and various forms of economic assistance, filling in gaps left by poorly functioning states. Further, beyond those traditional core activities, OCGs reach deep into the legal fabric of our economies. The United Nations estimates the proceeds from organized criminal activities at USD 2.1 trillion or 3.6% of the global GDP (UNODC, 2011). It is plausible to assume that OCGs face significant barriers to the reinvestment of their criminal profits into the criminal industry. For example, OCGs cannot spend significant resources on advertising to acquire new customers without increasing the risk of detection and can only rely on a few trusted counterparties to conduct their criminal business. These considerations suggest that the share of the legal economy that is infiltrated by OCGs is potentially substantial—though rigorous estimates of its size and motives remain elusive.

This final section seeks to establish a connection between the study of organized crime and the study of relational contracts. This area is quite unexplored and, therefore, the connections highlighted here will appear quite loose. Rather than a summary of existing work at the intersection of these two fields, the goal of this section is to provide an indication of possible areas for further work. There are potentially multiple lines connecting relational contracting and organized crime. A first, perhaps obvious, connection is that criminals cannot rely on courts to enforce contracts over their illegal transactions. As in the case of collusion, the governance of illegal transactions will need to be based on reputational forces, sometime of a relational nature. For example, Mehmood and Chen (2024, Contract enforcement in a stateless economy. Unpublished manuscript) conduct a field experiment at a large illegal gambling market in Pakistan. Contracts are inherently unenforceable due to the illicit nature of gambling. They find that reputational sanctions, such as excommunication from the group and social image sanctions, facilitate (illegal) betting in this context. Galenianos and Gavazza (2017) estimate a structural model of illicit drugs markets using data on purchases of crack cocaine. In their model, buyers search for high-quality (e.g., in terms of purity) drugs. Quality, however, can only be determined after it is consumed.

Sellers have the choice between selling low-quality drugs to their first-time buyers or offering higher-quality drugs to maintain a good reputation and induce buyers to purchase from them again. The structural model is transparently estimated and provides a framework to conduct counterfactual policy evaluations with respect to the impact of legalization or of larger penalties, while taking into account reputational forces. Beyond this type of mechanical connection between the importance of reputational forces in markets for illegal transactions, we highlight two additional potential connections between studies of organized crime and relational contracting. The first emphasizes the role that OCGs can play in supporting transactions when there is limited trust. The second emphasizes the relational nature of OCGs.

Gambetta's (1996) classic book on the Sicilian Mafia provides one of the most convincing conceptualizations of criminal organizations. The Sicilian Mafia is portrayed not merely as a criminal organization engaged in illegal activities, but rather as a sophisticated enterprise that thrives in a society where trust is scarce and the state's provision of security and justice is inadequate. Gambetta's analysis is grounded in economic theory and sociological insight, offering a unique perspective on the Mafia's role as a provider of private protection. He argues that the Mafia emerges to fill the vacuum created by the state's failure, selling a guarantee of safe conduct for parties, both in legal and illegal commercial transactions. Operating in the shadows of the legal system, the Mafia leverages its reputation for the credible (the threat of the) use of violence as a means of ensuring contractual compliance. While this service is essential in a society where trust is in short supply and democracy is weak, a central theme in Gambetta (1996) is how the Mafia is, paradoxically, both predatory and protective. The ambiguity in Mafia's role arises from the fact that the Mafia offers protection against any abuse—including those it directly threatens. The organization is thus simultaneously engaged in protection *and* in extortion.

Dixit (2004) provides a game-theoretic analysis of how property rights can be protected and contracts enforced by private institutions that transcend or supplement weak economic governance provided by the state. Dixit (2004) considers several privately provided mechanisms designed to enforce contracts. These include long-term relationships, arbitration, social networks for information dissemination, norms enforcement, and for-profit enforcement services. The latter is, of course, closely connected to the analysis in Gambetta (1996). Dixit (2004) emphasizes how these mechanisms operate alongside formal legal systems—often in the shadow of the law. In fact, the book elucidates both the inefficiencies associated with the for-profit private supply of contract enforcement but also why—even in the presence of strong legal frameworks—for-profit private supply of contract enforcement emerges as part of the equilibrium and should thus not be regarded as an anomaly, but rather as an essential component of economic governance. This observation resonates well with our earlier remark that criminal organizations are widespread, including in countries with strong legal systems. In other words, criminal governance has survived the growth of strong states.⁵

⁵ Empirical studies have confirmed the negative impact of OCGs on economic development. For the case of Italy, see, e.g., Pinotti (2015) and, more recently, Fenizia (2023).

Criminal organizations are thus inherently relational. A particularly neat illustration is provided by the 'Ndrangheta, an Italian criminal organization that originated in the peninsular region of Calabria and that dates back to the nineteenth century. The 'Ndrangheta is Italy's most powerful organized crime syndicate and is considered one of the largest and most sophisticated criminal organizations worldwide by enforcers such as the FBI and Europol. Although the 'Ndrangheta is involved in a wide-spectrum of illegal activities, such as fraud, extortion, high-level corruption, and trafficking in both humans and weapons, it is generally believed that its most lucrative revenue stream is given by the proceedings from international drug trafficking. The organization has been the subject of trials and investigations in many countries across continents—from Italy and Germany, to Canada and Australia. A striking aspect of the organization, emphasized by several judicial investigations, is its ability to transplant its original structures in its region of origin to the new areas in which it seeks to expand. This aspect—which has allowed the organization to become so powerful and entrenched—is often credited to the fact that, unlike the Sicilian Mafia (which is highly centralized), the 'Ndrangheta is organized as a more decentralized network of *'ndrina* (clans) that are heavily based on tight family ties. It is only relatively recently that investigations have ascertained how the 'Ndrangheta also comprises a centralized structure above the clans. The purpose of this committee is to coordinate the activities of, identify potential opportunities for collaboration between, and resolve disputes among the otherwise fairly autonomous clans.

Given the importance of relationships for OCGs, it would seem essential to better understand how such relationships are established and maintained. In a way somewhat reminiscent of the Genesove and Mullin (2001) study of the Sugar Institute, Gambetta (1996) gathers evidence from the confessions of eight Mafiosi. This evidence is used, among other things, to shed light on the organization's internal mechanisms to manage the “relational contracts” that are essential to its functioning. For example, Gambetta (1996) rationalizes the frequent use of blood ceremonies and other rituals and obscure symbols as a mechanism through which the organization asserts its authority. But, borrowing from Gibbons and Henderson's (2012) framework, we could perhaps interpret this symbolism as an essential way through which the organization builds *clarity* about the relational contract among its members, by means of reinforcing narratives that perpetuate the “organizational culture” (Gibbons & Prusak, 2020). In the case of the 'Ndrangheta, family ties constitute the primordial nucleus upon which the *'ndrina* is formed and operates. However, the execution of complex illegal activities—such as international drug trafficking or corruption of public procurement—necessitates relationships both with other members that are not part of the family, as well as with individuals that are *not* members of the organization and, yet, are willing to cooperate with it. Leveraging unusually detailed investigative records, Facchetti et al. (2024, The organizational economics of 'Ndrangheta. Unpublished manuscript) provide a systematic exploration of how 'Ndrangheta's *'ndrinas* not only specialize in certain types of crime, but also how this specialization translates into different allocation of tasks across individuals with different roles. In particular, complex criminal activities are organized by

assigning different types of crime (“tasks”) to members of different seniority, as well as individuals outside of the organization.

For criminal organizations, relationships are thus of critical importance not only to conduct illegal activities but also to infiltrate the legal economy. Arellano-Bover et al. (2024, *Mafias and firms*. Unpublished manuscript) provide a comprehensive analysis of OCGs’ infiltration of legal firms in the Italian context, where leading OCGs originate and unique information is available. The paper develops a novel conceptual framework for OCGs’ infiltration of legal firms—distinguishing between alternative motives of infiltration. Although the conceptual framework can be applied to other countries, the endemic presence of OCGs and a sophisticated legal and institutional apparatus to fight them make Italy an ideal setting for our analysis. The paper then leverages highly confidential data from the Italian anti-money laundering agency (the Financial Intelligence Unit of the Bank of Italy) to map OCGs’ investments in legal firms in Italy and test the predictions of the model. The data identify around 100,000 legal firms, approximately 2% of all Italian corporations and partnerships, as potentially connected to OCGs.⁶ The paper uncovers a previously unnoticed motive of infiltration. Traditionally, both the legislator and the existing literature have assumed that the infiltrated firm is either used for criminal activities (e.g., to facilitate money laundering) or to benefit from criminal activities (e.g., threats to competitors to acquire market shares). In both cases, the firm is directly “contaminated” by OCGs’ criminal activities. Alongside these traditional motives of infiltration, however, the conceptual framework highlights—and the empirical analysis supports—a novel infiltration motive in which the legal firm is *not* directly involved in criminal activities. In this motive—which the paper labels “pure” to distinguish it from the various “contaminated” motives—the infiltrated firm does not display drastic changes in its operation or performance, the hallmarks of the “contaminated” infiltration. In contrast, in the “pure” motive, the only detectable change is a substitution away from bank lending, consistent with the entry of fresh capital brought in by the criminal organization. Although there is some evidence that the OCG sometime targets firms that are experiencing liquidity problems, in many other cases, that doesn’t appear to be the case. Since the “pure” motive is detected on large firms that, presumably, have relatively good access to sources of capital, this begs the question of what are the main benefits behind this type of infiltration? The paper argues that the acquisition of *relational capital* by the criminal organization is likely an important factor. For example, infiltrated firms that

⁶Following the literature, the paper defines infiltration of an OCG into a legal firm as “any case in which a natural person belonging to a criminal organization or acting on its behalf, or an already infiltrated legal person, invests financial and/or human resources to participate in the decision-making process of a legitimate business” (Savona & Berlusconi, 2015, p. 19). The key feature of this definition is that a person tied to an OCG plays an active role in the decision process of a legal firm. Hence, in the data, a firm is infiltrated if it has at least an owner *or* an administrator on a highly confidential list of individuals available to the Financial Investigative Unit. Note that the accrual of financial resources is thus not a necessary condition for infiltration.

are more likely to reflect the “pure” motive have a particularly strong central position within the network of owners and board members and are disproportionately likely to have politicians on their boards. This suggests that the deliberate acquisition of *relational capital* outside the organization—facilitated by the availability of large sums of cash—is an absolutely central feature of OCGs’ strategy in Italy.

The evidence on the relational nature of OCGs’ activities, in particular in their relationships with the “legal” part of the economy, is so far only suggestive. More detailed evidence on how these relationships are formed, how they evolve, and how they adapt to new business opportunities as well as to enforcement actions should be a priority area for research to deepen our understanding of criminal organizations and help designing more robust policies to curb their financial crimes and money laundering activities. While this section has mainly focused on the Italian case, evidence in this area would be equally important from other contexts in which OCGs have a different portfolio of criminal activities, face different investment opportunities in the legal economy, and are contrasted by states with more limited capacity.

24.6 Conclusions

Relational contracts—defined as “informal agreements sustained by the value of future relationships”—are a pervasive feature of economy life and, both within and between organizations, likely govern how most resources are allocated. In recent years, a body of empirical work has begun to emerge that has documented the prevalence of relational contracting in a wide variety of industries and contexts and has rigorously tested many of the implications of the underlying theory. Most of this work has focused on understanding how relational contracting benefits participating parties, enabling them to cooperate on important aspects of their relationship that would be otherwise impossible, or prohibitively costly, to formally contract upon. Among other things, relational contracts have been shown to facilitate the reliability of supply and the resilience of supply chains, a better planning of capacity utilization, the financing of working capital, the enforcement of voluntary sustainability standards, and better working conditions. There is no reason to believe, however, that relational contracts will only be used to achieve socially desirable forms of cooperation. Almost by definition, transactions that support illegal activities cannot rely on formal enforcement of contracts and will have to be arranged through relational contracts. This chapter has therefore argued that the empirical analysis should be extended beyond *the good*, where relational contracts facilitate desirable transactions and the efficient allocation of resources, to also consider *the bad*, where relational contracts are used to sustain uncompetitive practices such as collusion and *the ugly*—where relational contracts support the activities of complex criminal organizations that threatens our societies.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Adelman, M. A. (1955). Concept and statistical measurement of vertical integration. In G. J. Stigler (Ed.), *Business concentration and price policy*. Princeton University Press.
- Antràs, P., & Foley, C. F. (2015). Poultry in motion: A study of international trade finance practices. *Journal of Political Economy*, 123(4), 853–901.
- Antràs, P., & Chor, D. (2022). Global value chains. *Handbook of International Economics*, 5, 297–376.
- Arrow, K. J. (1972). Gifts and exchanges. *Philosophy and Public Affairs*, 1(4), 343–362.
- Athey, S., & Bagwell, K. (2001). Optimal collusion with private information. *The Rand Journal of Economics*, 32(3), 428–465.
- Athey, S., & Bagwell, K. (2008). Collusion with persistent cost shocks. *Econometrica*, 76(3), 493–540.
- Baker, G., Gibbons, R., & Murphy, K. J. (2002). Relational contracts and the theory of the firm. *Quarterly Journal of Economics*, 117(1), 39–84.
- Barron, D., Gibbons, R., Gil, R., & Murphy, K. J. (2020). Relational adaptation under real authority. *Management Science*, 66(5), 1868–1889.
- Besley, T., Fontana, N., & Limodio, N. (2020). Antitrust policies and profitability in nontradable sectors. *American Economic Review: Insights*, 3(2), 251–265.
- Bloom, N., & Van Reenen, J. (2007). Measuring and explaining management practices across firms and countries. *The Quarterly Journal of Economics*, 122(4), 1351–1408.
- Bloom, N., & Van Reenen, J. (2010). Why do management practices differ across firms and countries?. *Journal of Economic Perspectives*, 24(1), 203–224.
- Blouin, A., & Macchiavello, R. (2019). Strategic default in the international coffee market. *The Quarterly Journal of Economics*, 134(2), 895–951.
- Boudreau, L., Cajal-Grossi, J., & Macchiavello, R. (2023). Global value chains in developing countries: A perspective from coffee and garments. *Journal of Economic Perspectives*, 37(3), 59–86.
- Boyd, R., & Richerson, P. J. (1988). *Culture and the evolutionary process*. University of Chicago Press.
- Brancaccio, G., Kalouptsi, M., Papageorgiou, T., & Rosaia, N. (2023). Search frictions and efficiency in decentralized transport markets. *The Quarterly Journal of Economics*, 138(4), 2451–2503.
- Bresnahan, T. F. (1982). The oligopoly solution concept is identified. *Economics Letters*, 10(1–2), 87–92.
- Butler, J. V., Giuliano, P., & Guiso, L. (2016). The right amount of trust. *Journal of the European Economic Association*, 14(5), 1155–1180.
- Cajal-Grossi, J., Del Prete, D., & Macchiavello, R. (2023a). Supply chain disruptions and sourcing strategies. *International Journal of Industrial Organization*, 90, 103004.
- Cajal-Grossi, J., Macchiavello, R., & Noguera, G. (2023b). Buyers' sourcing strategies and suppliers' markups in Bangladeshi garments. *The Quarterly Journal of Economics*, 138(4), 2391–2450.
- Chassang, S., & Ortner, J. (2019). Collusion in auctions with constrained bids: Theory and evidence from public procurement. *Journal of Political Economy*, 127(5), 2269–2300.
- Chassang, S., & Ortner, J. (2022). Regulating collusion. *Annual Reviews of Economics*, 15, 177–204.
- D'Aspremont, C., Jacquemin, A., Gabszewicz, J. J., & Weymark, J. A. (1983). On the stability of collusive price leadership. *The Canadian Journal of Economics/Revue Canadienne d'Economie*, 16(1), 17–25.
- Dixit, A. K. (2004). *Lawlessness and economics: Alternative modes of governance*. Princeton University Press.
- Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2002). The regulation of entry. *The Quarterly Journal of Economics*, 117(1), 1–37.

- Fahn, M., Bentley MacLeod, W., & Muehlheusser, G. (2023). Past and future developments in the economics of relational contracts. *Journal of Institutional and Theoretical Economics*, 179(3–4), 413–440.
- Forbes, S. J., & Lederman, M. (2009). Adaptation and vertical integration in the airline industry. *American Economic Review*, 99(5), 1831–1849.
- Galenianos, M., & Gavazza, A. (2017). A structural model of the retail market for illicit drugs. *American Economic Review*, 107(3), 858–896.
- Gambetta, D. (1996). *The Sicilian mafia: The business of private protection*. Harvard University Press.
- Genesove, D., & Mullin, W. P. (2001). Rules, communication, and collusion: Narrative evidence from the sugar institute case. *American Economic Review*, 91(3), 379–398.
- Ghani, T., & Reed, T. (2022). Relationships on the rocks: Contract evolution in a market for ice. *American Economic Journal: Microeconomics*, 14(1), 330–365.
- Gibbons, R., & Henderson, R. (2012). What do managers do? Exploring persistent performance differences among seemingly similar enterprises. In R. Gibbons & J. Roberts (Eds.), *The handbook of organizational economics*. Princeton University Press.
- Gibbons, R., & Prusak, L. (2020). Knowledge, stories, and culture in organizations. *AEA Papers and Proceedings*, 110, 187–192.
- Gil, R., & Zanarone, G. (2016). New frontiers in empirical research on informal contracting. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, 172(2), 390–407.
- Gil, R., & Zanarone, G. (2017). Formal and informal contracting: Theory and evidence. *Annual Review of Law and Social Science*, 13, 141–159.
- Gil, R., Kim, M., & Zanarone, G. (2022). Relationships under stress: Relational outsourcing in the U.S. airline industry after the 2008 financial crisis. *Management Science*, 68(2), 1256–1277.
- Green, E. J., & Porter, R. H. (1984). Noncooperative collusion under imperfect price information. *Econometrica*, 52(1), 87–100.
- Igami, M., & Sugaya, T. (2022). Measuring the incentive to collude: The vitamin cartels, 1990–99. *The Review of Economic Studies*, 89(3), 1460–1494.
- Joskow, P. L. (1988). Price adjustment in long-term contracts: The case of coal. *The Journal of Law and Economics*, 31(1), 47–83.
- Ksoll, C., Macchiavello, R., & Morjaria, A. (2021). Electoral violence and supply chain disruptions in Kenya's floriculture industry. *Review of Economics and Statistics*, 105(6), 1335–1351.
- Lafontaine, F., & Slade, M. (2012). Inter-firm contracts. In R. Gibbons & J. Roberts (Eds.), *The handbook of organizational economics*. Princeton University Press.
- Lessing, B. (2021). Conceptualizing criminal governance. *Perspectives on Politics*, 19(3), 854–873.
- Macchiavello, R. (2022). Relational contracts and development. *Annual Review of Economics*, 14, 337–362.
- Macchiavello, R., & Morjaria, A. (2015). The value of relationships: Evidence from a supply shock to Kenyan rose exports. *American Economic Review*, 105(9), 2911–2945.
- Macchiavello, R., & Morjaria, A. (2021). Competition and relational contracts in the Rwanda coffee chain. *The Quarterly Journal of Economics*, 136(2), 1089–1143.
- Macchiavello, R., & Morjaria, A. (2022). Empirical approaches to trust and relational contracts. *International Journal of Industrial Organization*, 84(C), 102849.
- Macchiavello, R., & Morjaria, A. (2023). Relational contracts: Recent empirical advancements and open questions. *Journal of Institutional and Theoretical Economics*, 179(4), 673–700.
- Macchiavello, R., Reardon, T., & Richards, T. J. (2022). Empirical industrial organization economics to analyze developing country food value chains. *Annual Review of Resource Economics*, 14, 193–220.
- Mailath, G. J., & Samuelson, L. (2006). *Repeated games and reputations: Long-run relationships*. Oxford University Press.
- Malcomson, J. M. (2012). Relational incentive contracts. In R. Gibbons & J. Roberts (Eds.), *The handbook of organizational economics*. Princeton University Press.

- McMillan, J., & Woodruff, C. (1999). Interfirm relationships and informal credit in Vietnam. *The Quarterly Journal of Economics*, 114(4), 1285–1320.
- Michler, J. D., & Wu, S. Y. (2020). Relational contracts in agriculture: Theory and evidence. *Annual Review of Resource Economics*, 12, 111–127.
- Mitton, T. (2008). Institutions and concentration. *Journal of Development Economics*, 86(2), 367–394.
- Pinotti, P. (2015). The economic costs of organised crime: Evidence from southern Italy. *The Economic Journal*, 125(586), F203–F232.
- Savona, E. U., & Berlusconi, G. (Eds.). (2015). *Organized crime infiltration of legitimate businesses in Europe: A pilot project in five European countries*. Transcrime, Università degli Studi di Trento.
- Schelling, T. C. (1956). An essay on bargaining. *The American Economic Review*, 46(3), 281–306.
- Simon, H. A. (1951). A formal theory of the employment relationship. *Econometrica*, 19(3), 293–305.
- Smith, A. (1776). *An inquiry into the nature and causes of the wealth of nations*. W. Strahan and T. Cadell.
- Stigler, G. J. (1964). A theory of oligopoly. *The Journal of Political Economy*, 72(1), 44–61.
- Taylor, C. R., & Wiggins, S. N. (1997). Competition or compensation: Supplier incentives under the American and Japanese subcontracting systems. *American Economic Review*, 87(4), 598–618.
- UNODC. (2011). *Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes*. United Nations Office on Drugs and Crime.
- Watson, J. (2021). Theoretical foundations of relational incentive contracts. *Annual Review of Economics*, 13, 631–659.
- Williamson, O. E. (2005). The economics of governance. *American Economic Review*, 95(2), 1–18.
- World Bank. (2016). *Breaking down barriers: Unlocking Africa's potential through vigorous competition policy (English)*. World Bank Group.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Governing Relationships: Theory, Evidence, and Opportunities 25

Ricard Gil and Giorgio Zanarone

Contents

25.1	Introduction	622
25.2	Theoretical Framework	625
25.2.1	The Canonical Model of Relational Incentive Contracts	625
25.2.2	The Clarity Problem and the Role of Formality	634
25.3	From Theory to the Data	636
25.3.1	Surplus Increases in the Shadow of the Future (for Given Contract Provisions)	637
25.3.2	Under Relational Contracting, Responses to and the Design of Court-Enforceable Contracts Differ from Arm's-Length Contracting	639
25.3.3	A Better Contractual Technology May Crowd Out Relational Contracts	640
25.3.4	Relational Contracts and Organizational Structure	642
25.3.5	Future Challenges: Empirical Analyses of the Clarity Problem	643
25.4	Concluding Remarks	645
	References	646

Abstract

In this chapter, we provide a framework to investigate the interaction between the “hardware” (formal contracting) and the “software” (relational contracting) of economic collaborations, both within and between firms. First, we present a simple formulation of the “canonical model” of relational contracting and highlight its key empirical implications. Next, we discuss selected empirical works that shed light on the relevance and testability of the model’s predictions. While significant progress has been made at measuring the complementarity between formal and relational contracts, our investigation highlights several implications

R. Gil (✉)
IESE Business School, Barcelona, Spain
e-mail: rgils@iese.edu

G. Zanarone
Hautes Etudes Commerciales, University of Lausanne, Lausanne, Switzerland
e-mail: giorgio.zanarone@unil.ch

of the model that are consistent with anecdotes yet lack systematic empirical verification. Testing these implications is an exciting opportunity for future scholars in the field.

Keywords

Governance · Formal contracts · Relational contracts · Enforcement · Evidence

25.1 Introduction

What makes self-interested individuals cooperate, both within and across firm boundaries? Economists have developed two answers to this question. On the one hand, incentives to cooperate can be created through carefully designed, farsighted contracts, which a court (or another third party, such as an arbitrator or, in certain societies, a community/customary judge) will enforce.¹ This is the perspective adopted by much of agency theory, contract theory, and transaction cost economics. On the other hand, people can develop trust and long-term relationships that make their promises to cooperate credible even when third-party enforcement is absent or dysfunctional. This perspective, initially championed by legal scholars and sociologists (e.g., Granovetter, 1985; Macaulay, 1963), has been later incorporated into models of self-enforcing agreements, which economists have come to label “relational contracts.”²

The economic literatures on contracts and relationships as mechanisms to govern cooperation are vast and have been extensively reviewed elsewhere (Gil & Zanarone, 2017, 2018; Lafontaine & Slade, 2013; Macchiavello, 2022; Malcomson, 2013). A key question that remains empirically understudied, and to which this essay is devoted, is how these two dimensions of governance interact with each other. This question is especially important in developed economies, where well-functioning courts and third-party enforcement systems make it possible for firms to use both

¹Both transaction cost economics and agency theory acknowledge that most contracts are inherently incomplete and thus cannot directly specify desired cooperation under all possible contingencies. Accordingly, these branches of economics emphasize contractual provisions that realign the parties' incentives, thereby indirectly facilitating or incentivizing cooperation. See Lazear and Oyer (2013) and Lafontaine and Slade (2013) for a comprehensive review of the empirical evidence, respectively, within firms and between.

²While most economists (including ourselves) use the term “relational contracts” as a synonym of self-enforcing agreements, the term was originally coined by legal scholars to define nonstandard agreements that are meant to flexibly govern an ongoing and idiosyncratic relationship (MacNeil, 1978). This definition was also adopted by some prominent transaction-cost economists (Williamson, 1979, 1991). According to this second usage of the term, relational contracts may well be enforced by courts but based on a special body of contract law that aims to fill gaps in the agreement in line with the parties' intentions (Schwartz, 1992). See Gil and Zanarone (2017) and Masten (2021) for a deeper discussion of the different definitions of relational contracting in the literature.

contractual and relational mechanisms to govern their collaborations. Examples of the interaction between the two are ubiquitous and diverse. On the one hand, there are studies showing that even in relationships governed by detailed and court-enforceable contracts, the extent of cooperation is sensitive to the “shadow of the future,” that is, to the existence and value of a long-term relationship between the parties (see, most recently, Barron et al., 2020, on distributor-exhibitor agreements in the Spanish movie industry, and Gil et al., 2022, on major-regional airline contracts in the US airline industry, both of which are extensively discussed below). Moving from interfirm collaborations to employment, leading manufacturers like Lincoln Electric have become famous for jointly using contractual, court-enforceable piece rates and discretionary subjective bonuses to incentivize their employees (Bartlett, 1998; Fast & Berg, 1975). Does the presence of court-enforceable contractual provisions in these interfirm and intrafirm relationships strengthen (and complement) the parties’ trust-based, self-enforcing cooperation promises? If so, how?

On the other hand, recent surveys of business-to-business collaborations in the USA show that even in contexts where litigating disputes in court is impractical, and hence the enforcement of cooperation promises must fully rely on relational mechanisms, firms enter sophisticated formal contracts that are drafted by lawyers and specify performance targets, monitoring mechanisms, and penalties and rewards in great detail (Bernstein & Peterson, 2023; Hadfield & Bozovic, 2016). A similar phenomenon occurs within firms. A company’s chief executive has ultimate authority over all decisions that involve the use of the company’s resources, implying that he or she cannot contractually commit not to exert such authority (Williamson, 1991). Thus, when chief executives delegate spending and decision authority to divisional managers—and sometimes, as at Netflix (Hastings & Meyer, 2020) and Oticon (Foss, 2003), even to middle managers and junior employees—they are implicitly or explicitly making a promise that the latter need to trust, that is, a relational promise. Indeed, as discussed below, there is evidence that delegation of authority from headquarters to divisional managers is more likely in the presence of mutual trust (Bloom et al., 2012). Despite the intrinsically relational and non-contractual nature of authority delegation, company executives spend considerable time crafting detailed and highly formal organizational charts that specify the allocation of decision authority within their organizations. What role do contractual and organizational formalism play in collaborations where the enforcement of cooperation promises is purely relational?

While there are theoretical papers addressing the two sets of questions above (we further elaborate on them below), the number of large-sample empirical studies is remarkably scarce. This paucity of empirical research is in a way not surprising. Obtaining data on contracts and formal governance (across relationships or over time) is difficult due to the private and confidential nature of this information. Measuring relational promises and the shadow of the future in long-term collaborations is even harder because these variables are almost by definition not reported in company documents. Thus, constructing a dataset that combines both sets of information becomes almost a heroic task. The goal of this chapter is to draw upon existing theoretical and empirical research to make the case that, albeit not easy,

investigating the interaction of formal and relational contracts (and governance, more broadly) is both feasible and worthwhile.

We begin by presenting a simple version of the “canonical model” of relational contracts, adapted from Gil and Zanarone (2018). While early theoretical analyses of relational contracts go back to the 1980s (Bull, 1987; Klein, 1980; MacLeod & Malcomson, 1989; Telser, 1981), the modeling of this phenomenon has later converged toward a standard tractable framework developed by the subsequent papers of Baker, Gibbons, and Murphy (1994, 1999, 2002) and Levin (2003). Because most recent models draw upon and enrich the Baker-Gibbons-Murphy-Levin framework (sometimes explicitly stating which of its assumptions they modify or relax), we refer to this framework as the canonical model.

Our model captures the interaction of relational contracts with both formal contracts (as in Baker et al., 1994) and formal organization design (as in Baker et al., 1999) and, thus, connects to all the dimensions of our discussion above. The model suggests three key insights. First, when contractual provisions alone cannot support efficient cooperation, cooperation levels increase in the expected duration and value of future interactions between the parties. This value is lost if promises are broken and/or the relationship terminates so the greater it is, the more the parties can trust that neither of them has an incentive to defect. Importantly, this insight applies both to the case in which cooperation is a supplier’s or employee’s promise to exert productive effort (first part of the model) and to the case in which cooperation is a chief executive’s promise to respect the organizational chart’s allocation of authority (second part of the model).

The second insight from the canonical model is that parties may enter imperfect court-enforceable contracts not because they plan to behave according to their terms but because, given the amount of cooperation that they target, these contracts reduce the portion of such cooperation that must be enforced relationally and, thus, enable the parties to select high levels of cooperation even if the value of their future relationship and their mutual trust are somewhat limited. Moreover, we show that when formal contracts are used to support an overarching relational agreement, their provisions look quite different from those in arm’s-length contracts (i.e., contracts that are supposed to fully describe and dictate the parties’ behavior).

The third insight is that regardless of whether they are court-enforceable or not, formal contractual provisions play an important role in ensuring that relational cooperation promises are clear to both parties, thereby reducing the risk that good-faith misunderstandings are interpreted as breach of trust and lead to the termination of valuable relationships. That is, in the words of two leading scholars in the field, formal contracts may play both a “credibility-building” role (our second insight) and a “clarity-building” role (our third insight) in sustaining superior relational agreements between collaborating parties (Gibbons & Henderson, 2012).

In the second part of the paper, we discuss recent empirical studies in the light of the canonical model of relational contracting. Our purpose in doing so is not to offer a detailed and comprehensive review of the relevant literature (which, as mentioned above, is available elsewhere) but rather to highlight promising data and empirical strategies that have been developed to address some questions on the interaction

between formal and relational contracts and how these data and strategies could be further leveraged to investigate theoretical predictions and questions that are still awaiting empirical scrutiny.

25.2 Theoretical Framework

In this section, we analyze a simple version of the canonical Baker-Gibbons-Murphy-Levin model of relational incentives, adapted from Gil and Zanarone (2018), to articulate the literature's key insights on the interaction between contractual and relational enforcement. Before we proceed, a few observations are in order. First, we use a model (as opposed to verbal reasoning) to precisely lay down the economic mechanisms underlying these insights. The model is therefore directed not so much to economists already versed in the literature on relational contracts but rather to readers (economists and not) who are approaching this literature for the first time. Second, as in any model, there is a trade-off between generality and analytical tractability. While the canonical model (analyzed in Sects. 25.2.1.1, 25.2.1.2, and 25.2.1.3) focuses on a relatively specific contracting problem—namely, incentive provision through pay-for-performance—the underlying economic forces are quite general. We discuss generalizability of the model's insights to other settings in Sect. 25.2.1.4. Lastly, while the canonical model focuses on how formal contractual provisions complement relational ones through the threat of court-enforcement, recent research emphasizes a different and complementary role of such formal provisions—namely, building a shared understanding about the content of the relational agreement. We discuss this “clarity-enhancing” role of contractual formalism in Sect. 25.2.2.

25.2.1 The Canonical Model of Relational Incentive Contracts

25.2.1.1 Environment

Consider a “principal” (P), who hires an “agent” (A) to perform a task for an indefinitely long number of periods. Both parties discount next-period payoffs by the common factor $\delta \in [0, 1]$. A convenient interpretation of δ is as the exogenous probability that, assuming both parties cooperate while in a relationship (in a way that will be clarified momentarily), their relationship survives from one period to another—for instance, because neither party dies, retires, or goes bankrupt. P and A can be equivalently interpreted as parties in an interfirm collaboration (for instance, a buyer and a supplier) or in an employment relationship (for instance, a manager and a subordinate).

P hires A because A's effort e (a real number) produces a valuable contribution to P, y , which we will call P's “individual output” (Gibbons, 2005) or, simply, output. Formally, we assume $y = e$, that is, output increases in A's effort. Exerting effort costs $c(e) \equiv \frac{1}{2}e^2$ to A, who must therefore be compensated for doing so. In principle,

P could agree on a simple contract with A, which specifies the amount of effort to be provided, in exchange for a fixed payment. Unfortunately, it is too costly for P to directly measure A's effort (think of an employee's initiative and creativity, or an offshore supplier's investment in responsible labor practices). Thus, the only way in which P can contractually induce A to provide effort is through an "incentive contract" that links A's compensation to some outcome that is sensitive to his effort.

The natural candidate outcome for such an incentive contract is the output y , that is, A's individual contribution to P. While it is plausible that P observes such contribution, however, it is less plausible that she can describe it and record it in a contract in such a way that a third party with little knowledge of the firm (a court) can assess compliance with the contract's terms. For instance, while a manufacturer might realize through her repeated interaction with various suppliers that supplier A's components are better customized and engineered than those of supplier B, it might be difficult for the manufacturer to objectively demonstrate that B's components have contributed more than A's components to the market success of the final product. We capture this basic idea by assuming that A's output y is observable but non-verifiable: the parties can informally agree on an incentive contract that rewards A based on output but cannot expect a court to enforce such a contract, which must therefore be relational.

While output-based incentive contracts cannot be enforced by courts, there may be variables other than individual output that are sensitive to A's effort and that can be objectively and easily measured (real-world examples are an agent's sales, a unit's or firm's accounting profits, or a supplier's timely delivery). For simplicity, we assume there is one such variable, which we label p (for "measured performance"). Importantly, we assume that while p is sensitive to A's effort, it is a noisy proxy for such effort, in the sense that A may also be able to affect it through an unproductive action m (again, a positive real number), which we call "manipulation." For instance, agents on a sales commission may make inaccurate statements to induce more customers to buy the product, and a supplier incentivized for timeliness may economize on quality control. Formally, we assume $p = e + gm$, where the parameter g (for "gaming") measures the extent of misalignment between individual output and measured performance. Like effort, manipulation also requires some time and dedication, and hence it is costly to A. We assume for expositional simplicity that manipulation has the same cost as productive effort: $c(m) \equiv \frac{1}{2}m^2$.

Given the environment described above, the contractual relationship between P and A proceeds as follows. At the beginning of each period, P offers an incentive contract to A, which specifies A's compensation as a function of output and measured performance: $w(y, p) \equiv s + bp + By$. The term s in A's compensation is a contractual fixed payment (a salary if A is a subordinate employee, an input price if A is a supplier). The term b is a contractual "piece rate" paid for each realized unit of the performance measure p . Lastly, the term B is a relational bonus, which is paid based on P's observation of A's realized individual output, y . Because output cannot be verified by a court (as discussed above), the relational bonus B must be self-enforcing—that is, A must trust P to pay it as specified by the compensation formula.

Thus, the model allows for arm's-length incentive contracts that solely rely on court enforcement ($b > 0$ and $B = 0$), purely relational contracts ($B > 0$ and $b = 0$) that solely rely on self-enforcement, and a combination of the two ($b > 0$ and $B > 0$).

Before analyzing the interaction between contractual and relational incentives, it is useful to briefly discuss, as a benchmark, the ideal allocation of effort for the principal-agent relationship. If there were no contracting frictions (i.e., if effort were directly contractible), P and A would select the "first-best" actions, denoted by e^* and m^* , that maximize the sum of their individual payoffs, or "joint surplus":

$$V(e, m) \equiv y - c(e) - c(m) = e - \frac{1}{2}(e^2 + m^2). \quad (25.1)$$

Clearly, the first-best level of manipulation is zero given that manipulation is costly but does not produce any value: $m^* = 0$. Regarding productive effort, it is efficient for A to exert it up to the point where its marginal contribution to the joint surplus (derivative of y with respect to e) equals its marginal cost (derivative of $c(e)$), that is, $e^* = 1$. The joint surplus resulting from these first-best actions is therefore $V^* \equiv \frac{1}{2}$.

25.2.1.2 Arm's-Length Contracting

To build some intuition on how contracting frictions (the fact that effort and output are non-contractible and that the performance measure p is manipulable) affect incentive provision, we begin with analysis of the case where P can only use contractual incentives: $b > 0$ and $B = 0$. In this case, if A accepts the compensation offered by P, she chooses effort and manipulation to maximize his own utility, that is, the difference between compensation and the actions' costs:

$$u = s + bp - c(e) - c(m) = s + b(e + gm) - \frac{1}{2}(e^2 + m^2). \quad (25.2)$$

The solution of A's problem is given by the effort and manipulation response functions $e^F(b)$ and $m^F(b)$, which equate A's marginal benefit under a given contractual piece rate b (i.e., the derivatives of the variable compensation bp with respect to e and m , respectively) to the marginal cost (defined as above):

$$e^F(b) \equiv b, m^F(b) \equiv bg. \quad (25.3)$$

Note that, because contractual incentives must be based on the distortionary performance measure p , an increase in the piece rate b raises not only A's effort but also her manipulation, and more so the greater the misalignment between measured performance and A's individual output, as measured by parameter g .

Anticipating A's response to contractual incentives, P chooses the payment terms s and b to maximize her own profit (the difference between output and compensation),

$$\pi = y - s - bp, \quad (25.4)$$

subject to the constraint that A's total compensation is high enough to cover his costs (or else A would reject P's contractual offer). This condition, which can be stated mathematically as $u \geq 0$, is known in the economics of contracts as "participation constraint." Clearly, it is optimal for P to offer the minimum fixed payment that satisfies A's participation constraint, which given (25.2) is:

$$s^F \equiv c(e^F(b)) + c(m^F(b)) - bp.$$

After substituting s^F into (25.4), P's problem simplifies to choosing the piece rate to maximize the joint surplus, given A's anticipated response to the piece rate, that is, P maximizes:

$$V(e^F(b), m^F(b)) = e^F(b) - c(e^F(b)) - c(m^F(b)).$$

After substituting the response functions from (25.3), P's problem can be further rewritten as that of maximizing:

$$b - \frac{1}{2}b^2(1 + g) \quad (25.5)$$

The solution is given by the piece rate that equates P's marginal benefit (derivative of output, the first term in (25.5)) to her marginal cost (derivative of A's cost of responding to incentives, the second term in (25.5)), that is:

$$b^F \equiv \frac{1}{1 + g^2},$$

which generates joint surplus $V^F \equiv \frac{1}{2(1+g^2)}$.

When output and measured performance are perfectly aligned, such that A cannot increase the latter through manipulation ($g = 0$), the optimal arm's-length contract features high-powered incentives ($b^F = 1$) that elicit first-best actions from A (substituting $b = 1$ into (25.3), we obtain $e^F = 1$, $m^F = 0$). As the misalignment parameter g grows, the contractual piece rate decreases ($b^F < 1$), and the (now second-best) optimal contract sacrifices some productive effort to contain unproductive manipulation.

25.2.1.3 Optimal Combination of Contractual and Relational Incentives

Let us now consider the more general case in which P can use both contractual piece rates (b) and relational bonuses (B) to incentivize A—that is, P offers compensation $w(y, p) \equiv s + bp + By$. If A accepts, she chooses the two actions to maximize his payoff given the contract, that is:

$$u = s + bp + By - c(e) - c(m) = s + b(e + gm) + Be - \frac{1}{2}(e^2 + m^2). \quad (25.6)$$

As usual, the solution is given by the actions that equate A's marginal benefit and cost. For manipulation, these are the same as under arm's-length contracting (because the relational bonus is conditional on output, which does not depend on manipulation). Thus, A's manipulation continues to be $m^F(b)$. For productive effort, A's marginal cost is the same as before, but the marginal benefit is now larger because if A exerts one more unit of productive effort, she now increases his compensation through both the piece rate and the relational bonus:

$$e^R(b, B) \equiv b + B.$$

Anticipating A's response (and taking into account, as before, that the optimal salary is the minimum value of s that induces A to accept the agreement), for a given relational bonus B , P chooses the piece rate b to maximize the joint surplus:

$$e^R(b, B) - c(e^R(b, B)) - c(m^F(b)). \quad (25.7)$$

After substituting A's response functions into the surplus function above and equating P's marginal benefit and marginal cost of b , the optimal piece rate for a given relational bonus is:

$$b(B) \equiv \frac{1 - B}{1 + g^2}. \quad (25.8)$$

It is also useful to define the joint surplus that P appropriates as a function of the piece rate in (25.8) and the relational bonus B :

$$V^R(B) \equiv e^R(b(B), B) - c(e^R(b(B), B)) - c(m^F(b(B))).$$

While our analysis is not complete yet, the piece rate formula in (25.8) already provides some insights. If output and measured performance are completely misaligned (g tends to infinity), or if $B = 1$, so that the relational bonus alone can induce A to exert first-best effort without manipulation, P does not benefit from adding a distortionary contractual piece rate to A's compensation: $b(B = 1) = 0$. Aside from these limit cases, adding a contractual piece rate to the relational bonus is typically beneficial: $b(B) > 0$, decreasing in the misalignment parameter g and in the size of the relational bonus B .

To complete our analysis of the interaction between contractual and relational incentives and draw empirical implications from the model, we now study P's choice of the relational bonus B . Recall that, because P's promise to pay B cannot be enforced by a court, the relational bonus must be self-enforcing. In relational contracting models (which are special cases of repeated game models), self-enforcement arises from the threat of terminating the relationship: both parties agree at the outset (explicitly or implicitly) that if P ever fails to pay the promised bonus, A will never believe any of P's promises in the future, and thus P will revert to the arm's-length incentive contracts analyzed in Sect. 25.2.1.2 (which do not

require trust and credibility to be enforced), earning the joint surplus V^F forever after. This threat to terminate the relational contract ensures self-enforcement if in any given period, P's present discounted gains from the relational contract are larger than the relational bonus she is tempted to default on in the current period. Taking into account that P's per period gain from honoring the promised bonus is the difference between surplus under relational contracting, $V^R(B)$, and surplus under arm's-length contracting, V^F , and that P always discounts next-period gains by δ , the condition for a relational bonus B to be self-enforcing can be formulated mathematically as:

$$\delta[V^R(B) - V^F] + \delta^2[V^R(B) - V^F] + \delta^3[V^R(B) - V^F] + \dots \geq B,$$

which can be rewritten as

$$\frac{\delta}{1 - \delta} [V^R(B) - V^F] \geq B. \quad (25.9)$$

Thus, P's problem is to choose the relational bonus B to maximize the per period joint surplus $V(B)$, subject to self-enforcement condition (25.9).

The problem above highlights the trade-off P faces when choosing how much to rely on the relational bonus to incentivize A. On the one hand, increasing B (and thus decreasing the contractual piece rate b) is attractive for P, as it allows her to elicit more productive effort without having to also compensate A's unproductive manipulation that the piece rate contract incentivizes. On the other hand, promising a larger relational bonus B increases P's present temptation to renege and, hence, makes it harder to keep the relational contract self-enforcing (the self-enforcement condition (25.9) becomes harder to satisfy as B grows).

If the parties value the future of their relationship enough (because the exogenous probability of continuation, δ , or the relationship's per period value, $V^R(B)$, is high, or alternatively, because the "fallback option" of reverting to arm's-length contracting, V^F , is not too attractive), then a relational bonus high enough to induce A to exert the first-best effort level is self-enforcing. Thus, there is no need for distortionary piece rates—the optimal compensation agreement is purely relational: $B^R = 1$, and $b^R = 0$. For lower value of the parties' future relationship, the self-enforcement constraint (25.9) becomes binding—that is, a bonus that elicit first-best effort is not credible for A, and the highest relational bonus consistent with (25.9) is $B^R < 1$, the largest positive number that satisfies (25.9) with equality. In such a scenario, adding a contractual piece rate $b^R > 0$ to the relational bonus becomes valuable (despite the wasteful manipulation activity it induces) because given the limited credibility of P's promises, the piece rate allows P to elicit a higher effort from A than the relational bonus alone.

It is interesting to contrast these findings with the common view of collaborative relationships as "hand-shake agreements" (Macaulay, 1963). While economic modeling does support Macaulay's insight that collaborating parties do not "play by the book," relying on contracts to dictate their behavior, it also suggests that

court-enforceable contracts play an important role in making relational collaboration credible (Klein, 2000).

25.2.1.4 Empirical Implications

The simple model developed above suggests specific predictions on the complementarity between contractual and relational incentives that one can test empirically. The first prediction is that among similar relationships governed by the same contract terms, those with a longer shadow of the future (higher levels of δ in self-enforcement condition (25.9)) will be more cooperative (higher effort e , lower manipulation m) and more productive (higher output and surplus). If, instead, these collaborations did not rely on relational incentives to complement contractual ones (i.e., if they were pure spot-market, arm's-length relationships), the shadow of the future should not affect their productivity *holding the contractual incentives constant*. One can test this prediction by regressing effort, cooperation, or surplus in collaborative relationships governed by similar contract terms on measures of the shadow of the future. We discuss examples of how to develop such an empirical strategy in the next section.

A second prediction of the model is that parties respond to contractual incentives differently when these are used in combination with relational ones. One can see that from our formal analysis above by noticing that under a given contractual piece rate b , A chooses effort $e^F(b) = b$ if no relational bonus is present, whereas A chooses the higher effort $e^R(b, B) = b + B$ if a relational bonus $B > 0$ is also present. One could test this prediction on the complementarity between contractual and relational incentives by checking whether the observed cooperation in a collaboration is systematically profitable or unprofitable under the court-enforceable contract terms used by the parties. Finding that observed cooperation is unprofitable under the contract would be consistent with relational incentives playing a complementary role. We elaborate further on this empirical strategy in the next section.

A third prediction of the model is that relational incentives (B) affect the parties' choice of contractual incentives (b). To see this point analytically, recall from Eq. (25.8) that the optimal contractual piece rate is $b(B) \equiv \frac{1-B}{1+g^2}$, which varies (in our setting, decreases) in the relational bonus B . Thus, collaborations characterized by different levels of the optimal relational bonus (for instance, due to different shadows of the future, as measured by δ) will select different contractual piece rates. The insight that the design of court-enforceable contracts is affected by the presence of an overarching relational contract is quite general and has been analyzed in several refinements or applications of the canonical model.³

³See, for instance, Battigalli and Maggi (2008) and Kvaløy and Olsen (2009) for general models of the interaction between contractual and relational provisions in incomplete agreements. See also Zanarone (2013) for a model of the interaction between relational contracts and the contractual allocation of decision rights in franchising. Finally, Argyres et al. (2020) provide for a model of the interaction between relational contracts and outsourcing scope in buyer-supplier relationships.

A fourth and last prediction of the canonical model pertains to the relationship between contractual technology (monitoring, legal systems, courts) and relational contracting. Specifically, the model predicts that when relational incentives are used, improvements in the contractual technology may increase or decrease surplus, that is, contractual technology and relational contracts may be substitutes or complements (Baker et al., 1994, 2002). To see this point in our formal analysis, consider a reduction in the misalignment parameter g . If P were solely relying on the contractual piece rate b to incentivize A , a reduction in misalignment would reduce unproductive manipulation ($m^F(b) = gb$ decreases in g). This reduction in manipulation would encourage P to raise the piece rate, which in turn would increase productive effort and surplus (joint surplus under the optimal arm's-length piece rate contract, $V^F = \frac{1}{2(1+g^2)}$, decreases in g). When relational incentives are also used, however, arm's-length contracting is P 's fallback option after reneging on the relational bonus. Thus, an increase in V^F (due to the lower misalignment g) reduces P 's net long-term gain from the relational contract (the left-hand side of condition (25.9)), making it harder to keep a given relational bonus B self-enforcing.

Therefore, the overall effect of a reduction in misalignment on surplus is ambiguous and non-monotonic. While initial improvements in the contractual technology (starting from very low levels) may increase productive effort and surplus (i.e., the manipulation-reduction effect dominates the fallback-option-enhancement effect), further improvements may backfire: the possibility for P to revert to arm's-length contracting may become too attractive, and as a result, relational bonuses that were self-enforcing under poorer contractual technologies may become unsustainable (i.e., may no longer satisfy self-enforcement condition (25.9) above). To put it more colorfully, a decent-but-not-perfect contractual technology may crowd out superior relational contracting. This result provides a possible microeconomic foundation for legal arguments in favor of a pure private ordering in business transactions, whereby parties willingly "opt out of the legal system" and commit themselves to solely rely on relational enforcement mechanisms (Bernstein, 1992). Despite its important influence on relational contracting theory, this implication of the canonical model has so far escaped serious empirical scrutiny, as discussed below.

25.2.1.5 Reinterpreting the Canonical Model: Adaptation and Organization Design

The insights of the canonical model of relational incentive contracts, as discussed above, extend to different but related contracting problems. For instance, several models analyze the role of relational contracts in settings where court-enforceable contracts are incomplete both *ex ante* (parties cannot specify and commit to desired actions under all possible future contingencies) and *ex post* (once it becomes clear what actions should be taken, it is not possible to fill contractual gaps via spot negotiation, for instance, because there is insufficient time). Examples of such settings are franchising networks, supply chains, strategic alliances, and other organizational arrangements that are "hybrids" between spot market contracting

and intrafirm hierarchies (for a comprehensive review and critical discussion of the literature on hybrids, see Ménard, 2013).

For instance, Baker et al. (2011) develop a general model of relational contracts in which multiple parties agree to adapt non-contractible decisions to realized contingencies and explore which contractual allocation of the authority to make those decisions among the parties supports the best relational contract. Zanarone (2013) develops a related model of distribution networks in which decisions are contractible, but pro-distributor laws constrain the set of feasible contracts, and studies the interaction between relational contracts and the contractual allocation of decision rights. Argyres et al. (2020) model adaptation in buyer-supplier relationships and show that broad-scope contracts, whereby a given buyer concentrates its purchases into one supplier, enable the parties to sustain more efficient relational adaptation, compared to narrow-scope contracts. Like the canonical model discussed above, these models of relational adaptation find that contractual and relational provisions should coexist and that the design and performance of court-enforceable contracts is affected by the presence of an overarching relational contract.

A second area toward which the canonical model of relational contracts has been extended is organization design and particularly the choice between centralization and delegation of decision authority in organizations. Firms sometimes incentivize their employees (or suppliers) through an appropriate organization design rather than through incentive contracts. Aghion and Tirole (1997) made this point clear by showing that if a “boss” delegates to a “subordinate” the authority to implement a decision, she can incentivize the subordinate to exert creative efforts that may ultimately benefit the boss.⁴ While Aghion and Tirole (1997) make the simplifying assumption that a boss can contractually commit to delegate authority, in most organizations, delegation (and, more broadly, any configuration of the organizational chart) is the boss’s discretionary decision and, thus, can always be retracted. In other words, delegation of authority is a relational promise, pretty much like incentive bonuses in the canonical model. Baker et al. (1999) developed the first model of relational delegation, studying how the expected duration and value of the relationship between a boss and a subordinate affects the credibility of the boss’s promise to delegate. This model shares one of the insights we derived from the canonical model above: as the expected duration and value of the boss-subordinate relationship increases, the promise to delegate authority to the subordinate becomes more credible and, thus, the boss can elicit greater levels of initiative and effort from the subordinate.

Drawing upon Baker et al. (1999), Li et al. (2017) develop a model showing that organizations may fail to efficiently recentralize previously delegated authority for fear that employees will interpret centralization as a breach of trust and quit the firm.

⁴While Aghion and Tirole (1997), like us in this section, study the incentive-provision function of delegation, organizational economists have explored other important functions such as coordination (Garicano, 2000), adaptation (Dessein, 2002), and the interaction between the two (Alonso et al., 2008; Rantakari, 2008). See Bolton and Dewatripont (2013); Gibbons et al. (2013); Garicano and Van Zandt (2013); and Mookherjee (2013), for extensive reviews of this literature.

Rantakari (2023) develops a “radical” model of relational organization design in which even the boss’s authority over her/his subordinates (which is assumed to be contractual in both Aghion & Tirole, 1997, and Baker et al., 1999) must be part of a self-enforcing relational agreement. In this agreement, centralized authority arises in equilibrium when the employee makes a self-enforcing promise to follow the boss’s instructions (without attempting to investigate the matter and second-guess the boss).

25.2.2 The Clarity Problem and the Role of Formality

Models of relational contracts (including the one presented here) make the simplifying assumption that parties agree at the outset on what constitutes cooperation or defection (under all contingencies). More recent papers (Gibbons & Henderson, 2013; Helper & Henderson, 2014) have questioned this assumption, pointing out that like court-enforceable contracts, relational contracts are also “incomplete.” This would mean that parties cannot foresee at the outset all the contingencies their relational contract will have to cover as time goes by. Additionally, language ambiguities and imperfections may further limit the parties’ ability to reach a common understanding and interpretation of what actions their agreement prescribes.

This problem of conflicting interpretations, also known as the “clarity problem” (Gibbons & Henderson, 2012), appeared to be present at First Boston in the 1990s: while the company’s new owner (Credit Suisse) believed bankers’ bonuses should be reduced in a year of poor financial results, First Boston bankers believed that, because they were as talented and hardworking as their peers at other banks, they should receive similar bonuses regardless of profitability differences. After receiving bonuses below their expectations for two consecutive years, most top bankers at First Boston quit the firm (Stewart, 1993). Another example of the clarity problem, in the context of organizational design rather than contracting (see our discussion above), is the “Spaghetti organization” experimented by Oticon, a Danish producer of hearing aid equipment, in the 1990s (Foss, 2003). Oticon decentralized decision-making at the worker level to foster innovation by allowing workers to form teams, compete for human capital within the organization, and pursue the projects they desired. After a promising start, duplication of effort drove the organization to overrule several decisions made by workers. Importantly, while Oticon’s top management viewed such overruling as perfectly consistent with the decentralized organizational structure they created, employees interpreted it as breach of a promise. Foss (2003) reports that, following the overruling, several employees lost morale and motivation and some of them decided to quit the firm.

As the First Boston and Oticon examples illustrate, the clarity problem is especially consequential in relational contracts because it may lead not only to occasionally inefficient decisions but also to the complete breakdown of valuable collaborations. To understand this point, recall that in a relational contract, promises of cooperation are credible only if they are backed by a threat to terminate the

relationship in the event of defection. For instance, in our relational incentive model, both parties believe that if P fails to pay the bonus when A has produced the agreed-upon output, A will no longer trust P's promises in the future, and relational incentives will never again be sustainable. While this mutually understood threat of termination is useful to keep relational contracts self-enforcing, it may lead to unnecessary and wasteful termination when a clarity problem is present. Imagine that P and A in our model disagree on the definition and/or interpretation of output. Then, it could happen that after A chooses his effort, P "sees" low output, $y = \underline{y}$, and hence pays a low bonus, $B\underline{y}$, whereas A "sees" high output, $y = \bar{y}$, and thus expects a high bonus, $B\bar{y}$. When that is the case, A may interpret P's payment of $B\underline{y}$ (instead of $B\bar{y}$) as defection and terminate the relational contract in response. Then, despite the fact that nobody has willingly defected, P and A will be forced to revert to arm's-length contracting, and earn low surplus, in all subsequent periods.

Given the relationship-disruption risks posed by the clarity problem, a frontier research question for economists and social scientists interested in this phenomenon is how contractual language and design together with management practices could help parties to reach a shared framing of the relationship and a shared interpretation of mutual obligations. For instance, recent papers study corporate culture as a set of shared understandings among collaborating parties and the role of organizational leaders in moving from one culture to another (Gibbons et al., 2021a, b).

An interesting case study of clarity-enhancing management practices is Netflix, which revolutionized the entertainment industry and attracted hundreds of millions of subscribers in over 190 countries. In a recent book (Hastings & Meyer, 2020), cofounder Reed Hastings explains that because the company promotes the autonomy of workers and managers—from deciding how to travel to a business event to deciding whether to hire an expensive consultant—it has developed a principle that limits unnecessary or opportunistic decisions: "When spending company money, always act in Netflix's best interest." Employees who are found to be in breach of this principle can be fired. Hastings soon realized that the definition of "Netflix's best interest" is highly incomplete: how can employees know in each unforeseen situation what course of action the company's executive consider to be consistent with the principle? Hastings explains that Netflix has solved this clarity problem by asking its managers to provide extensive history-based examples to new hires, which clarify what the company considered to be its "best interest" in past circumstances, and hopefully offer guidance on what such best interest could be in not-yet-foreseen future circumstances that share some similarities with past ones. As examples accumulate, clarity should build up. Hastings also notices how managers typically do not fire subordinates after behaviors that appear to be against the company's interest—they instead give those subordinates the benefit of the doubt, explaining why their behavior was inappropriate, and resort to the threat of termination only after the subordinate has been long enough at Netflix, such that clarity concerning the "best interest" principle can be more safely presumed.

A few recent papers at the frontier between law, economics, and sociology have also explored how contractual formalism may help parties in a collaborative

relationship to achieve clarity. For instance, Hadfield and Bozovic (2016) suggest that because formal and legalistic language is explicit and clear, it may enable the parties to agree on a common classification of contingencies. This may explain why formal agreements, often drafted by lawyers, are commonly used even in relationships where the threat of litigation is not credible, and hence formalism should not serve the purpose of facilitating courts' verification and enforcement. Bernstein and Peterson (2023) similarly argue that long-term buyer-supplier contracts include a variety of "managerial provisions" that appear to clarify to the parties themselves (rather than to courts) how to define good performance, and what supplier's conducts deemed unsatisfactory by the buyer may lead to termination or forgiveness. A clarity role of contractual formalism is also invoked by Frydlinger and Hart (2024), who argue that by including formal guiding principles (such as integrity and loyalty) in their contract, parties frame their relationship as one of mutual collaboration, thus activating some innate principles of mutual altruism and cooperation that would remain inert in the absence of such provisions.

Besides contracts, the clarity-building role of formalism can also be seen in the context of organization design. Clearly framed organizational charts and decision-making processes may ensure that managers and employees in a firm develop a shared understanding of who makes what decisions in each contingency. This, in turn, may prevent employees from interpreting a manager's decision to correct or reject some of their proposals as the breach of a promise of delegation, as in the Oticon case discussed above. In fact, Oticon itself is a good illustration of how a more formalized organizational structure may help firms to sustain a relationally agreed decision-making process. After employees lost motivation due to the frequent overruling of their proposals by the upper management, Oticon formalized and restructured the scope of employees' delegated authority. On the one hand, under the new structure, employees' projects would be routinely approved, and hence there was no ambiguity left on the existence of a promise to delegate. On the other hand, projects would have to be chosen within formal company guidelines and overseen by centrally appointed team leaders. Thus, employees would be clearly informed about the fact that there are limits to delegation under the agreed-upon organizational structure and what those limits are.

25.3 From Theory to the Data

A classic study by Poppo and Zenger (2002) provides the first evidence that court-enforceable contracts and relational contracts complement each other. Using survey instruments to measure the extent and complexity of contracts and the strength of relational governance in buyer-supplier dyads, they find that more detailed and complex formal contracts are used in relationships that the parties perceive as more collaborative, and trust based. While this is an important pioneering study, it leaves an important question open: which economic mechanism drives the co-presence of contractual and relational provisions?

The canonical model developed in Sect. 25.2 focuses on facilitation of self-enforcement as the key mechanism underlying the complementarity that studies like Poppo and Zenger (2002) uncover. As discussed above, the model also provides predictions that allow one to test for this mechanism. In this section, we discuss (1) a selection of empirical papers that provide insight on how to test these predictions and (2) research opportunities arising from untested implications or extensions of the canonical model as well as the clarity problem discussed in Sect. 25.2.1.4 above. Before we proceed, it is important to acknowledge that the list of empirical papers discussed here is not (and does not intend to be) an exhaustive review of the literature (such type of reviews, as discussed in the introduction, can be found in Gil & Zanarone, 2017, 2018, and Macchiavello, 2022). Instead, we selected papers that in our view address most clearly and directly the theoretical insights we aim to shed light on. Moreover, while we cite some early papers (see right below!), our discussion primarily focuses on recent work that is not extensively covered in the existing reviews.

We organize our discussion of the empirical evidence below around each of the theoretical predictions highlighted in Sect. 25.2.1.4.

25.3.1 Surplus Increases in the Shadow of the Future (for Given Contract Provisions)

The first prediction of the canonical model is that *even when collaborations are governed by court-enforceable contracts (and holding those contracts fixed)*, their performance and surplus increase in the present discounted value (hereafter, PDV) of the future relationship between the parties. Finding evidence in support of these predictions would be a way to empirically test for the interaction and complementarity between court-enforceable and relational contracts.

The first paper providing a rigorous test of this prediction (i.e., one based on plausibly exogenous variation in the PDV of contractual relationships) is Gil and Marion (2013).⁵ They study how contractors' relationships with subcontractors affect the competitiveness of their bids for procurement contracts (a proxy for the generated surplus). The importance of contractor-subcontractor relationships in this industry makes it an ideal setting to study informal agreements (Eccles, 1981). Using data for 10 years of highway procurement contracts from Caltrans (California Department of Transportation) auctions, they use the number and value of local projects (at the construction district level) auctioned by Caltrans over the next calendar year as an exogenous proxy for changes in the PDV of contractor-subcontractor relationships. This is an exogenous proxy because projects are offered

⁵Earlier papers, some of which we discuss below (Corts & Singh, 2004; Kalnins & Mayer, 2004), have used past interactions as a proxy for future interactions. A problem with this approach is that past interactions may also proxy for mutual learning. Other papers (e.g., Poppo et al., 2008) have used endogenous measures such as the parties' expectation on the duration of their business relationship.

when highway and road repairs are needed, and not based on the locally available pairs of contractors and subcontractors. Gil and Marion (2013) first notice that contractors and subcontractors do use court-enforceable contracts, which are disclosed as part of the contractor's auction bids. Then, they show that despite the presence of these contracts, the shadow of future interactions between contractors and subcontractors (proxy for the PDV of their relationship) significantly improves performance. Then, they show that, holding past interactions (and hence accumulated mutual learning) constant, contractors post lower bids in highway procurement auctions when more and more valuable forthcoming projects are scheduled (shadow of the future). Because a contractor interacts with several subcontractors during the temporal span of the data, and because there are several contractors in any given construction district, the regression analysis in Gil and Marion (2013) can include contractor and construction district fixed effects—that is, they hold constant all possible time-invariant contractor and location characteristics, including the boiler-plate court-enforceable contract used by a contractor with its subcontractors.

An important methodological contribution to the empirical literature on relational contracting is Macchiavello and Morjaria (2015). This paper does not study the interaction between contractual and relational provisions and is covered extensively in a companion chapter of this handbook. Nevertheless, we briefly discuss it here because its methodology has been used by subsequent papers studying the interaction problem, as discussed below. Macchiavello and Morjaria (2015) analyze a rich dataset of contracts between Kenyan rose exporters and buyers. Their setting allows them to compute the exporters' temptation to renege on their informal agreements with buyers as the stipulated quantity times the spot market auction price of roses. Because self-enforcement implies that the PDV of a given buyer-exporter dyad must be at least as large as the reneging temptation (see, for instance, self-enforcement condition (25.9) in our model above), Macchiavello and Morjaria (2015) use their estimates of the reneging temptation as lower bounds for the PDV. To validate their interpretation of the reneging temptations as lower bounds for the PDV, they provide evidence that following an unforeseen increase in the spot market price, informal contracts between exporters and buyers proportionally reduce the stipulated quantity, such that the self-enforcement condition is restored. This "lower-bound approach" to measuring the PDV of contractual relationships is potentially replicable in other settings because present reneging temptations, unlike future interactions, can often be measured based on market and contractual data (such as the rose spot market price and quantity data used by Macchiavello & Morjaria, 2015).

Gil et al. (2022) use this methodology to study the collaborations between US major airlines and their outsourced regional partners around the 2008 financial crisis. They show that although majors and regionals sign long and detailed formal contracts, the extent of their collaboration around the crisis is importantly affected by the PDV of their future relationship. In particular, Gil et al. (2022) observe that the 2008 financial crisis reduced profitability in the airline industry (as in many other industries), calling for a downsizing of the majors' route portfolios. Then, they show empirically that majors selectively cut regional routes from the low-PDV relationships while preserving those in the high-PDV relationships. This result is robust to

the inclusion in their empirical specifications of route (market) and major airline fixed effects—that is, Gil et al. (2022) look at the effect of PDV on the resilience of route portfolios while holding constant all possible characteristics of the focal route and, most importantly, all possible characteristics of the major airline, including the boilerplate court-enforceable contract it uses with its regional partners.

To measure the PDV of major-regional relationships, Gil et al. (2022) use the lower-bound approach developed by Macchiavello and Morjaria (2015) and discussed above. We briefly describe their implementation as that can be insightful for empirical researchers interested in applying the lower-bound approach to other relational contracting settings. As noted by Forbes and Lederman (2009), a key non-contractible cooperative decision for major and regional airlines is the exchange of landing slots during the slot rationing that follows inclement weather (70% of winter days at several major airports in the USA). Gil et al. (2022) assume that because slot exchanges between majors and outsourced regionals are non-contractible, they must be governed by relational agreements. They proxy the regionals' maximum temptation to renege on such agreements (i.e., the regional's present savings from denying slots to the major in the contingency in which inclement weather strikes all airports served by the partnership) by aggregating bad weather conditions (rainfall and snow) across all routes in the major-regional partnership. Similar to Macchiavello and Morjaria (2015), Gil et al. (2022) use their proxy for the regional's reneging temptation as a lower bound for the PDV of the major-regional relationships because in a self-enforcing relational contract, the PDV must be at least as large as the reneging temptation.

25.3.2 Under Relational Contracting, Responses to and the Design of Court-Enforceable Contracts Differ from Arm's-Length Contracting

A few early papers provide evidence consistent with the prediction that the court-enforceable provisions one observes under relational contracting differ from those that one should observe under arm's-length contracting. For example, Kalnins and Mayer (2004)—in the context of IT service procurement—show that downstream firms rely less on formal fixed-price terms to motivate their suppliers to cut costs when they are engaged in repeated interactions with the suppliers. Similarly, in a study of offshore oil drilling contracts, Corts and Singh (2004) find that projects are less likely to be governed by a turnkey contract (akin to fixed-price) than by a day-rate contract (akin to cost-plus) when the oil company and the driller have worked together in the past. More recently, Ryall and Sampson (2009) study joint technology development contracts in the telecommunications and microelectronics industries and find that court-enforceable provisions, including penalties, are more frequent and detailed when the firm and the partner interact more frequently. A common feature of these studies is that the extent of relational contracting is proxied by past interactions between the parties rather than measures of the shadow of the future. This is somewhat problematic given that the self-enforcement mechanism

underlying relational contracting relies on the latter. A clear opportunity for future researchers is therefore to leverage on the methodologies discussed in Sect. 25.3.1 above to develop tests for how the shadow of the future affects contractual design.

A recent paper by Kosova and Sertsios (2018) uses a different approach to study how relational contracting affects the design of court-enforceable contracts in the context of hotel franchising. Rather than measuring the extent of relational contracting via repeated interactions, they measure it through the franchisor's difficulty of monitoring franchisees, and hence of enforcing their desired performance contractually. They find that hotel franchisors grant upper-tier franchise contracts (and higher future quasi-rent rents that can be lost upon termination) to locations that are farther from the headquarters and hence harder to monitor. Their interpretation of the results is that as monitoring costs increase, a relational contract requiring franchisees to exert effort may cease to be self-enforcing, and this, in turn, induces franchisors to raise the franchisees' quasi-rents to keep the relational contract self-enforcing.

Barron et al. (2020) provide rare evidence on a related implication of our model—namely, that parties in a relational contract cooperate beyond the letter of their court-enforceable agreement. They study contracts between movie distributors and a large Spanish exhibitor, which are governed by formal revenue sharing provisions specifying the share of each movie reel's revenues that is kept by the exhibitor in any given week the movie is shown. Barron et al. (2020) find that when the exhibitor shows movies whose continuation would be unprofitable given their formally contracted revenue share and box office turnouts, the distributor rewards the exhibitor by unilaterally and retroactively renegotiating the exhibitor's revenue share upward after the movie has ended its run at its theaters. This result indicates that the exhibitor responds to relational incentives (the discounts) rather than contractual ones (the revenue share) in making its movie-showing decisions. It also indicates that to sustain relational contracts, distributors initially select contract terms that would not be optimal in the absence of relational contracting: once a movie ends its run, a retroactive increase in the revenue share clearly cannot affect the exhibitor's contractual incentives to show that movie.⁶

25.3.3 A Better Contractual Technology May Crowd Out Relational Contracts

A key challenge in testing the “crowding out” prediction of the canonical model is how to measure the quality of arm's-length contracting (the fallback option two

⁶Moreover, Barron et al. (2020) also proxy the lower bound of the PDV of distributor-exhibitor relationships through the maximum discount granted by the exhibitor in the first half of the sample period. They find that the exhibitor is more likely to keep showing high-risk movies when the relationship with the distributor has higher PDV. This result further supports the hypothesis that distributors and exhibitors govern their relationships through a combination of formal and relational contracts.

parties would revert to if their relational agreement broke down). A natural measure is the quality of third-party enforcement, that is, the speed and reliability with which contracts are enforced in a given country or industry at a certain point in time. All else equal, higher-quality third-party enforcement should increase the surplus two parties can achieve by relying on arm's-length contracts that solely rely on court enforcement. Several studies have attempted to measure variations in court quality both across countries (Djankov et al., 2003; Johnson et al., 2002) and within countries (Berkowitz & Clay, 2006; Choi et al., 2008). In addition to developing contemporary measures of court quality based on the flexibility and low formalism of court proceedings, Djankov et al. (2003) have also argued that the common-law legal tradition exhibits lower formalism than the civil-law tradition due to exogenous historical reasons. Accordingly, they have proposed to use common-law legal origin as an instrumental variable for contemporary court quality that may allow to identify the causal effect of court quality on outcomes of interest. The historical analysis underlying this approach rests on strong assumptions and has been criticized (Arruñada, 2007; La Porta et al., 2007).

Despite the possibility to proxy for variations in the value of arm's-length contracting, we are not aware of papers that have developed an empirical strategy to test the crowding out hypothesis. Two studies that provide interesting insights (albeit without developing a full test) are Antràs and Foley (2015) and Michler and Wu (2020). Antràs and Foley (2015) study a large US frozen food exporter's provision of trade credit to its international clients. They find that repeated interactions between the exporter and a given client (a proxy for relational contracting) increase trade credit but only if court quality in the client's country is low enough. Their evidence is consistent with crowding out: when arm's-length contracting becomes less dysfunctional, a too-attractive fallback option for the client may prevent the buyer from offering superior relational trade credit agreements, and as a result, marginal improvements in relational enforcement no longer affect trade credit provision.

Using survey data on groundwater irrigation contracts in Bangladesh, Michler and Wu (2020) find that households rely less on fixed-price contracts and more on output sharing contracts, in villages where the punishment of contract breach is more severe (social ostracism instead of negotiated symbolic penalties or fines). Since the main problem in groundwater irrigation is to ensure consistent and reliable water delivery, Michler and Wu (2020) interpret fixed-price agreements—which do not provide incentives for water delivery—as relational incentive contracts (in the sense that water delivery under these agreements must be incentivized through a self-enforcing arrangement), and crop sharing agreements, which do provide such incentives, as court-enforceable incentive contracts. Under this interpretation, their results are consistent with crowding out: improvements in the value of arm's-length contracting (more severe punishments in case of breach of the output sharing obligation) reduce the parties' reliance on relational incentives.

While the results in Antràs and Foley (2015) and in Michler and Wu (2020) are consistent with the crowding out hypothesis, they are also consistent with a much simpler hypothesis: if third-party enforcement becomes good enough (due to higher

court quality, as in Antras and Foley, or due to stronger punishments, as in Michler and Wu), parties may be able to achieve the first-best level of cooperation (high trade credit in Antras and Foley's setting, reliable water delivery in Michler and Wu's setting), and hence they may simply not need relational contracting. In terms of our model from Sect. 25.2, if g drops down to zero, arm's-length contracting achieves the first best, and there are no gains from adding a relational bonus to the contractual piece rate. It would therefore be desirable to replicate the kind of test that Antràs and Foley (2015) and Michler and Wu (2020) perform using continuous measures of both relational contracting and third-party enforcement. Observing that marginal improvements in third-party enforcement first increase and then reduce the use of relational contracts would be consistent with crowding out but not with first-best formal contracting.

25.3.4 Relational Contracts and Organizational Structure

As discussed in Sect. 25.2.1.5, the canonical model of relational contracting can be reinterpreted to analyze a corporate boss's promise to delegate authority to a subordinate (Baker et al., 1999). Several empirical papers have documented how firms allocate decision authority efficiently, delegating it to subordinates when that is likely to add value. Some examples are McElheran (2014) on delegation of IT investments; Lo et al. (2016) on pricing authority in the industrial equipment industry; Guadalupe et al. (2014) on finance and human resource decisions in Fortune 500 companies; Kala (2022) on strategic decisions in Indian state-owned enterprises; and Yang et al. (2023) on invention development at a global ICT company. However, no paper (to the best of our knowledge) has directly and explicitly investigated "relational delegation," that is, whether delegation requires self-enforcing relational contracts to be sustainable and, particularly, whether the shadow of the future (i.e., the PDV of boss-subordinate relationships) facilitates delegation.

The empirical study that perhaps gets closest to documenting relational delegation is Bloom et al. (2012). Using cross-country plant-level data on about 4000 firms, they find that higher levels of bilateral trust between a multinational's home country and a subsidiary's country increase decentralization of authority from the headquarters to the subsidiary's managers. These results are consistent with the hypothesis that relational contracts are necessary to sustain delegation. However, while their explanatory variable, bilateral trust between two countries (as measured by the World Values Survey), probably correlates with the PDV of the relationship between a boss in country 1 and a subordinate in country 2, alternative interpretations of this variable, and of the findings in Bloom et al. (2012), are possible. For instance, as discussed by the authors, lack of bilateral trust may also reflect a CEO's belief that the subsidiary's manager has misaligned interests, which would lead the CEO to reduce delegation even if the decision to delegate were contractible. Future studies should therefore aim to develop direct measures of the PDV or

discount factor in a parent-subsidiary relationship (perhaps drawing upon the measurement methodologies discussed in Sect. 25.3.1 above) and integrate these measures into the cross-country delegation data of Bloom et al. (2012), which are publicly available.

Overall, the paucity of empirical evidence on how relational contracts sustain delegation and corporate organizational structures represents in our view an important opportunity for future research.

25.3.5 Future Challenges: Empirical Analyses of the Clarity Problem

In the field of relational contracting, empirical research typically comes after a wave of theoretical models. Given the difficulty of collecting data on this phenomenon, creative and sometimes indirect tests need to be designed, and theoretical models provide guidance in the development of such tests. Theoretical research on the clarity problem is still in its infancy, and we do not have yet a canonical model of this problem and its solutions, like we do for the credibility/self-enforcement problem (see Sect. 25.2 of this chapter). It is therefore not surprising that we have virtually no empirical investigations of clarity. This is in a sense exciting because we foresee that clarity will be one of the main areas of research on contracting in the upcoming years.

Perhaps the only published empirical paper on clarity to date is Gibbons et al. (2023). In their lab experiment, a buyer and a seller play a trading game in several periods, unable to predict how their environment will change over time. Gibbons et al. (2023) begin by asking players what pricing and quality decisions they would consider to be consistent with the relational contracts under different hypothetical scenarios. Then, they find that after one of these scenarios is actually realized in the experiment, buyer-seller pairs that lacked a shared understanding of the relational contract (i.e., pairs plagued by a clarity problem) are less likely to make decisions that match the other player's expectation (i.e., they are less likely to reach an equilibrium), resulting in lower joint surplus.

In the second part of the experiment, Gibbons et al. (2023) explore how formal agreements the parties may reach at the outset on how to play the game may attenuate the clarity problem. They find that buyer-seller pairs who agree on a broad principle (i.e., a general criterion on how to adapt price and quality to future contingencies) rather than a narrow rule (specific price and quality levels) are more likely to reach a shared understanding of the relational contract and efficient decisions as the game proceeds.

One potential consequence of the clarity problem, as documented by Gibbons et al. (2023), is failure to reach an equilibrium and to efficiently adapt decisions to realized contingencies. An even worse consequence, as discussed in Sect. 25.2 above, might be the misinterpretation of cooperation as defection (due to the lack of a common understanding of the relational contract) and the consequent breakdown of valuable relationships. Because of the severity of such a consequence,

parties may attempt to preempt the clarity problem by choosing actions that the other party is less likely to interpret as defection, even if these actions are inefficient.

A recent working paper by Espinosa, Groshjean, Yuan, and Zanarone (2024) provides evidence consistent with such preemptive behavior. They use US data on individual meetings between lobbyists and Congress legislators to investigate collaborative relationships between the two sets of parties. In these collaborations, the politician provides political support for the lobbyist's clients, and the lobbyist provides useful information and advice to the politician. Because a typical lobbyist is time-constrained and has many clients, the efficient relational contract is one where the lobbyist meets the politician only in the "right" periods—that is, those where the politician sits in a committee that is relevant to the lobbyist's clients. However, politicians switch committees (and lobbyists switch clients) with some frequency, implying that *ex ante*, the two parties may lack a shared understanding of what "right period" might mean. Espinosa et al. (2024) hypothesize that to prevent relationships from breaking down due to this clarity problem, lobbyists may engage in hyper-conservative behavior, meeting a given politician even when the period is not right. Consistent with this hypothesis, they find that lobbyist-politician pairs that have interacted frequently in the past (and hence have had a chance to develop a common understanding) efficiently interrupt meetings when the politician rotates to a committee that does not matter to the lobbyist's client (and resume meetings when the opposite kind of rotation occurs). In contrast, pairs that have had few past interactions (and hence are less likely to share a common understanding of the relational contract) do not interrupt meetings after the politician rotates to a less relevant committee.

Future empirical investigations of the clarity problem (and the role of contractual formality in solving this problem) could collect high-frequency data on cooperative actions that are contingent on hard-to-predict future states and investigate whether clarity-enhancing formal provisions increase the frequency of efficient outcomes (i.e., adaptation of decisions to the realized state) and/or decrease the frequency of wasteful disagreement and relationship termination. Implementing this kind of empirical test poses at least three challenges. First, it is difficult to obtain high-frequency data on state-contingent cooperation as that information is typically proprietary and confidential. Second, it is difficult to find sources of exogenous variation in contractual formalism that allow the researcher to identify causal effects (as opposed to correlations). Third, given that contractual provision also play an enforcement-enhancing role in relational agreements (as discussed in Sect. 25.2), it is hard to assess whether any observed effect of formal contracts on adaptation and termination is due to the clarity-enhancing or the enforcement-enhancing role of such contracts.

Regarding the first challenge (collecting high-frequency data on state-contingent cooperation), a promising empirical setting to tackle it would be the US airline industry where, as discussed above, a key cooperation problem is the coordinated exchange of landing slots between major airlines and regional partners under continuously changing weather conditions. Real-time data on slot exchanges at US airports are available and are being used to study governance (Argyres et al., 2024).

Regarding the second challenge (finding exogenous variation in contractual formalism), researchers could exploit natural experiments such as regulation-induced changes in contractual design (Zanarone, 2009) or conduct field experiments to generate random variations in formalism. Several field experiments on the effects of organizational forms and management practices have been successfully conducted in the development literature (Macchiavello, 2022). The methodologies developed in this literature could serve as a guidance for empirical researchers interested in exploring the clarity-enhancing role of formalism. Regarding the third challenge (disentangling the clarity-enhancing and enforcement-enhancing roles of formalism), a possible strategy would be to identify settings in which the threat of litigation is not feasible nor credible, and thus any enforcement-enhancing rule of formalism can be excluded by design. For instance, it could be argued that certain cross-border contractual agreements are not backed by a credible threat of litigation. Moreover, Ben-Shahar and Bernstein (2000) argue that companies for whom trade secrets and proprietary technology and know-how are especially valuable do not have a credible threat to sue their suppliers for non-performance because the risk that confidential information will become public through the discovery process is prohibitive for them.

While not an easy task, these empirical strategies are potentially feasible and may enable future researchers to make progress (beyond laboratory experiments) toward understanding the role of contractual formality in generating shared understandings and interpretations of relational contracts, thereby solving the clarity problem.

25.4 Concluding Remarks

In this paper, we have presented a simple version of the canonical model of relational contracting (Baker et al., 1994, 2002; Levin, 2003), which provides predictions on the complementarity between contractual and relational incentive provisions, and we have discussed selected empirical studies through the lenses of this model. While we have found support for the hypothesis that contractual and relational provisions interact with and complement each other, we have also noted several opportunities for future empirical research on governance. In particular, future researchers may investigate the possibility that imperfect court enforcement institutions crowd out relational contracting, the role of relational contracting in sustaining delegation promises in organization, and the role of contractual formalism in solving clarity problems (as opposed to enforceability problems) in relational agreements. We hope that our discussion of theoretical and empirical approaches to relational contracting will encourage future cohorts of researchers within and outside economics to continue to expand our understanding of this important and fascinating phenomenon.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Aghion, P., & Tirole, J. (1997). Formal and real authority in organizations. *Journal of Political Economy*, 105(1), 1–29.
- Alonso, R., Dessein, W., & Matouschek, N. (2008). When does coordination require centralization? *American Economic Review*, 98(1), 145–179.
- Antràs, P., & Foley, C. F. (2015). Poultry in motion: A study of international trade finance practices. *Journal of Political Economy*, 123(4), 809–852.
- Argyres, N., Bercovitz, J., & Zanarone, G. (2020). The role of relationship scope in sustaining relational contracts in Interfirm networks. *Strategic Management Journal*, 41, 222–245.
- Argyres, N., Gil, R., & Zanarone, G. (2024, Forthcoming). Does supply concentration encourage cooperation? Evidence from airlines. *Management Science*. <https://doi.org/10.1287/mnsc.2022.03455>.
- Arruñada, B. (2007). Pitfalls to avoid when measuring institutions: Is ‘doing business’ damaging business? *Journal of Comparative Economics*, 35(4), 729–747.
- Baker, G., Gibbons, R., & Murphy, K. J. (1994). Subjective performance measures in optimal incentive contracts. *Quarterly Journal of Economics*, 109, 1125–1156.
- Baker, G., Gibbons, R., & Murphy, K. J. (1999). Informal authority in organizations. *The Journal of Law, Economics, and Organization*, 15(1), 56–73.
- Baker, G., Gibbons, R., & Murphy, K. J. (2002). Relational contracts and the theory of the firm. *Quarterly Journal of Economics*, 117, 39–84.
- Baker, G., Gibbons, R., & Murphy, K. J. (2011). *Relational adaptation*. SSRN: <https://ssrn.com/abstract=4256008>
- Barron, D., Gibbons, R., Gil, R., & Murphy, K. J. (2020). Reel authority: Relational renegotiation in the movie industry. *Management Science*, 66(5), 1868–1889.
- Bartlett, A. (1998). Lincoln electric: Venturing abroad. *Harvard Business School Case* #398–095.
- Battigalli, P., & Maggi, G. (2008). Costly contracting in a long-term relationship. *RAND Journal of Economics*, 39, 352–377.
- Ben-Shahar, O., & Bernstein, L. (2000). The secrecy interest in contract law. *Yale Law Journal*, 109, 1885–1925.
- Berkowitz, D., & Clay, K. (2006). The effect of judicial Independence on courts: Evidence from the American states. *The Journal of Legal Studies*, 35, 399–440.
- Bernstein, L. (1992). Opting out of the legal system: Extralegal contractual relations in the diamond industry. *The Journal of Legal Studies*, 21(1), 115–157.
- Bernstein, L., & Peterson, B. (2023). Managerial contracting: A preliminary study. *Journal of Legal Analysis*, 14, 176.
- Bloom, N., Sadun, R., & Van Reenen, J. (2012). Americans do IT better: US multinationals and the productivity miracle. *American Economic Review*, 102(1), 167–201.
- Bolton, P., & Dewatripont, M. (2013). Authority in organizations. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics*. Princeton University Press.
- Bull, C. (1987). The existence of self-enforcing implicit contracts. *Quarterly Journal of Economics*, 102, 147–159.
- Choi, S., Gulati, G., & Posner, E. (2008). Professionals or politicians: The uncertain empirical case for an elected rather than an appointed judiciary. *Journal of Law, Economics and Organization*, 26, 290–336.
- Corts, K., & Singh, J. (2004). The effect of repeated interaction on contract choice: Evidence from offshore drilling. *Journal of Law, Economics, and Organization*, 20, 230–260.
- Dessein, W. (2002). Authority and communication in organizations. *The Review of Economic Studies*, 69(4), 811–838.
- Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2003). Courts. *Quarterly Journal of Economics*, 118(2), 453–517.
- Eccles, R. (1981). The Quasifirm in the construction industry. *Journal of Economic Behavior and Organization*, 2(4), 335–357.

- Espinosa, M., Groshjean, T., Yuan, Z., & Zanarone, G. (2024). How the clarity problem affects relational contracts: Evidence from Lobbying. mimeograph.
- Fast, N., & Berg, N. (1975). The Lincoln electric company. *Harvard Business School Case* #376–028.
- Forbes, S., & Lederman, M. (2009). Adaptation and vertical integration in the airline industry. *American Economic Review*, 99, 1831–1849.
- Foss, N. (2003). Selective intervention and internal hybrids: Interpreting and learning from the rise and decline of the Oticon spaghetti organization. *Organization Science*, 14, 331–349.
- Frydlinger, D., & Hart, O. (2024). Overcoming contractual incompleteness: The role of guiding principles. *Journal of Law, Economics and Organization*, 40(3), 625–647.
- Garicano, L. (2000). Hierarchies and the Organization of Knowledge in production. *Journal of Political Economy*, 108(5), 874–904.
- Garicano, L., & Van Zandt, T. (2013). Hierarchies and the division of labor. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics*. Princeton University Press.
- Gibbons, R. (2005). Incentives between firms (and within). *Management Science*, 51, 2–17.
- Gibbons, R., & Henderson, R. (2012). Relational contracts and organizational capabilities. *Organization Science*, 23, 1350–1364.
- Gibbons, R., & Henderson, R. (2013). What do managers do? In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics*. Princeton University Press.
- Gibbons, R., Matouschek, N., & Roberts, J. (2013). Decisions in organizations. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics*. Princeton University Press.
- Gibbons, R., LiCalzi, M., & Warglien, M. (2021a). What situation is this? Shared frames and collective performance. *Strategy Science*, 6(2), 124–140.
- Gibbons, R., Siegel, J., & Weber, R. (2021b). Strategy meets culture (for breakfast): Understanding the relationship and highlighting its potential. *Strategy Science*, 6(2), 111–118.
- Gibbons, R., Grieder, M., Herz, H., & Zehnder, C. (2023). Building an equilibrium: Rules vs. principles in relational contracts. *Organization Science*, 34(6), 2231–2249.
- Gil, R., & Marion, J. (2013). Self-enforcing agreements and relational contracts: Evidence from California highway procurement. *Journal of Law, Economics and Organization*, 29, 239–277.
- Gil, R., & Zanarone, G. (2017). Formal and informal contracting: Theory and evidence. *Annual Review of Law and Social Science*, 13, 141–159.
- Gil, R., & Zanarone, G. (2018). On the determinants and consequences of informal contracting. *Journal of Economics & Management Strategy*, 27(4), 726–741.
- Gil, R., Kim, M., & Zanarone, G. (2022). Relationships under stress: Relational outsourcing in the U.S. airline industry after the 2008 financial crisis. *Management Science*, 68(2), 1256–1277.
- Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology*, 91(3), 481–510.
- Guadalupe, M., Li, H., & Wulf, J. (2014). Who lives in the C-suite? Organizational structure and the division of labor in top management. *Management Science*, 6, 824–844.
- Hadfield, G. K., & Bozovic, I. (2016). Scaffolding: Using informal contracts to support informal relations in support of innovation. *Wisconsin Law Review*, 2016(5), 981–1032.
- Hastings, R., & Meyer, E. (2020). *No Rules Rules*. Penguin Press.
- Helper, S., & Henderson, R. (2014). Management practices, relational contracts, and the decline of general motors. *Journal of Economic Perspectives*, 28(1), 49–72.
- Johnson, S., McMillan, J., & Woodruff, C. (2002). Courts and relational contracts. *Journal of Law, Economics, and Organization*, 18, 221–277.
- Kala, N. (2022, forthcoming). The impacts of managerial autonomy on firm outcomes. *Econometrica*.
- Kalnins, A., & Mayer, K. (2004). Relationships and hybrid contracts: An analysis of contract choice in information technology. *Journal of Law, Economics, and Organization*, 20, 207–229.
- Klein, B. (1980). Transaction cost determinants of "unfair" contractual arrangements. *American Economic Review Papers and Proceedings*, 70(2), 356–362.

- Klein, B. (2000). The role of incomplete contracts in self-enforcing relationships. *Revue d'Économie Industrielle*, 92, 67–80.
- Kosova, R., & Sertsios, G. (2018). An empirical analysis of self-enforcement mechanisms: Evidence from hotel franchising. *Management Science*, 64, 43–63.
- Kvaløy, O., & Olsen, T. (2009). Endogenous verifiability and relational contracting. *American Economic Review*, 99, 2193–2208.
- La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2007). *The economic consequences of legal origins* (NBER working paper 13608).
- Lafontaine, F., & Slade, M. (2013). Inter-firm contracts: The evidence. In R. Gibbons & J. Roberts (Eds.), *The handbook of organizational economics*. Princeton.
- Lazear, Edward P., & Oyer, Paul. (2013) “Personnel Economics”. *The Handbook of Organizational Economics*, edited by Robert Gibbons and John Roberts, Princeton: Princeton University Press, pp. 479–519.
- Levin, J. (2003). Relational incentive contracts. *American Economic Review*, 93(3), 835–857.
- Li, J., Matouschek, N., & Powell, M. (2017). Power dynamics in organizations. *American Economic Journal: Microeconomics*, 9(1), 217–241.
- Lo, D., Dessein, W., Ghosh, M., & Lafontaine, F. (2016). Price delegation and performance pay: Evidence from industrial sales forces. *The Journal of Law, Economics, and Organization*, 32(3), 508–544.
- Macaulay, S. (1963). Non-contractual relations in business: A preliminary study. *American Sociological Review*, 28, 1–19.
- Macchiavello, R. (2022). Relational contracts and development. *Annual Review of Economics*, 14, 337–362.
- Macchiavello, R., & Morjaria, A. (2015). The value of relationships: Evidence from a supply shock to Kenyan rose exports. *American Economic Review*, 105, 2911–2945.
- MacLeod, B., & Malcomson, J. (1989). Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica*, 57(2), 447–480.
- MacNeil, I. (1978). Contracts: Adjustments of long-term economic relations under classical, neoclassical, and relational contract law. *Northwestern University Law Review*, 72, 854–906.
- Malcomson, J. (2013). Relational incentive contracts. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics*. Princeton University Press.
- Masten, S. (2021). Adaptation, adjudication, and private ordering: Contractual relations through the Williamson lens. *Journal of Institutional and Theoretical Economics*, 2, 283–296.
- McElheran, K. (2014). Delegation in multi-establishment firms: The organizational structure of IT purchasing authority. *Journal of Economics and Management Strategy*, 23(2), 225–257.
- Ménard, C. (2013). Hybrid modes of organization. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics*. Princeton University Press.
- Michler, J. D., & Wu, S. Y. (2020). Governance and contract choice: Theory and evidence from groundwater irrigation markets. *Journal of Economic Behavior and Organization*, 180, 129–147.
- Mookherjee, D. (2013). Incentives in hierarchies. In R. Gibbons & J. Roberts (Eds.), *Handbook of organizational economics*. Princeton University Press.
- Poppo, L., & Zenger, T. (2002). Do formal contracts and relational governance function as substitutes or complements? *Strategic Management Journal*, 23, 707–725.
- Poppo, L., Zhou, K., & Ryu, S. (2008). Alternative origins to Interorganizational trust: An interdependence perspective on the shadow of the past and the shadow of the future. *Organization Science*, 19, 39–55.
- Rantakari, H. (2008). Governing adaptation. *The Review of Economic Studies*, 75(4), 1257–1285.
- Rantakari, H. (2023, forthcoming). Simon says? Equilibrium obedience and the limits of authority. *Journal of Law, Economics, and Organization*. <https://doi.org/10.1093/jleo/ewad026>.
- Ryall, M., & Sampson, R. (2009). Formal contracts in the presence of relational enforcement mechanisms: Evidence from technology development projects. *Management Science*, 55, 906–925.

- Schwartz, A. (1992). Relational contracts in the courts: An analysis of incomplete agreements and judicial strategies. *Journal of Legal Studies* 21(2), 271–318.
- Stewart, J. (1993). Double dare at First Boston. *Financial Review*, September 17. <https://www.afr.com/politics/double-dare-at-first-boston-19930917-jl05w>
- Telser, L. (1981). Margins and futures contracts. *Journal of Future Markets*, 1(2), 225–253.
- Williamson, O. (1979). Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics*, 22(2), 233–61
- Williamson, O. (1991). Comparative economic organization: The analysis of discrete structural alternatives. *Administrative Science Quarterly*, 36, 269–296.
- Yang, J., Criscuolo, P., & Silverman, B. (2023). The Organizational Structure of Development and Firm Innovation. mimeograph.
- Zanarone, G. (2009). Vertical restraints and the law: Evidence from automobile franchising. *Journal of Law and Economics*, 52, 691–700.
- Zanarone, G. (2013). Contract adaptation under legal constraints. *Journal of Law, Economics and Organization*, 29, 799–834.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Contracting with Public Authorities: Recent Developments 26

Jean Beuve and Stéphane Saussier

Contents

26.1	Introduction	652
26.2	Specificities of Public Contract	653
26.2.1	Corruption Risks and Contract Rigidity	654
26.2.2	Third-Party Opportunism and Contract Rigidity	654
26.2.3	Political Contestability and Contract Rigidity	656
26.3	Public Contract Renegotiations	657
26.3.1	Renegotiations as Opportunistic Behavior	658
26.3.2	Renegotiations as Relational Tools for Public Contracts	659
26.4	Rules Versus Discretion	661
26.4.1	The Benefits of Discretion	662
26.4.2	The Detrimental Effects of Discretion	664
26.5	Quality of Institutions and Capabilities of Public Buyers	666
26.5.1	Institutional Framework	666
26.5.2	Buyers' Competences and Integrity	666
26.6	Conclusions and Policy Implications	668
	References	669

Abstract

Recent theoretical and empirical developments suggest that public contracts are characterized by intrinsic differences from private contracts, due to significant oversight and control by political opponents and interest groups, who have an interest in challenging and disrupting the contractual relationship. As a result, public contracts are more rigid and more frequently renegotiated than private contracts. This

J. Beuve
Lille Economie Management, University of Lille, Lille, France
e-mail: jean.beuve@univ-lille.fr

S. Saussier (✉)
Sorbonne Business School, Institut d'administration des entreprises, University Paris I Panthéon-Sorbonne, Paris, France
e-mail: Saussier@univ-paris1.fr

chapter provides a brief progress report on these recent studies and highlights the need to improve the tools currently available to public procurement departments by introducing more flexibility at the award stage (e.g., negotiation, discretion at the award stage) and at the execution stage (e.g., renegotiation of contracts). The appropriate utilization of this higher flexibility is contingent upon the quality of the institutional framework, as well as the integrity and competencies of public officials.

Keywords

Public procurement · Public contract · Public scrutiny · Contractual rigidity · Renegotiation · Discretion

26.1 Introduction

Given that it accounts for approximately 13% of GDP in OECD countries and 15% in EU countries (OECD, 2023), it is vital for economists and policy-makers to understand public procurement.

For a long time, studies have focused on issues of asymmetric information, examining strategies to get private companies to reveal their costs through competitive tendering (Tadelis, 2012). However, considering transaction costs complicates the contracting process and often makes incomplete contracts and renegotiation the rule more than the exception. Adding ad hoc assumptions in incentive makes it possible to account for various forms of transaction costs (Malin & Martimort, 2000) leading to incomplete contracts. Unforeseeable contingencies, the cost to write the (high number of) foreseeable ones into a contract, the fact that some contingencies may be private information (i.e., not observable), or the impossibility for the contracting parties to commit not to renegotiate, all those elements make the specification of “incentive-compatible” mechanisms of information transmission particularly expensive and render incomplete contracts and renegotiation an attractive alternative (Tirole, 1986, 2009).

In this vein, transaction cost economics (Williamson, 1985) as well as property rights theory (Hart & Moore, 1988) have emphasized that complex contracts are inevitably incomplete and therefore renegotiated, relegating questions of incentives to the background (Bajari et al., 2014). These two theories have significantly enhanced our comprehension of public procurement and its efficiency factors. For example, Transaction Cost Theory has proven extremely useful in understanding the choice of management modes and contractual choices made by public authorities; as for property right theory, it has, among other things, highlighted the significant influence of verifiability on the ability of public authorities to outsource certain services.

While the abovementioned contract theories can contribute to the analysis of public contracts, they often fail to capture an important and crucial aspect. Indeed, one of the fundamental characteristics of public contracts is that they are subject to public control because they involve public funds.

Public scrutiny and the need for transparency might be justified by corruption risks, and in this case, political opponents, representative of taxpayers like auditors

or even competing suppliers, can be considered as useful watchdogs for accountability of public buyers to taxpayers. However, public oversight is not always free from bad intentions, and abovementioned competitors, auditors, and political opponents can be opportunistic third parties, more interested in the failure than in the success of public decision-makers' actions. Indeed, as Moszoro and Spiller (2014) argue, public contracts are characterized by intrinsic differences arising from significant oversight and control by political opponents and interest groups, who have an interest in challenging and disrupting a contractual relationship. Thus, public contracts are often subject to political considerations and public scrutiny, which can limit the flexibility of public authorities in negotiating and implementing such contracts (Beuve et al., 2019). One of the consequences is that public contracts are subject to inherent rigidities, independent of corruption issues.

The recognition that public oversight is not always free from bad intentions and has consequences on public contracts is also reshaping the way renegotiation is viewed. Thus, while proponents of incentive theory believe that "high rates of contract renegotiation have raised serious questions about the viability of the concession model in developing countries" (Guasch et al., 2008, p. 421), others, who recognize that public contracts must be renegotiated because of their inherent differences from private contracts, conclude that "the frequency of contract renegotiation may give concessions a 'relational' quality" (Spiller, 2009, p. 63). A small but developing body of empirical research corroborates this (Beuve et al., 2023; Beuve & Saussier, 2021) and suggests that granting public procurement officers greater discretion during tender procedures may prove beneficial (Decarolis et al., 2016; Coviello et al., 2018; Kang & Miller, 2021). However, it is of paramount importance to recall that the potentially beneficial effects of granting greater discretion to public officials at various stages of the contract life cycle are contingent upon the quality of the institutional framework as well as the integrity and competencies of said officials.

In this chapter, we focus on recent developments that are changing the way we perceive the challenges associated with public contracts, recognizing that (1) public contracts are intrinsically different from private contracts and (2) their renegotiation can bring good news, but only under certain specific conditions concerning (3) discretion and (4) the quality of the institutional framework and the capabilities and probity of public officials.

26.2 Specificities of Public Contract

When contracting for or with complex objectives (e.g., uncertain quality or unpredictable future events), the survival of a contract may be at risk. One way to deal with contract complexity is to rely on informal commitments, that is, relational contracting (Macaulay, 1963). However, to rely on relational contracts in the public domain is more complicated than in the private one, if not unfeasible. As Spiller (2009) points out, "when faced with unforeseen or unexpected circumstances, private contractees, as long as the relation remains worthwhile, adjust their required performance without the need for costly renegotiation or formal recontracting. Public contracting, on the other hand, seems to be characterized by formalized,

standardized, bureaucratic, rigid procedures” (p. 45). Indeed, the awarding of public contracts is governed by a series of procedural rules, reflecting a bureaucratic approach. This stiffness is characterized by the inclusion of contractual specifications that enforce strict compliance after the fact, allowing little room for flexibility, showing a low tolerance for changes, and imposing penalties for any deviations. But where does this rigidity come from?

26.2.1 Corruption Risks and Contract Rigidity

Public procurement represents a significant area of government operations highly susceptible to corruption. This vulnerability stems not only from the sheer number of transactions and the substantial financial interests involved but also from the intricate nature of the procurement process. The risks of corruption are further heightened by the intensive interaction between government personnel and private sector entities, along with the involvement of numerous stakeholders.

Obviously, corruption is more likely to occur in situations where procedures for awarding public procurement contracts and the contracts themselves lack precision. Iossa and Martimort (2016) highlight how corruption influences the tendency toward incomplete contracts, which do not specify state-contingent clauses, thereby saving on verifiability costs but increasing agency costs. Their study models the impact of corruption on the selection between complete and incomplete contracts. They show that corruption leads to a preference for incomplete contracts, which are less efficient in risk management but may facilitate corrupt practices by allowing public officials to channel additional compensation to contractors.

This preference for incomplete contracts can lead to inefficiencies in risk allocation and increased costs for taxpayers, who ultimately bear the financial burden of corruption through higher project costs or reduced quality of public services.

Even when complete state-contingent contracts are not feasible, incomplete contracts might opt for rigidity (to limit renegotiations) by stipulating rigid options that applies under all circumstances and do not rely on any communication from the contractor. The use of standardized contracts and the reduction of discretion in contract management can be viewed as measures to mitigate the negative impacts of corruption on contract completeness and efficiency, ultimately protecting taxpayers’ interests.¹

26.2.2 Third-Party Opportunism and Contract Rigidity

Recent developments in the literature on public contracting emphasize the specificities of public versus private contracting, arguing that public contracts are subject to third-

¹ Obviously, an incomplete flexible contract giving a public official the right to decide for a public authority depending on its local information is preferred to a rigid one as long as the public officer is not corrupt and his preferences are aligned with the public authority (Iossa & Martimort, 2016).

party opportunism (Moszoro & Spiller, 2019; Spiller, 2009). Indeed, public contracts are often subject to political considerations and public scrutiny, which are not only carried out by the designated agencies in charge of contract supervision but also by interested third parties (e.g., consumer associations, lobbyists, political opponents, excluded contractors) that may opportunistically question the probity of public officials. In anticipation of potential challenges from third parties, both the public administration and the private contractor have an incentive to increase the rigidity of public contracts.

In addition, private contractors are concerned about the risk of governmental opportunism, i.e., the ability of governments to change the rules through ordinary administrative powers to extract quasi-rents from investors (Spiller, 1995). Sunk investments provide politicians with the opportunity to behave opportunistically toward the investing company, exposing it to the risk of expropriation. Investors will therefore need special guarantees to deploy their capital by putting in place institutional arrangements that limit the government's ability to behave opportunistically. While this type of risk of opportunistic behavior is more frequently observed in developing countries, where the institutional framework may be of lower quality, it is by no means absent in developed countries (see Sect. 5).

Ultimately, both third parties and governmental opportunism increase the incentives of public officials and private investors to raise contractual rigidity. As a consequence, by their very nature, public contracts are less flexible than purely private contracts (Spiller, 2009). While this argument appears to be quite convincing from a theoretical standpoint, putting it to the test is much more difficult. In particular, the difficulty of comparing contractual terms in the public and private spheres explains why public and private sector contracts have only been studied extensively in their respective fields and why there has been no empirical analysis of public versus private contracts.

Beuve et al. (2019) have sought to fill this empirical gap. They have collected a unique and complete set of contracts concerning car park agreements signed between 1985 and 2009 in France by the biggest operator, and they compare supply contracts in which the buyer is either a public administration ("public contracts") or a private company ("private contracts"). Similar to Schwartz and Watson (2012), they analyzed this rich sample of contracts using textual analysis to construct ordinal measures of contractual rigidity. This paper is the first to investigate the intrinsic properties of public contracts. Using contracts for the supply of a standard product with measures of rigidity and a clear identification between public and private contracting partners, results clearly indicate that public contracts contain many more rigidity clauses² than private ones. These results strongly support the idea

²Rigidity clauses are defined based on the methodology developed by Moszoro et al. (2016). The authors assess rigidity levels of contracts by constructing "dictionaries" of machine-read rigidity categories including arbitration, certification, evaluation, litigation, penalties, termination, and contingencies. These categories capture relevant contractual clauses intended to signal probity and lower the likelihood of challenges by third parties. Each category encompasses specific search terms related to its domain, for example, arbitration includes terms like "appeal," "arbitration," and "conciliation," among others. This approach allows for a systematic analysis of contracts to measure the degree of difference in rigidity between contracts by analyzing the normalized frequencies of the total count of search terms across these rigidity categories.

that public contracts are more rigid by nature. In addition, the authors also collect local election data for the 132 public authorities covered by public contracts in the dataset, allowing them to explore the political dimension of contract rigidity in greater depth, as described in the following subsection.

26.2.3 Political Contestability and Contract Rigidity

As previously emphasized, Moszoro and Spiller (2014) consider that public contracts are characterized by intrinsic differences stemming from a substantial amount of supervision and control by political opponents and interest groups, who hold a stake in challenging and disrupting a contractual relationship. This is consistent with Ring and Perry's (1985) seminal contribution that public organizations are more permeable to the external environment than private organizations, not least because they face the scrutiny of the media and their constituents. Third-party opportunism can come from a multitude of sources such as consumer associations, lobbying activities, political opponents, or even private companies that have not been retained for the work. Of these sources, political opponents play an important role, not least because they can take advantage of general disapproval to replace the incumbent politician. For the incumbent politician, even simple projects can be difficult to implement if the political environment is hostile. Consequently, politics, which is generally alien to private-to-private contracts, becomes essential to understanding private-to-public contracts. Unlike private managers, public officials are subject to exogenous time constraints (i.e., elections). Since public officials do not want to be exposed to potential scandals or accused of mismanagement in front of their constituents, they prefer to increase the rigidity of a contract. This effect may be greater where electoral competition is fierce. Consequently, the need for contract rigidity is less in uncontested political contexts.³

Beuve et al. (2019) empirically test the permeability of public contracts to their political environment. As discussed in the previous subsection, they collect information about French local elections for the 132 public authorities covered by public contracts of the dataset. Then, they examine the interplay between the following two dimensions: the nature of contract (public or private depending on the nature of the co-contractor) and the intensity of political contestability that characterized the elections preceding signature within the city where the contract was taking place. Their results indicate that the rigidity of public contracts (i.e., evaluated through the frequency of rigidity clauses) rises when political competition is tight, thus revealing a link between public contract rigidity and political contestability. On the contrary, private contracts are not more rigid when they are signed in cities where political contestability is high. Furthermore, for a causal interpretation of the impact of political contestability on contractual rigidity, the

³ It is important to emphasize here that, most of the time, elected officials are not directly responsible for the administration and oversight of public contracts, which fall within the purview of public managers. However, as recipients of the externalities – both positive and negative – resulting from the effects of these contracts, they may influence or even dictate the activities of public managers.

researchers leveraged a legal change aimed at enhancing public procurement's scrutiny and transparency. They employed a difference-in-differences methodology to examine how political scrutiny influences contract rigidity. The Transparency, Anti-Corruption, and Economic Modernization Act, known as the "Sapin Act" (named after the Minister of Finance who advocated for it) and enacted in January 1993, aimed at ethical improvement in public tendering processes. A significant outcome of this legislation was the requirement for a public and transparent bidding process, heightening oversight on how public funds are utilized. The enactment of this law rendered municipalities with intense political contestability more susceptible to scrutiny from political adversaries. Consequently, in areas with high political competition, there was an expectation for newly executed contracts to exhibit greater rigidity as a precautionary measure against opportunistic legal challenges by third parties. Empirical results strongly corroborate this hypothesis: public officials in municipalities where political competition was structurally high⁴ increased the contractual rigidity of new contracts (i.e., contracts signed under the "Sapin Law") on average by two standard deviations.

Interestingly, in their study, the authors control for the possible influence of a corrupted environment on public contract rigidity. To do so, they use information from Transparency International France on the number of cases of corruption implicating the mayor or a member of the municipal council in the 3 years preceding the date the contract was signed. Even if this measure has limitations, as it is conditional upon corruption cases being detected and prosecuted which does not necessarily reflect endemic corruption, it appears that a corrupted environment is not statistically correlated to a higher level of contract rigidity. Consequently, the authors conclude that corruption does not play a role in their setting.

Altogether, these findings support the argument that a significant part of the differences in contractual rigidity among private and public contracts is a political risk adaptation to mitigate challenges from political contesters. The corollary is that the perceived inefficiency of public contracting is mostly the result of contractual adaptation to different inherent hazards and is, thus, not directly remediable (Williamson, 1999). Another important corollary is the fact that renegotiations cannot be analyzed in the same way when it comes to public contracts.

26.3 Public Contract Renegotiations

Because they address services of general interest, public contracts and their renegotiations are closely scrutinized. Many examples of renegotiations in public-private agreements are provided by Guasch (2004). By studying more than 1300 concession

⁴The authors use longitudinal electoral data to identify cities where political contestability is stable through time. Political contestability is measured through the win margin during elections, i.e., the differences between the number of votes for the winner and the number of votes for the runner-up candidate, divided by the total number of expressed votes. Then, political contestability within a city is considered as "structurally high" when this measure of political contestability is in the lowest quartile for six elections in a row.

contracts that were signed in Latin American and Caribbean countries between the mid-1980s and 2000, he found that 55% of transportation contracts and 74% of water and sanitation contracts were renegotiated, 2 years on average after their signature. Some of them led to some contract terminations. Similar frequencies of public contracts renegotiations have been observed in other countries and other sectors (Estache & Saussier, 2014).

26.3.1 Renegotiations as Opportunistic Behavior

Guasch (2004) highlights that renegotiation, at first glance, is a consequence of opportunistic behavior. Indeed, whether initiated by the private or public party, they resemble the desire to deviate from the initial contractual conditions and renegotiate them in one's favor. When initiated jointly by both parties, these renegotiations can also manifest as corrupt agreements between the parties to the detriment of end users. In their empirical analysis of renegotiation of concession contracts in Latin America, Guasch et al. (2008) show that the three most common outcomes of renegotiations are extension of deadlines in investment requirements (69%), reduction in investment requirements (62%), and price increases (62%); those empirical results are confirmed by a more recent study with updated data—see Guasch et al. (2017).

Regarding private operator opportunism, Guasch et al.'s (2008) interpretation is that renegotiations are a consequence of aggressive bids in the context of an *ex ante* lack of commitment from the government. Indeed, they show that the probability that a public contract is renegotiated increases when the institutional quality is lower, especially in developing countries. Moreover, recent empirical contributions seem to corroborate this view even in the case of developed countries. For instance, Brogaard et al. (2021) show that, in the USA, politically connected firms opportunistically win public contracts through aggressive bids knowing they will engage and succeed in renegotiation strategies to maximize the value of the whole-of-life transaction.⁵ This effect disappears abruptly as soon as the concerned politician withdraws from political life.

Such a result that might also reflect collusive behaviors between government officials and local bidders is consistent with the recent work by Coviello and Gagliarducci (2017). In their analysis of Italian public works, the authors analyze the delays in delivery, which represent a measure of *ex post* renegotiation of the original deadline of the contract. While almost 90% of the works are delivered on time, their evidence suggests that tenure in office (measured as the number of terms in office at the time of the bid) raises the number of days of delay in the delivery of the public works about 1 month per term.

⁵ A “connected firm” is identified through its political contributions, focusing on its relationship with politicians in Congress, specifically where the firm has contracts. This connection is defined based on contributions to political action committees (PACs) of House representatives or senators in locations relevant to the firm's contracts.

Lastly, renegotiations can also be a tool enabling incumbent governments to circumvent budgetary rules before elections (Engel et al., 2019). For instance, Le Squeren and Moore (2015) highlight the existence of a political cycle of renegotiations in public procurement in France. The authors use a difference-in-difference methodology to show that, compared with private contract renegotiations, public contract renegotiations significantly increase before local elections. Renegotiations aiming at modifying the end-user fees or the financial side of the contract (i.e., the remuneration of one of the parties) increase before an election, whereas all other types of renegotiations do not.

In the end, generalized renegotiations are often described as perverse outcomes of the growing participation of the private sector in public services (Albalade & Bel, 2009; Sarmiento & Cruz, 2018), and regardless of who is at the origin of the renegotiation process, the scant empirical literature on renegotiations mostly considers renegotiations as negative events and almost never as good news.⁶ This is confirmed by several empirical studies, including those we have mentioned so far.

Nevertheless, examining contractual renegotiations solely through the lens of opportunistic motives of the parties involved might appear overly simplistic, particularly considering that public contracts are often complex transactions for which objectives are changing and for which it is not easy to anticipate future events that may disturb the relationship. Renegotiations might also be viewed as necessary adaptations to fill in the contractual blanks (Grossman & Hart, 1986). Hence, renegotiation is simultaneously a necessary tool associated with greater surplus and a risky adaptation process that may lead to opportunistic behaviors (Fehr et al., 2011; Frydinger et al., 2019; Williamson, 1985). It is also consistent with the needs the contract is intended to meet for the parties. While they require a formal agreement to secure their specific investments, this security should not come without any contractual flexibility to adapt to unanticipated events. This trade-off leads to an optimal second-best level of contractual completeness (e.g., Crocker & Reynolds, 1993; Saussier, 2000) which translates to an optimal level of contractual renegotiations. Moreover, as emphasized in the previous section, the analysis of public contract renegotiations cannot be made without taking political issues, and *de facto* associated levels of contractual rigidity, into account. Recent contributions by Beuve et al. (2019) and Beuve and Saussier (2021) suggest a less negative perspective on the renegotiation of public contracts.

26.3.2 Renegotiations as Relational Tools for Public Contracts

Beuve et al. (2023) introduce a model of public procurement where the likelihood of renegotiations is influenced by contractual flexibility and the political willingness to accept deviations from the contract terms. In this model, contractual flexibility

⁶Renegotiations are seen as negative events for social welfare, not for the parties that may do well at appropriating a large amount of public funds by renegotiating.

facilitates adjustments without the need for formal renegotiation, whereas the political acceptance of deviations diminishes as political competition intensifies. Subsequently, they evaluate the rates of renegotiation in procurement contracts, distinguishing between those managed by public administrations and those overseen by private corporations. As previously discussed, contracts managed by public entities tend to exhibit greater rigidity, particularly when public officials operate under the shadow of tight political competition and the potential for opportunistic legal challenges against their contractual actions. From the viewpoint of the public agent, a strict contract reduces the likelihood of politically charged disputes. Similarly, for contractors, it lowers the risk of government opportunism, which could include unjust administrative actions or gradual encroachment on property rights (Moszoro & Spiller, 2019; Spiller 2009). Consequently, Beuve et al. (2023) hypothesize that increased political competition, by augmenting contractual rigidity, is likely to result in a higher incidence of formal renegotiations in public contracts. Their findings provide solid support for the model's expectations: public contracts undergo renegotiation more frequently than their private counterparts (by a margin of 7–13%), a trend that is particularly pronounced in jurisdictions with high political contestability.

In a similar approach, Beuve and Saussier (2021) posit that the design of public contracts is affected by the challenge of including the appropriate level of flexibility for renegotiation to occur when needed. Undesirable opportunistic renegotiations are likely to occur when there is too much flexibility, but a lack of it can also lead to the loss of opportunities for welfare-enhancing renegotiations. To empirically test their argument, they analyze the impact of contractual renegotiations, notably by disaggregating their frequencies and their types, on the likelihood of contract renewals. Because it is nearly impossible to assess how renegotiations influence contractual surplus, they instead use contract renewals as a proxy to indirectly assess parties' perceptions about their previous relationships and ultimately their cooperative adaptation and contractual surplus creation during renegotiations. If renegotiations result in a significantly negative outcome, parties are not prone to contract again. Their empirical analysis of public contracts in the car park sector allows them to reach two main conclusions. In a first step, a deliberately simplistic approach highlights a negative correlation between the likelihood of renewal and the absence of prior renegotiation (even once) during the contract's lifetime. This finding already introduces a nuance to the argument that renegotiations stem solely from opportunistic behavior of the parties involved. In a second step, the authors delve into a detailed analysis of the impact of frequency and nature of renegotiations, ultimately concluding the existence—at least within their dataset—of an optimal threshold for renegotiation frequency (0.7 renegotiations per year or, equivalently, one renegotiation every year and a half). These findings reinforce the insight that contracts are governance mechanisms that should be rigid enough to reflect real commitment from contracting parties but flexible enough to adapt to a changing environment.

Obviously, consistent with the previous subsection, we cannot exclude a potential link between contract renewal and corruption. Indeed, in such a case, the contract would be renewed if the two corrupt parties find themselves satisfied with the arrangement, sharing rents from unwarranted quality downgrades or price upgrades

through a bribe, damaging third parties like taxpayers and honest and possibly more efficient suppliers. To disentangle the opportunistic story (of collusion between private operators and elected officials) from the optimistic story (that assumes that if a contract is renewed, performance must have been good for the taxpayer), the authors consider the number of corruption cases implicating the mayor or a member of the city council between 1980 and 2010 to identify a potential corrupted environment. Notwithstanding the limitations of this measure—which is, as previously mentioned, conditional on cases being detected and prosecuted and does not necessarily reflect endemic corruption—the results indicate that corruption is of minor concern in their setting. This is especially the case since their analysis focuses on contracts between various public authorities and France’s largest car park company. Unless the entire sector is dominated by this operator, it is reasonable to assume that the negative reputation effects of corruption allegations likely outweigh any potential benefits from engaging in illicit activities. Moreover, the authors also control for the kind of renegotiations that the contract went through and highlight that while renegotiations about quality dimensions are associated with higher likelihood of contract renewal, renegotiations concerning tariffs and financial equilibrium of the contract (i.e., renegotiation to reduce quality or to increase the price) are statistically negatively correlated with contract renewal.

Altogether, the results of Beuve et al. (2023) and Beuve and Saussier (2021) indicate that previous empirical studies pointing to the inefficiencies of public contracts related to high renegotiation rates may have overlooked parts of the story. Such inefficiencies may not be remediable (Williamson, 1999), as the frequent renegotiations observed in public contracts can be understood as a consequence of their specific rigid nature instead of a manifestation of corruption, governmental opportunism, or even government incapacity to avoid renegotiation when dealing with opportunistic private corporations. To the contrary, the frequent renegotiation of public contracts may provide a relational quality of adaptability to contingencies in politically contestable environments. However, it is important to underline that the positive and significant relationship between renegotiations and renewals of public contracts, identified by Beuve and Saussier (2021), is contingent upon the abilities of public agents to renegotiate contracts and their capacity to base the decision of contract renewal (or termination) on the events that occurred during the contract’s lifespan. Therefore, the existence of an optimal level of renegotiations that is maximizing contract renewal stands if public authorities have discretion over their choice of partner. Such an argument about the potential benefits of the judicious use of discretionary power by public authorities echoes the dynamic literature on the debate between rules and discretion in public procurement.

26.4 Rules Versus Discretion

Stringent procedural regulations that limit the discretion of public buyers, such as the requirement for open competitive procedures, are widespread. These regulations are widely acknowledged as effective tools to curb corruption and favoritism, particularly

in environments with weak institutions (Bosio et al., 2022; Rose-Ackerman, 1999). However, the setting of rules to reduce abuses by the few often determines costs and suboptimal choices for the many. In this vein, Banfield (1975) and, specifically, Kelman (1990) for public procurement have argued long ago that restricting bureaucrats' discretion hampers their effective performance. More precisely, Kelman (1990) highlighted the drawbacks of excessively stringent regulations in US government procurement, positing that procurement regulation frequently poses more problems than solutions. In contrast, granting discretion to contracting authorities allows them to select, reward, and penalize contractors based on performance information that cannot be verified, which is vital for achieving quality that cannot be explicitly stipulated in contracts, leveraging relational and reputational dynamics (Decarolis et al., 2016; Spagnolo, 2012). Up to now, the extent of regulation and the level of discretion to leave to buyers are still some of the most controversial questions in public procurement. Recent research highlights the potential benefits of discretion, while others insist on the fact that risk of its abuse remains high, especially in the case of weak institutions and (even in the case of better institutional quality) in the context of emergencies.

26.4.1 The Benefits of Discretion

Carril (2021) examined the bunching of US federal contracts for goods and services below a simplified acquisition threshold. While the measurement of bunching was complicated by these contracts being negotiated and not having a reserve price, substantial bunching at the threshold was documented, attributed to contract manipulation and the deterrence of purchases.⁷ Contracts awarded with less discretion were found to perform worse in terms of ex post outcomes, and a significant increase in the threshold level was deemed optimal.

In another recent study by Coviello et al. (2018), Italian construction works data were examined using a regression discontinuity design around a regulatory threshold in the project value to identify the causal effects of increased procurement discretion on procurement outcomes.⁸ Below this threshold, regulations enhanced contracting authorities' ability to utilize restricted auctions with a minimal number of invited bidders, allowing them to exercise discretion in selecting who to invite. The study found that greater discretion did not worsen procurement outcomes, sometimes even improved them, and led to more frequent contract renewals with the same suppliers (i.e., incumbency) after good performance. The results indicated that discretion

⁷This refers to the avoidance of transactions that would exceed the threshold value, due to the anticipation of more cumbersome processes, increased scrutiny, or higher costs associated with contracts above that threshold. As a result, entities may deliberately limit the size or scope of their contracts to avoid crossing this threshold, leading to an accumulation of contracts just below it.

⁸Regression Discontinuity Design (RDD) is a quasi-experimental research design used to estimate the causal effect of a treatment or intervention by exploiting a natural or imposed threshold. In RDD, individuals or units are assigned to treatment or control conditions based on whether their value of a continuous assignment variable crosses a predetermined threshold or cutoff point.

increased the probability of the same firm winning repeatedly but did not worsen (and may have improved) the observed procurement outcomes. This qualitative finding aligns with the notion that repeated interactions enable renegotiating contracts without overly opportunistic behaviors.

In the same vein, Deodato Domingos et al. (2023) examine the links between the way contracts are awarded (formal versus discretionary approach) and the way they are renegotiated within the European Union. They find that when contracts are awarded through procedures allowing some discretion for the government, the time for renegotiation is longer, and renegotiations are less likely to be viewed as adversarial by contracting parties.

Hoekman and Onur Tas (2024) look at comprehensive procurement award data across 33 European nations. They found that limitations on the exercise of discretion correlate with higher average contract prices, with price increases being more pronounced in countries characterized by above-average government effectiveness. They further demonstrate that leveraging the legal allowance for greater discretion to achieve price reductions diminishes the chances of small firms winning contracts, a pattern that persists over time. These results highlight a fundamental trade-off between the potential for lowering prices through discretionary practices and the objective of policies designed to enhance the likelihood of small- and medium-sized enterprises (SMEs) securing contracts.

This also resonates with the observation that incentives based on reputation can significantly enhance supplier performance and mitigate the adverse impacts of price competition on contract fulfillment, as proposed by Decarolis et al. (2016). In their study, the authors examine a distinctive experiment conducted by a major utility company in Italy, which implemented a new vendor rating system. This system evaluated the historical performance of its suppliers and connected these evaluations to the assignment of forthcoming contracts. They find that responses in both price and performance to the announcement of the switch from price-only to price-and-rating auctions were particularly strong. Notably, the average performance enhanced dramatically, rising from 25% to 90% compliance with the audited parameters. These improvements, which spanned all parameters and suppliers, proved to be enduring, persisting for at least a decade following the initial experiment, and resulted in enhanced service quality from the utility. Meanwhile, contract prices remained relatively stable overall, although there was evidence of a temporary reduction in prices immediately following the announcement—as suppliers vied for contracts to receive scores—and an increase in prices once suppliers had secured a strong reputation. The main takeaway from their study is that a reputational mechanism based on objective past performance can be a powerful tool to curtail supplier moral hazard and, consequently, that discretion over their choice of participants can be strongly beneficial.

We can also mention the recent contribution by Kang and Miller (2021) in which they question the reason behind the low level of competition in government procurement. To do so, they develop, identify, and estimate a principal-agent model in which the procurer chooses whether to solicit bids and how much effort to exert to attract more bids, and then she negotiates with bidders to choose a winner and reach an agreement on a contract. Their theoretical model predicts that, for any given numbers of bidders, procurement agencies can extract more rent from a winning seller when they negotiate

compared to when they run an auction. Then, their empirical findings show that stripping the agencies of their discretion in designing and negotiating contracts would more than double the average number of bids but would hardly reduce the size of the payment to winning sellers. In other words, allowing procurement agencies to exercise some discretion to use their knowledge of the supply side can reduce procurement costs, even if they simultaneously engage in some rent-seeking behavior.

Here again, the political nature of public contracts cannot be set aside. In this vein, Coviello et al. (2022) delve into the mechanisms and actors behind the use of “illegal” discretion achieved through illicit alterations of contract values in public procurement, examining the implications for final procurement outcomes. The authors utilize a substantial dataset of Italian procurement transactions to investigate the phenomenon of contract value manipulation (bunching) aimed at staying below the legislative thresholds that would otherwise limit the autonomy of contracting authorities. Their findings reveal significant bunching just beneath these discretion-limiting thresholds, predominantly within public administrations whose officials are appointed by the government, like ministries and road authorities. Interestingly, this bunching effect vanishes in cases where officials, such as those in municipalities and provinces, are directly elected by the public. This result aligns with the political economy argument that electoral discipline is weaker for the former (i.e., officials appointed by the government), making it less costly to circumvent rules through contract value manipulation. Concerning procurement outcomes, Coviello et al. (2022) also emphasize that bunching led to a higher reliance on restricted auctions and a decrease in the number of bidders. It had a varied or nonexistent impact on discounts and prices. Additionally, it resulted in the choice of financially more robust suppliers and enhanced post-contractual outcomes, including delivery times, contractual delays, and cost overruns. They also show that appointed bodies exhibited a higher frequency of purchasing from the same sellers who performed better in the past. In sum, their findings align with the observation that administrations with appointed officials may breach procedural rules to adjust contract values and preserve discretion. This discretion is frequently employed to foster ongoing interactions with suppliers who are perceived as less risky and exhibit better performance, resulting in an overall improvement in procurement outcomes. On the contrary, administrations led by elected officials show no signs of such manipulation, likely leading to poorer procurement outcomes, which could be attributed to the more stringent accountability enforced by electoral discipline.

26.4.2 The Detrimental Effects of Discretion

The phenomenon of bunching of procurement contract values below discretion-restricting thresholds has also been studied by Palguta and Pertold (2017)—with data from the Czech Republic and Hungary, respectively—but with very different results. Indeed, in their cases, discretion-enhancing procurement manipulations lead to higher prices and the selection of suspicious (anonymous) or inefficient and politically connected suppliers.

Such results are consistent with other works which emphasize the perilous connections between discretion and the risk of corruption. For instance, Decarolis

and Giorgiantonio (2022) investigate the relationship between contract award discretion and corruption in public procurement, focusing on the Italian context. They found a systematic association between certain tender characteristics, such as multi-parameter awarding criteria and higher corruption risks. In the same way, Decarolis et al. (2020b) used a confidential database linking firms and officials investigated for corruption to over 200,000 procurement auctions and find that discretionary procedures (negotiated rather than open bidding) are associated with corruption, especially when competition at bidding stage is low.

Other recent contributions highlight the fact that increased discretion may enhance the risk of its abuse, in particular in the context of Covid-19 (Bandiera et al., 2021). For instance, Fazekas et al. (2021) explore how regulatory changes in Romania in March 2020, which allowed for simplified contracting processes for emergency goods, impacted corruption risks in public procurement. They monitored various risk indicators, such as the prevalence of direct contracting or abbreviated advertisement periods, as well as whether suppliers were registered in tax havens or entered the market without prior experience. Their analysis indicates a significant rise in corruption risks across all sectors after the regulatory adjustments, despite the emergency regulations being specifically intended for healthcare supply procurement.

From a political economics perspective, the cost of discretion is also analyzed by Baranek and Titl (2021), who investigate how shifts in the ruling party within the Czech Republic influence the contract values secured by companies that make campaign contributions. Their findings indicate that a company enhancing its political donations by 10% to a winning party can expect, on average, a 0.5–0.6% increase in the value of its public procurement contracts in the subsequent year. Firms that contribute to campaigns tend to secure more contracts in scenarios where discretion is permissible. The research further reveals that authorities often adjust contract values to stay below the discretion-allowing threshold. Additionally, it's shown that contracts awarded to these politically connected firms are overpriced by 11%, without any accompanying improvement in quality.

Such potential misallocation of public funds due to political interference facilitated by discretionary power of public administrations is also emphasized by Baltrunaite et al. (2021). Using Italian data on municipal public works contracts tendered between 2009 and 2013, the authors analyze the impact of a reform that broadened bureaucrats' discretion on supplier selection. They discover an increase in the proportion of contracts awarded to firms that count a local politician among their administrators or shareholders, alongside a decrease in the (ex ante) labor productivity of the winning firms. Crucially, the study also points out that these effects are predominantly observed in procurements of lower quality.⁹

⁹The quality of procuring agencies is measured through a synthetic indicator that combines (1) an indicator of corruption risk based on the number of reported crimes against the public administration in the period 2004–2011, on citizens' trust in local public institutions and on perceptions of administrations' integrity; (2) the educational level of local politicians and local public officials (as a proxy of human capital and competences); and (3) the percentage of tenders for which information on the execution stage is available (as a measure of transparency of public administration).

In the end, the contrasting findings emphasizing either the detrimental or the beneficial effects of increased discretion are consistent with Bosio et al.'s (2022) argument that the effects of rules and discretion depend significantly on the quality of the institutional framework.

26.5 Quality of Institutions and Capabilities of Public Buyers

As emphasized in the previous section, beneficial and detrimental uses of discretion coexist and, *de facto*, entail a trade-off whose strictness heavily depends on the institutional framework. Alongside the paramount importance of institutional quality, the competences and the probity of public buyers also play a fundamental role.

26.5.1 Institutional Framework

Following the contribution of Bosio et al. (2022), several recent research argue that the quality of surrounding institutions is crucial in determining whether discretion ends up being beneficial. For instance, Coccio et al. (2021) offer a comprehensive examination of emergency public procurement measures during the Covid-19 crisis, utilizing a survey conducted with 136 procurement experts from 103 countries during April to August 2020. Their findings indicate that countries more severely impacted by Covid-19 implemented a higher proportion of emergency procedures, alongside a consistent decline in transparency and accountability in Covid-related procurement. Aligning with Bosio et al. (2022), the research establishes that the quality of a country's institutions significantly influences the promptness and quality of its procurement response to the pandemic. Furthermore, it shows that robust institutions help mitigate the erosion of transparency and accountability standards in emergency procurement scenarios.

On a different and complementary approach, Decarolis et al. (2020b) illustrate that the advantages of bureaucratic discretion hinge on whether it is utilized for the public good or exploited for personal advantage. They explore the link between discretion and corruption within Italian government procurement auctions, drawing upon a confidential dataset of companies and procurement officials under investigation for corruption by Italian authorities. Their research reveals that auctions conducted under discretionary procedures are only correlated with corruption when there's a low level of competition during the bidding process. Additionally, they observe that while officials under investigation for corruption tend to favor these potentially corrupt discretionary auctions, such auctions are less commonly employed within procurement administrations where at least one official is under investigation for corruption.

26.5.2 Buyers' Competences and Integrity

If inefficient bureaucracy can represent a major obstacle (Djankov et al., 2002) to economic activities, it can also negatively affect the ability of private firms to do

business with governments. In accordance with such statement, the World Bank has recently begun to release its *Benchmarking Public Procurement*, examining public procurement laws and regulations across 180 economies; the report uncovers a wide variety of practices across countries, which can lead to considerable inefficiencies and waste. As emphasized by Decarolis et al. (2020a), these inefficiencies are often rooted in a lack of competence, particularly because procurement—especially of non-standardized goods or services—entails complex activities. Public buyers are tasked with identifying their needs, designing, and managing the award process, balancing risks and incentives in contracts, monitoring contractors, and overseeing contract execution, which can be lengthy and intricate. If these tasks require knowledge of products and market characteristics and of legal rules, they also demand strategic abilities to design the tender and negotiate the contract, good organization competencies, time management practices, and the coordination of experts at the different stages of the projects. In order to empirically assess the importance of competence and professionalization of bureaucrats, Decarolis et al. (2020a) conducted a survey among the US federal workforce and analyzed combined data from US procurement and federal employees. Their empirical analysis indicates that enhancing the competence of the procurement institution by one standard deviation leads to a 23% reduction in delays, a 29% decrease in cost overruns, and a 50% reduction in renegotiations. They also discovered that managerial competencies are particularly crucial for managing complex procurement processes, such as those aimed at fostering innovation. These processes often require greater discretion both during the award phase and throughout contract management.

Corroborating results are also put forward by Bandiera et al. (2009). Considering the price paid for standardized goods and services by different classes of Italian public buyers, the authors find that the expenditure would be reduced by 21%—corresponding to a saving of between 1.6% and 2.1% of the Italian GDP—if all public buyers paid the same prices as the public buyers at the tenth percentile, that is to say, if all the public buyers could exhibit comparable levels of effectiveness. Bandiera et al. (2009) assess that at least 82% of the estimated waste is related to excessive regulation, red tape, and bureaucratic inefficiency (i.e., passive waste). They argue that passive waste is more significant than active corruption in Italy and that more autonomous administrations have better procurement processes and are less corrupt than central public bodies.

Lastly, it is also interesting to highlight the fact that if the benefits of discretion—and also, more broadly, the government expenditure outcomes—are shown to improve with institutional quality and buyers competences, they also depend on the integrity of public officers.¹⁰ Indeed, several contributions mentioned in this chapter clearly point out the permeability of public contracts to their environment

¹⁰ Competences and integrity of public officials might be correlated, but not necessarily. A public official may be honest but, unfortunately, incompetent, whereas a competent public official may choose to leverage their skills for dishonest behavior.

and the risk of elected officials to distort the use of public procurement in order to reach their personal—rather than the general—interest. Therefore, elected public decision-makers with high integrity may make a more appropriate use of discretion and renegotiations. In a particularly interesting contribution, Bosio et al. (2022) aim to investigate this issue and posit that there is a degree of complementarity between rules on transparency and competition in procurement practices and rules on the types of income and asset disclosures that politicians must make public. More precisely, they argue that these procurement-specific rules function in a synergistic manner: procurement-specific regulations serve as preventative measures at what they call “entry” stage, aiming to mitigate inefficiency and corruption, whereas disclosure regulations are designed to further deter such behavior at what they call the “exit” stage. Collectively, they foster an environment more conducive to delivering high-quality public services.

To assess the empirical validity of their proposition, they use data on public procurement and disclosures of income and asset declarations of politicians in 175 countries to show that these two sets of regulations are indeed complementary in achieving better procurement outcomes and ensuring the proper behavior of politicians and government officials. In this way, they echo the older concept of probity transactions put forward (Williamson, 1999) and characterized by “their need for loyalty (to leadership and mission) and integrity of process.” In such transactions, the public authority needs a high level of professionalism from the contract holder. Indeed, any failure on the part of the latter to meet technical or scheduling objectives is likely to cause great loss to the public authority and to the public at large.

26.6 Conclusions and Policy Implications

In their policy paper commissioned by the French government to strengthen public procurement practices, Saussier and Tirole (2015) noted that greater flexibility for public purchasers could be beneficial, provided it is part of a wider move toward greater transparency, effective competition, and the development of specific public expertise. Recent developments in research on public contracting point in the same direction. Indeed, those developments suggest that public contracts are characterized by intrinsic differences from private contracts, due to significant oversight and control by political opponents and interest groups, who have an interest in challenging and disrupting the contractual relationship. As a result, public contracts are more rigid and more frequently renegotiated than private contracts. One implication is the need to introduce, for public buyers, more flexibility at the award stage (e.g., negotiation, discretion at the award stage) and at the execution stage (e.g., renegotiation of contracts), provided that the buyer’s capacity is adequate, and transparency is effective (i.e., control and sanctions in case of abuse are planned and enforced).

The EU public procurement directives passed in 2014 and applied since 2016 in EU countries can be analyzed in the light of these recent developments as a step in

the right direction. Indeed, the EU directives promote greater discretion on the part of the public purchaser when awarding and executing public contracts (e.g., by opening the way for contract negotiation at the award stage and allowing significant renegotiations at the execution stage) and, at the same time, improve transparency with the mandatory online publication of contract renegotiations, although without control and sanctions till now. From this point of view, there are rich lessons to be learned from future evaluations of the impact of the directives on the effectiveness of public procurement.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Albalade, D., & Bel, G. (2009). Regulating concessions of toll motorways: An empirical study on fixed vs. variable term contracts. *Transportation Research Part A*, 43, 219–229.
- Bajari, P., Houghton, S., & Tadelis, S. (2014). Bidding for incomplete contracts: An empirical analysis of adaptation costs. *American Economic Review*, 104(4), 1288–1319. <https://doi.org/10.1257/aer.104.4.1288>
- Baltrunaite, A., Giorgiantonio, C., Mocetti, S., & Tommaso, O. (2021). Discretion and supplier selection in public procurement. *Journal of Law, Economics, and Organization*, 37(1), 134–166. <https://doi.org/10.1093/jleo/ewaa009>
- Bandiera, O., Prat, A., & Valletti, T. (2009). Active and passive waste in government spending: Evidence from a policy experiment. *American Economic Review*, 99(4), 1278–1308.
- Bandiera, O., Bosio, E., Spagnolo, G., & Centre for Economic Policy Research. (2021). *Procurement in focus rules, discretion, and emergencies*. CEPR Press. https://books.google.fr/books/about/Procurement_in_Focus_Rules_Discretion_an.html?id=nrzJzgEACAAJ&redir_esc=y
- Banfield, E. C. (1975). Corruption as a feature of governmental organization. *The Journal of Law & Economics*, 18(3), 587–605.
- Baranek, B., & Titl, V. (2021). Political connections in public procurement. In O. Bandiera, E. Bosio, & G. Spagnolo (Eds.), *Procurement in focus: Rules, discretion, and emergencies*. CEPR Press.
- Beuve, J., & Saussier, S. (2021). Renegotiations and renewals of public contracts. *Review of Industrial Organization*, 59(3), 461–482. <https://doi.org/10.1007/s11151-021-09819-w>
- Beuve, J., Moszoro, M. W., & Saussier, S. (2019). Political contestability and public contract rigidity: An analysis of procurement contracts. *Journal of Economics & Management Strategy*, 28(2), 316–335. <https://doi.org/10.1111/jems.12268>
- Beuve, J., Moszoro, M. W., & Spiller, P. T. (2023). Doing it by the book: Political contestability and public contract renegotiations. *Journal of Law, Economics, and Organization*, 39(1), 281–308. <https://doi.org/10.1093/jleo/ewab039>
- Bosio, E., Djankov, S., Glaeser, E., & Shleifer, A. (2022). Public procurement in law and practice. *The American Economic Review*, 112(4), 1091–1117. <https://doi.org/10.1257/aer.20200738>
- Brogaard, J., Denes, M., & Duchin, R. (2021). Political influence and the renegotiation of government contracts. *Review of Financial Studies*, 34(6), 3095–3137. <https://doi.org/10.1093/rfs/hhaa093>
- Carril, R. (2021). *Rules versus discretion in public procurement* (Economics Working Paper). Universitat Pompeu Fabra.
- Cocciolo, S., Di Maro, V., & Samaddar, S. (2021). Public procurement at the onset of the Covid-19 pandemic. In O. Bandiera, E. Bosio, & G. Spagnolo (Eds.), *Procurement in focus: Rules, discretion, and emergencies*. CEPR Press.

- Coviello, D., & Gagliarducci, S. (2017). Tenure in office and public procurement. *American Economic Journal: Economic Policy*, 9(3), 59–105. <https://doi.org/10.1257/pol.20150426>
- Coviello, D., Guglielmo, A., & Spagnolo, G. (2018). The effect of discretion on procurement performance. *Management Science*, 64(2), 715–738. <https://doi.org/10.1287/mnsc.2016.2628>
- Coviello, D., Guglielmo, A., Lotti, C., & Spagnolo, G. (2022). *Procurement with manipulation* (CEPR Discussion Paper No. 17063).
- Crocker, K. J., & Reynolds, K. J. (1993). The efficiency of incomplete contracts: An empirical analysis of air force engine procurement. *RAND Journal of Economics*, 24(1), 126–146.
- Decarolis, F., & Giorgiantonio, C. (2022). Corruption red flags in public procurement: New evidence from Italian calls for tenders. *EPJ Data Science*, 11(1), 1–38. <https://doi.org/10.1140/epjds/s13688-022-00325-x>
- Decarolis, F., Spagnolo, G., & Pacini, R. (2016). *Past performance and procurement outcomes* (NBER Working Paper). National Bureau of Economic Research.
- Decarolis, F., Giuffrida, L. M., Iossa, E., Mollisi, V., & Spagnolo, G. (2020a). Bureaucratic competence and procurement outcomes. *The Journal of Law, Economics, and Organization*, 36(3), 537–597.
- Decarolis, F., Fisman, R., Pinotti, P., & Vannutelli, S. (2020b). *Rules, discretion, and corruption in procurement: Evidence from Italian government contracting*. NBER. <https://doi.org/10.3386/w28209>
- Deodato Dominguos, F., Heinrich, C. J., Saussier, S., & Shiva, M. (2023). *The interplay of discretion and complexity in public contracting and renegotiations* (Sorbonne Business School Working Paper).
- Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2002). The Regulation of entry. *The Quarterly Journal of Economics*, 117(1), 1–37.
- Engel, E., Fischer, R., & Galetovic, A. (2019). Soft budgets and endogenous renegotiations in transport PPPs: An equilibrium analysis. *Economics of Transportation*, 17, 40–50.
- Estache, A., & Saussier, S. (2014). Public-private partnerships and efficiency: A short assessment. *CESifo DICE Report*, 12(3), 8–13.
- Fazekas, M., Nishchal, S., & Soreide, T. (2021). Public procurement under and after emergencies. In O. Bandiera, E. Bosio, & G. Spagnolo (Eds.), *Procurement in focus: Rules, discretion, and emergencies*. CEPR Press.
- Fehr, E., Hart, O. D., & Zehnder, C. (2011). Contracts as reference points—Experimental evidence. *American Economic Review*, 101(2), 493–525. <https://doi.org/10.1257/aer.101.2.493>
- Frydlinger, D., Hart, O. D., & Vitasek, K. (2019). A new approach to contracts. *Harvard Business Review*.
- Grossman, S. J., & Hart, O. D. (1986). The cost and benefits of ownership: A theory of vertical and lateral intégration. *Journal of Political Economy*, 94(4), 691–719.
- Guasch, J.-L. (2004). *Granting and renegotiating infrastructure concession: Doing it right*. The World Bank.
- Guasch, J.-L., Laffont, J.-J., & Straub, S. (2008). Renegotiation of concession contracts in Latin America. Evidence from the water and transport sector. *International Journal of Industrial Organization*, 26, 421–442.
- Guasch, J.-L., Benitez, D., Portabales, I., & Flor, L. (2017). The renegotiation of public-private partnership contracts: An overview of the recent evolution in Latin America. In *Public private partnerships for transport infrastructure* (pp. 57–77). OECD. <https://doi.org/10.1787/9789282108130-5-en>
- Hart, O. D., & Moore, J. (1988). Incomplete contracts and renegotiation. *Econometrica*, 56(4), 755–785.
- Hoekman, B., & Tas, B. (2024). Discretion and public procurement outcomes in Europe/ast. *European Journal of Political Economy*, 82, 102525. <https://doi.org/10.1016/j.ejpoleco.2024.102525>

- Iossa, E., & Martimort, D. (2016). Corruption in PPPs, incentives and contract incompleteness. *International Journal of Industrial Organization*, 44, 85–100. <https://doi.org/10.1016/j.ijindorg.2015.10.007>
- Kang, K., & Miller, R. A. (2021). Winning by default: Why is there so little competition in government procurement? *The Review of Economic Studies*, 89(3), 1495–1556.
- Kelman, S. (1990). *Procurement and public management: The fear of discretion and the quality of government performance*. AEI Press. <https://doi.org/10.1787/5km91p7s1mxv-en>
- Le Squeren, Z., & Moore, J. (2015). *The permeability of public contracts: Empirical evidences from renegotiations in the French Car Park sector* (Chaire EPPP Working Paper).
- Macaulay, S. (1963). Non-contractual relations in business: A preliminary study. *American Sociological Review*, 28, 55–70.
- Malin, E., & Martimort, D. (2000). Transaction costs and incentive theory. *Revue d'Économie industrielle*, 92(1), 125–148. <https://doi.org/10.3406/rei.2000.1043>
- Moszoro, M. W., & Spiller, P. T. (2014). Third-party opportunism and the theory of public contracts: Operationalization and applications. In E. Brousseau & J.-M. Glachant (Eds.), *The manufacturing of markets: Legal, political and economic dynamics*. Cambridge University Press.
- Moszoro, M. W., & Spiller, P. T. (2019). Political contestability and public contracting. *Journal of Public Economic Theory*, 21(5), 945–966. <https://doi.org/10.1111/jpet.12325>
- Moszoro, M. W., Spiller, P. T., & Stolorz, S. (2016). Rigidity of public contracts. *Journal of Empirical Legal Studies*, 13(3), 396–427. <https://doi.org/10.2139/ssrn.2469270>
- OECD. (2023). *Government at a glance 2023*. OECD. <https://doi.org/10.1787/3d5c5d31-en>
- Palguta, J., & Pertold, F. (2017). Manipulation of procurement contracts: Evidence from the introduction of discretionary thresholds. *American Economic Journal: Economic Policy*, 9(2), 293–315.
- Ring, P. S., & Perry, J. L. (1985). Strategic management in public and private organizations: Implications of distinctive contexts and constraints. *Academy of Management Review*, 10(2), 276–286.
- Rose-Ackerman, S. (1999). *Corruption and government: Causes, consequences, and reform*. Cambridge University Press.
- Sarmento, J. M., & Cruz, C. O. (2018). Renegotiations of water concessions: Empirical analysis of main determinants. *Journal of Water Resources and Planning Management*, 144(11). [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000999](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000999)
- Saussier, S. (2000). Transaction costs and contractual incompleteness: The case of Électricité de France. *Journal of Economic Behavior and Organization*, 42(2), 189–206.
- Saussier, S., & Tirole, J. (2015). *Strengthening the efficiency of public procurement*. French Council of Economic Analysis. Les notes du conseil d'analyse économique.
- Schwartz, A., & Watson, J. (2012). *Conceptualizing contractual interpretation* (Research Paper). Yale Law & Economics.
- Spagnolo, G. (2012). Reputation, competition, and entry in procurement. *International Journal of Industrial Organization*, 30(3), 291–296.
- Spiller, P. T. (1995). A positive political theory of regulatory instruments: Contracts, administrative law or regulatory specificity. *Southern California Law Review*, 69, 477–515.
- Spiller, P. T. (2009). An institutional theory of public contracts: Regulatory implications. In C. Ménard & M. Ghertman (Eds.), *Regulation, deregulation, reregulation: Institutional perspectives* (pp. 45–66). Edward Elgar.
- Tadelis, S. (2012). Public procurement design: Lessons from the private sector. *International Journal of Industrial Organization*, 30(3), 297–302. <https://doi.org/10.1016/j.ijindorg.2012.02.002>
- Tirole, J. (1986). Procurement and renegotiation. *Journal of Political Economy*, 94, 235–259.
- Tirole, J. (2009). Cognition and incomplete contracts. *American Economic Review*, 99(1), 265–294.
- Williamson, O. E. (1985). *The economic institutions of capitalism*. The Free Press.
- Williamson, O. E. (1999). Public and private bureaucraties: A transaction cost economics perspective. *Journal of Law, Economics and Organization*, 15(1), 306–342.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Agricultural Contracts and Transaction Costs

27

Douglas W. Allen and Dean Lueck

Contents

27.1	Introduction	674
27.2	Basic Features of Agricultural Contracts	677
27.2.1	Land Leases	677
27.2.2	Equipment Contracts	679
27.2.3	Labor Contracts	681
27.3	Transaction Costs and Institutional Economics	682
27.3.1	Transaction Costs and Property Rights	682
27.3.2	Transaction Costs and Agriculture	683
27.4	Transaction Cost Explanations of Agricultural Contracts	684
27.4.1	Simple Contracts for Land	684
27.4.2	Cropshare Versus Cash Rent	685
27.4.3	Input Sharing in Share Contracts	687
27.4.4	Equipment Contracting and Custom Operation	687
27.4.5	The Role of Customs and Traditions in Contract Design	689
27.5	Transaction Costs and/or Risk Aversion	689
27.6	Conservation and Contract Design	692
27.7	Conclusion	693
	References	694

Abstract

Agricultural contracts are agreements between the owners of various inputs over the biological stages of food production. Biological stages of production are often heavily influenced by nature both in terms of random shocks and seasonal constraints. Nature can create large opportunities for transaction costs, constrain production specialization, and impose risk on individuals. This chapter focuses

D. W. Allen (✉)

Department of Economics, Simon Fraser University, Burnaby, BC, Canada

e-mail: doug_allen@sfu.ca

D. Lueck

Department of Economics, Indiana University, Bloomington, IN, USA

on the two main types of land contracts: cropshares and cash rents. It identifies a number of characteristics of these contracts and their evolution over time. We argue that a transaction cost framework remains the most robust theory to understand the structure of contracts and note that evidence from the past 20 years confirms this. Indeed, the transaction cost model has shown its usefulness in the new environmental literature on soil conservation in light of climate change. Finally, we argue that when market selection effects are considered, the transaction cost model is also consistent with some of the new findings regarding risk aversion and contract choice.

Keywords

Contracts · Farming · Cropshares · Cash rents · Biological stages · Risk aversion · Soil conservation · Transaction costs · Seasonality

27.1 Introduction

The economic analysis of agricultural contracts has a long and storied history, reflecting farming’s relatively simple setting, its amenability to abstract modeling, and its familiar history, institutional, and empirical features. Furthermore, agriculture remains a dominant industry in much of the world, and the role of contracts in the management of soil and other land attributes have become even more important within the context of changing climate.

Agriculture typically refers to a sequence of production stages, which typically are thought to span the biological stages “controlled by the farm.”¹ Figure 27.1 presents various stages of production for a generic crop—animal or plant—and draws attention to the arbitrariness of what is meant by “agriculture.” Figure 27.1 also shows that the range of farm production activities has narrowed over time. In

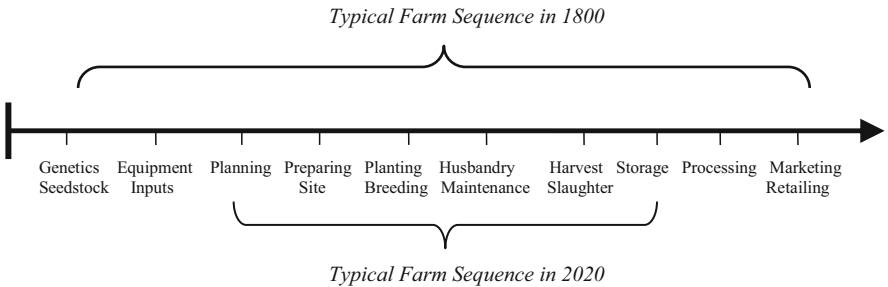


Fig. 27.1 The extent of agricultural production

¹ Among agricultural economics, the term “production agriculture” is often used.

1800 the typical North American farm would have broken land, produced seeds, and carried production through to some form of processing and marketing. Today, what we normally think of as “the farm” is a firm that essentially controls just the narrowest growth-based biological stages of production.

In this chapter, we consider the contracts that exist over the biological stages. Although we briefly discuss agricultural contracts over a broad number of stages, and some attention is given to international contracts, the focus is on farming contracts found in North America. This does not mean that agricultural contracts are not an important issue in other parts of the world but instead reflects our own expertise and knowledge.

Because agriculture is biological production, it is heavily affected by nature. The effects of nature in turn shape the structure of agricultural contracts because these effects influence the transaction costs involved. “Nature” refers to the aggregate of natural forces that can influence the outcome of agricultural production and includes the forces of climate and weather, pests, seasons, geology, and hydrology. In the aggregate, nature not only generates random havoc and blessings on crops and livestock, but nature also restricts the ability of farmers to specialize and exploit the gains from specialization.

In many cases, the constraints of nature, coupled with social and economic forces, lead to rather simple contracts. For example, contracts between farmers and landowners tend to be enforced by reputation. When agriculture includes the non-biological or non-growing stages of production, or when production is designed in such a way that nature’s role during the growing stage is minimized, contracts tend to become more complex, and the organization of the farm tends to move from family farms using simple contracts, to corporate firms with complicated contracts between relatively small, often family-based growers.²

Contracting in agriculture is not, of course, limited to agreements between farmers and landowners. There are contracts for equipment, labor, marketing, production, and services.³ These contracts can be simple and short term, or they can be complicated and long term. Table 27.1 summarizes some of the features of these contracts and shows where they are most found.⁴ For example, marketing

²For example, the introduction of antibiotics, which allowed for the intensive indoor husbandry of cattle, hogs, and poultry, dramatically limited major natural inputs (e.g., disease) in the production process. This allowed for corporate governance and complicated incentive contracts for the local growers. For other examples of complex agricultural production, see Frank and Henderson (1992), Fulton (1995), Knoeber (1989, 2000), Kliebenstein and Lawrence (1995), Knoeber and Thurman (1994), Lazzarini et al. (2001), Martinez and Reed (1996), Ménard (1996), Sauvee (2002), and Ménard and Klein (2004).

³These contracts use various methods of payment (e.g., hours for tractors, revenue share for marketing) and impose various duties on both farmers and other contracting parties. These duties include sharing input costs, using specified techniques, and performing tasks at specified times.

⁴Roy (1963) gives an early account of contracting practices in agriculture. See Perry et al. (1997) for a more recent account, which finds that such contracts cover almost one-third of the value of farm production in the United States.

Table 27.1 Features of typical agricultural contracts

Type	Features (parties, payment, duties)	Examples
Equipment	Farmer rents equipment on a time basis (day, month, year) or on an hourly rate of use	Field implements, tractors
Labor	Farmer hires labor for skilled and unskilled work. Payment can be hourly, monthly, or based on piece rates	Equipment operation, harvesting, field work
Land	Farmer rent land in return for cash or share of the crop. Input costs are sometimes shared	Small grains, pasture, hay, row crops
Marketing	Farmer agrees to deliver crops of a specific quantity and quality to a processor. Farmer and marketing firm share the revenues. Farmer control overall production and owns the crop	Apples and other fruit, dairy, sugar, cattle, vegetables
Production	Farmer ("grower") agrees to produce a crop under the direction of another firm. Farmer will be required to use certain inputs and techniques and may share costs. The contracting firm will generally control production and own the crop	Seed, vegetables, poultry, swine
Service	Farmer hires firm to perform specific tasks. Typically, the firm provides both the labor and the specialized equipment required	Harvesting, pest control, cattle feeding

contracts for fruits, dairy, sugar, and many vegetables are often established between the farmer and a processor. In these contracts, although the farmer (or "grower") typically controls production, he is partly integrated with the processing firm and agrees to deliver a specific quantity and quality of crop, usually at a specific time. For poultry, swine, and some vegetables, farmers enter production contracts where the farmer agrees to produce a crop under the direction of another firm. In these contracts, the farmer is often required to use certain inputs and techniques and may share costs with the firm.

In this chapter, our focus is on contracts for land because this remains the area of most interest to economists. We begin with some general descriptions of the structure and extent of contracts for land and equipment. Next, we describe the transaction cost framework for examining such contracts. It is our contention that this approach remains the best and most robust theory of agricultural contracts. We follow this description with a series of contract observations that are best explained by the mitigation of transaction costs.

Two main streams of literature in agricultural contracts have developed since our original (Allen and Lueck 2005) chapter. First, several papers reexamined the role of risk aversion in the context of endogenous matching between farmers and landowners. Second, the literature on soil conservation and environmental management has been drawn to the transaction cost problem of a farmer exploiting the land attributes of the landowner. Interestingly, evidence from this second literature has findings relevant to the role of risk aversion in contract design.

27.2 Basic Features of Agricultural Contracts

27.2.1 Land Leases

Farmland leases take two basic forms: cropshare and cash rent.⁵ A *cropshare* contract is so-called because the two parties share the harvested output, not the crop revenue, and they tend to have no additional cash side payments.

The farmer may pay for all the inputs (e.g., seed, fertilizer, harvesting) or share these input costs with the landowner. Sharing input costs is almost always done in a dichotomous fashion, with the input share either equaling the output share or equaling 100% for the farmer. This means that if the farmer and landowner are sharing output 50/50, then when inputs are being shared, they are also shared 50/50. If inputs are not shared, then the farmer almost always pays for them. When inputs are not shared, then the share of output to the farmer is always higher.

The primary alternative form of land contract in agriculture is a *cash rent* contract. In a cash rent contract, the farmer pays the landowner a fixed monetary sum for the exclusive use of the land for a specific period. Usually, the payment is made before the crop is harvested, but sometimes a series of payments is spread out over the year. Unlike cropshare contracts, in cash rent contracts, farmers typically do not share input costs with the landowner. Occasionally cash rent contracts will have an “adjustment” clause in them, increasing the rent if there is an exceptionally good or “bumper” crop.⁶

Although cropshare contract shares are quite limited to a few discrete fractions (0.5, 0.6, 0.75) and are usually no higher than $s = 0.75$ or $s = 0.8$, cash rents vary almost continuously and are more sensitive than cropshare contracts to differences in economic fundamentals, such as land quality, input prices, and land size. In this respect, a cash rent contract is *not* just an extreme cropshare contract where the share equals 1; a smooth continuum of share contracts does not exist.

In many non-agricultural settings, exchanges are governed by rather complicated contracts.⁷ Farmland contracts, on the other hand, are simple, often oral, and tend not to stipulate in detail how the land will be farmed; rather, they require the farmer use the land in a “good and husband-like” manner, a term used by lawyers and judges.⁸ Furthermore, these contracts are most often annual agreements, subject to automatic renewal, unless one party makes an early commitment not to renew. Sometimes the agreements are for several years, but rarely are they longer than 5 years. They are, however, typically renewed for extended periods, even up to 30 years. Most

⁵ In the law of property, a temporary agreement for the exclusive and possessory use of land is called a “lease”; thus, we use the terms contract and lease synonymously for farmland.

⁶ Allen and Lueck (2002) found that such clauses accounted for about 10% of the cash contracts in their data.

⁷ There are also simple contracts outside of agriculture, as Macauley (1963) first pointed out.

⁸ Goodhue et al. (2003) examine contracts for wine grape growers in California and find oral agreements are common, but are replaced more often with written agreements for high-quality grapes.

Table 27.2 Some features of modern farmland contracts

	United States 2014	Canada 2011	Ireland 2010	England 2010	France 2010	Belgium 2010
<i>General information,</i>						
Acres leased (%)	39	38.5	16	32	75	66
Average lease size (acres)	60	236				
Average land value (\$/acre)	\$4100	NA				
<i>Cash leases,</i>						
All leases (%)	70	91	57			
Leased acreage (%)	69	23.8	59			
<i>Share leases,</i>						
All leases (%)	15	9	21			
Leased acreage (%)	21	16.5	24			

Source: Statistics Canada, Census of Agriculture, 1976–2011; (US Census of Agriculture 1999, 2014; Swinnen et al., 2016)

interesting is that these contracts tend not to contain detailed descriptions regarding duties, techniques, or other farming practices.⁹

Both cash rent and cropshare leases are very common in North America.¹⁰ Table 27.2 shows some basic features of modern farmland contracts. The table shows that farmland leasing accounts for about 40% of all acres farmed in both Canada and the United States and a wide range across four European countries.¹¹ In terms of number of leases, about 70% of leases in the United States are cash rent in 2014 and 15% are cropshares.¹²

Cropshares vary, but most observed shares are from 50% to 75% and tend to be simple fractions (e.g., 1/2, 2/3, 3/5, and 3/4). Worldwide, 50% appears to be the most common output share, although in North America the three most common shares (the share kept by the farmer) are 50%, 60%, and 67%, which are often called “halves,” “fifths,” and “thirds” by farmers. Table 27.3 shows the discreteness and frequency of the shares to farmers in cropshare contracts across several regions in North America (Allen and Lueck 2009, Young & Burke, 2001).

⁹ Agricultural land leases are often between family members. Onofri et al. (2023) report that across Europe most of these features of American land leases carry over. Contracts tend to be informal, normally renewed for long durations, and family farms dominate.

¹⁰ Gray et al. (1924) note that leasing was widespread in early twentieth-century America and that cash rent and cropshare contracts were the dominant forms.

¹¹ The wide variation in land leasing in Europe is mostly driven by different types of agriculture and government regulations on land markets (Swinnen et al., 2016).

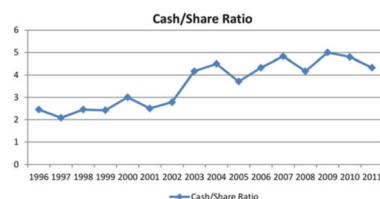
¹² The remainder are divided between cash leases with adjustment clauses (just 5.6% of the cash contracts), cash-share combination leases (about 10.5% of all leases), and “other” leases (about 10.5% of all leases). See 1999 Agricultural Economics and Landownership Survey, Table 98.

Table 27.3 Frequency (%) of farmer shares in cropshare contract terms across regions

<i>Farmer share (%)</i>	British Columbia 1979	British Columbia 1992	Louisiana 1992	Nebraska South Dakota 1992	Kansas 2000	Illinois 1995
9/10 (90)	5	4.4	0.3	0.12	0	0
17/20 (85)	7	20	0.6	0	0	0
5/6 (83)	0	0	12.6	0	0–0.6	0
4/5 (80)	21.9	8.9	38.6	0.12	0–1.2	0
3 / 4 (75)	26	15.6	23.1	1.40	0.4–1.2	0
2/3 (67)	19.8	22.2	0.9	32.8	68–79	9.7
3/5 (60)	1.2	13.3	6.8	30.16	11–15	6.7
1/2 (50)	11.2	6.7	2.2	30.92	9–15	82.3
2/5 (40)	0	0	0	1.32	0–2.1	2
Observations	242	45	324	2424	1449	935

Source: Allen and Lueck (2009)

Fig. 27.2 The growth of cash renting in the United States.
(Source: Allen and Borchers 2016)



Three important characteristics are shown in Table 27.3. First, shares below 50–50 for the farmer are less than 2%. Second, a variety of dominant share rates exist across different regions. And finally, all shares are simple fractions (e.g., ratios of small whole numbers). In each case, the shares are customary in that they are rigid locally but vary across regions.

Over the past 30 years, the most significant change in land leases in North America has been the rise of cash rent contracts. Figure 27.2 shows the growth of cash rent contracts relative to cropshare contracts in the United States (Canada is similar). The data for the figure includes all regions in the United States, nine different crops (cotton, rice, soy durum/spring/winter wheat, barley, corn, and sorghum), and runs between 1996 and 2011. The figure shows that the ratio of cash rent contracts to cropshare contracts (by acres) increased from just above two to around five—almost a doubling of the relative importance of cash rent contracts over this period.

27.2.2 Equipment Contracts

Although farmland leases are the most common agricultural contract analyzed by economists, farmers also contract for the use of equipment—both general (e.g.,

tractors) and specialized (e.g., swathers). The value of a major piece of equipment relative to a typical plot of land on a farm varies by location and type of farming. In some situations, farmers might have more wealth tied to their farm equipment than to other assets. For example, a quarter section (160 acres) of dry wheat land in the Canadian Prairies has an average worth of \$2250 (US) per acre or \$360,000 (Farm Credit Canada, 2022). In 2022, US cropland values averaged over \$5000 per acre and ranges from \$2000 to \$6000 per acre on the Great Plains states of Kansas, Nebraska, South Dakota, and North Dakota.¹³ A large modern combine harvester could be worth more than \$400,000, and a farmer might own more than one. Thus, equipment contracts are often as important as land contracts in terms of the value of the contracted assets.

Although data are limited, equipment is much more likely to be owned than rented and is much more likely to be owned compared to farmland. Farmers often hold large inventories of infrequently used equipment like cultivators, harvesters, and sprayers. In this sense, it is common to hear that farmers own “too much” capital. Many of these expensive assets are used for surprisingly short periods and are simply held in inventory for the remainder of the year.

When farmers rent equipment, such as combines and tractors, they are typically charged by the hour or some other unit of time (e.g., day, week, and month).¹⁴ The firms that lease out equipment tend to be local farm implement dealers that deal primarily in the sale of farm equipment. Equipment leases are usually written and usually last less than a year.¹⁵ In fact, equipment is seldom rented for an entire season and is often rented for only a particular stage of production. This means that for certain kinds of equipment, during crucial stages, farmers in a locale demand the equipment at the same time.

Even though farmers have underutilized equipment, they often use multiple pieces of the same or similar machines as well (e.g., combines, tractors, trucks). Farmers who use more than one machine are much more likely to rent the secondary machines than own them. Farmers often attempt to lower capital costs by sharing equipment with neighbors. However, since farmers in a locale often need the equipment at the same time, the types of shared equipment are limited to those machines for which timeliness costs are small.¹⁶ Equipment that is specialized to a crop (e.g., cotton gin) or to a specific stage (e.g., planting, harvesting) is seldom governed by contract.

Of particular interest among equipment contracts is the practice of custom contracting. In agriculture, the term “custom” refers to contracting for equipment

¹³See USDA Land Values Summary 2022 at https://www.nass.usda.gov/Publications/Todays_Reports/reports/land0822.pdf.

¹⁴Tractors prices vary with horsepower, so rates are often discussed in terms of “horsepower hours.” See Edwards and Meyer (1986) and Pflueger (1994) for some details on modern farm equipment leasing.

¹⁵Most long-term (greater than 1 year) equipment leases have options to buy (Pflueger, 1994).

¹⁶The contractual implications of timing are discussed in Sect. 27.3).

and an operator (often the owner) at the same time. In a typical case, a farmer may contract for a plane to do crop spraying and hire the pilot as well. Farmers also hire custom firms for baling, cultivating, fertilizer and pesticide application, feed grinding, fencing, hauling, land clearing, planting, plowing, and seed cleaning (Eaton & Shepherd, 2001). By far, the most important custom contracting is for grain harvesting. And by far the most important custom grain harvesting is the 1000-mile south-to-north migration of harvesters that takes place in the wheat belt of the Great Plains. A 1971 study by the USDA showed that nearly 3500 custom combiners or “cutters” harvested 35% of the wheat on the Great Plains (Lagrone & Gavett, 1975). In 1997, roughly 2000 custom cutters harvested one-half of the United States’ wheat crop (DuBow, 1999). Overall, however, all custom work comprises just 2% of all US farm expenses.¹⁷ Although some cutters work only locally, more than 90% make the long interstate journeys.

Contracts between custom combiners and farmers have a structure that originated in the 1940s when the industry had its first major expansion. Since that time, cutters have been paid according to a three-part formula that includes a per-acre fee for cutting, a per-bushel fee for hauling grain to a local storage site, and a per-bushel fee for high yield crops (usually over 20 bushels per acre).¹⁸ In unusual cases—drought, hail, long hauls, or wind—special harvest rates would be developed to suit both parties. As with agricultural contracts in general, farmers and cutters rely heavily on verbal agreements, enforced with handshakes and the possibility of renewal.¹⁹ As Isern (1981, p. 92) notes:

The unwritten code of custom combining was a flexible one, but the person who stretched it beyond reason suffered the consequences. The custom cutter who failed to live up to his obligations to a farmer found it hard to obtain work in the locality the next year. Likewise, if a farmer renegeed on an agreement, the word spread among custom cutters working the area, and the farmer might be left with no harvesters at all.

27.2.3 Labor Contracts

Of all contracts related to agriculture, labor contracts seem to resemble contracts outside agriculture the most. For land contracts, sharing is important, and for equipment contracts, underutilization is obvious; yet nothing particularly striking or unusual is apparent in agricultural labor contracts. No doubt, this results from the

¹⁷This shows how distinctive the custom grain harvest is. See 1997 Census of Agriculture, Table 3 “Farm Production Expenses.” The census does not collect data on specific custom services. Custom harvesters have a professional association, the US Custom Harvesters, which reports 649 operators in 2023.

¹⁸According to Isern (1981), this formula was typically 3/3/3 when it first emerged in the early 1940s. The high yield payment compensates the cutter for slower progress (in terms of acres) with higher yields and greater wear and tear on the machines.

¹⁹Our information on this is limited, but contracts for equipment rental from farm implement dealers are more formal, written agreements, much like those used when renting a car.

limited role that hired labor plays in most farming contexts. Farming is dominated by family farms where the farmer (or members of his family) provides most of the labor services.²⁰

Allen and Lueck (2002) found a pattern familiar to most labor contracts. When laborers engage in general farm duties, such as cleaning barns, mending fences, and operating basic equipment, wages are the dominant form of compensation. When workers assume supervisory or management roles, such as foreman, office worker, or manager, they are often paid by monthly salary. Finally, when workers are hired for harvest, both wage and piece rate contracts are used.

27.3 Transaction Costs and Institutional Economics

When transaction costs are zero, it does not matter how ownership of inputs and outputs is distributed by the terms of a contract. Farmers can control land through cash leases, share contracts, or ownership; farmers can own or rent their equipment in any proportion; and labor produces an equal amount of output under any compensation scheme. These applications of Coase (1960), of course, apply to other aspects of agricultural organization. If transaction costs are zero, then a farm could be a family-run sole proprietorship or a large-scale corporation. Farms could be integrated completely or disintegrated to the point of owning a wheat field for 1 day. Explanations of differences in how contracts assign property rights, therefore, *must* include differences in transaction costs. Institutional economists not only recognize the *necessity* of transaction costs, but their models also tend to treat transaction costs as a sufficient explanation.

27.3.1 Transaction Costs and Property Rights

Transaction costs are the costs of enforcing and maintaining property rights and include the deadweight losses that result from enforcing property rights (Allen, 1991, 2000), and the fundamental transaction cost hypothesis is that institutions, organizations, and contracts are designed to maximize wealth net of the relevant transaction costs. Therefore, within the context of agricultural production, various parts of the exchanges are managed in different ways. That is, various exchange margins will involve reputations, the law, third-party agencies, and private contracts.

Farm contracts are designed to mitigate transaction costs given the constraints imposed by the particular farming technology, the role of nature, and the potential gains from specialization, as the parties attempt to police certain interactions with

²⁰The main exception to this would be migrant farm labor for the harvesting of fruits and vegetables, especially along the Pacific Coast.

each other.²¹ Farmers can hide bales of hay that were intended to be shared with landowners; harvest crews can arrive late, causing a reduction in crop value; farmers can exploit landowner assets like land attributes and equipment; and of course, workers can shirk. Efforts to engage and prevent these activities, along with any lost gains from trade that result, are transaction costs.

27.3.2 Transaction Costs and Agriculture

Four transaction costs factors are important for the study of agricultural contracts. First, *agricultural inputs and outputs are complex, in the sense that they are composed of many attributes*. When goods are complex, they create an opportunity for transaction costs to arise for every attribute. This subsequently allows for ownership to be divided over the various attributes because multidimension goods are costly to measure (Allen & Barzel, 2023). Land, for example, is comprised of features that include terrain, nutrients, moisture, and soil type, but contracts for land typically only specify the amount of land in terms of surface area. Many, if not most, of the attributes of the land are not explicitly specified and can then be used by either party. Different ownership and contract types affect the various attributes in different ways, creating trade-offs.

Second, *uncertainty is an input into production*. The biological element of farm production means that nature is always present. Nature causes uncertainty, and this is the source of moral hazard on both sides of the contract. Nature allows one party to exploit an exchange at the expense of the other party because it masks their actual effort. This factor is important in agriculture because weather, pests, and other natural phenomena contribute so much to the final output. In a land lease, for example, uncertainty from weather and other natural forces means that the farmer has the opportunity to “exploit” the landowner in several ways: under-supplying effort; overusing soil quality attributes; and underreporting the shared crop, to name a few.

Third, nature also has a *systematic seasonality*. Seasonality refers to crop cycles, the number and length of stages, and timeliness costs. For any given crop or livestock, there is a natural order of production: planting comes before cultivation, which is always followed by harvest. For winter wheat in Texas, this means planting in the fall and harvesting in late spring or early summer. The predictable aspect of nature—its seasonality—limits the degree to which farmers can specialize in production (Allen & Lueck, 1998, 2002). Seldom is a farmer able to specialize in one task, due to seasonality. A farmer, who only plowed or only planted, would be unemployed for most of the year. Most types of farm output are greatly restricted by

²¹ This assumes that natural selection has resulted in the most valuable contract or organization being chosen (Alchian, 1950; Coase, 1960). The contracts observed are not the result of inert culture or custom, nor are they the result of one-sided bargaining. In the transaction cost approach, competition is assumed to extend to the nature of the contract. Thus, only those contracts which maximize the gains from trade will survive.

nature in terms of how they can be produced. Both plant and animal crops have “growing seasons” that restrict the nature of farm production. As a result, farmers are seldom able to realize potential forms of economies of size. And since individuals are engaged in many tasks, monitoring costs are high. Seasonality also influences the use and type of contracts because its presence makes the net gain from ownership and integration larger than it would be, otherwise.

Fourth, farming involves *timeliness* costs: the forgone values from lower crop yields or quality that result from not completing tasks (e.g., planting, cultivating, spraying, harvesting) at the optimal time (Short & Gifu, 1991). In many types of agriculture, these costs are substantial; waiting a day to harvest wheat might mean a complete crop loss from a hail or windstorm. The structure of contracts and farm organization is thus expected to create incentives to minimize these costs.²²

27.4 Transaction Cost Explanations of Agricultural Contracts

The transaction cost framework has been successful in explaining agricultural contracts. We report some of the basic findings below and add case study and historical data as well. In the process, we document the history of economic thought on the various contract questions.

27.4.1 Simple Contracts for Land

Why are farmland contracts so simple, even though the value of the assets at stake is quite large? Transaction cost features of agricultural contracts explain the seemingly naive contracts. First, specific assets (Klein et al., 1978; Williamson, 1979) are often relatively minor, so that long-term contracts are not needed to prevent holdup problems.²³ Second, in close-knit farming communities, market enforcement of contracts via reputation (Klein & Leffler, 1981) serves to reduce the need to use detailed written contracts.²⁴ Third, in areas where farming has been long practiced,

²² Allen and Lueck (2002) discuss the implications of timeliness costs.

²³ In cases where specific assets are important, long-term contracts and even vertical integration dominate (Crocker and Masten, 1988; Joskow, 1987). In farming specific investments are often made with irrigation. An important special case of specificity relates to “time specificity,” particularly in the harvest and delivery of perishable products. Specific investments may be more important with industrial farming methods, and in these contexts, we do not expect simple contracts. Again, these issues arise more in stages outside the biological stages we focus on. One area where specific asset might be important is for orchards in rather uneven terrain (e.g., Columbia River region in the Pacific Northwest). In this case, the subtleties of aspect (i.e., the land’s location with respect to the sun), elevation, and slope may generate specific knowledge about a plot of land that only a long-term farmer might have.

²⁴ Ellickson (1991) studies enforcement in the shadow of the law among members of homogeneous groups.

the common law has developed default rules that define the behavior of farmers and landowners, again rendering detailed contracts needless.²⁵

Even farmland contracts can be quite detailed when the conditions for simple contracts are not met. Consider the case of land leases on the Great Plains during the late 1800s and early 1900s. During this time, large so-called Bonanza farms arose in the Red River Valley along the Minnesota–North Dakota border, and their owners often leased out land in relatively small plots (e.g., 100–200 acres) (Drache, 1964). In sharp contrast to the simple (annual and oral) contracts that dominate modern farmland leases, these leases were remarkably detailed and always seem to have been written. Typically, these contracts were several single-spaced typed pages specifying the terms of payment; the method, depth, and time of plowing; the amount of fallow land; and the time and type of seeding (Allen & Lueck, 2002). The relative complexity of these contracts is consistent with the discussion above. In this region, the common law was simply undeveloped and lacked default rules, settlements were still being established and communities had not yet developed enough to exploit reputations, and so parties needed to specify, in detail, the sorts of things that would now simply fall under the category of good husbandry.

27.4.2 Cropshare Versus Cash Rent

Perhaps the oldest contractual question studied by economists has been: Under what conditions would a farmer and landowner contract for land with a cash rent or cropshare contract? From Adam Smith’s famous share-tax analogy, to Alfred Marshall and other neoclassical economists, the dominant view came to be that cropshare contracts were an inefficient choice compared to cash rent. Johnson (1950) questioned this view: if cropsharing is so inefficient, then why were so many Iowa farmers and landowners using it? Not until Cheung (1969) extended the Coase Theorem (1960) into share cropping did the modern analysis begin. Cheung (1969) and Stiglitz (1974) famously created the basic principal-agent framework of contract choice which depended on a trade-off between incentives and risk aversion.

A major shift came when Eswaran and Kotwal (1985) introduced a multiple margins moral hazard model with risk-neutral contracting parties. Other similar risk neutral models (Holmström & Milgrom, 1991; Allen & Lueck, 1992a, b; Laporini & Parigi, 1996; Prendergast, 2002) showed the power and simplicity of transaction cost models to explain contract choice.

Within this risk-neutral, multiple margins framework, the choice between cash rent and cropshare contracts can be explained by focusing on simple production

²⁵ Burkhart (1991) notes that farm leases contain an implied covenant to work the farm in a “farmer-like manner”—meaning that if the tenant does not use good farming practices, the covenant is breached and the landlord can recover damages. What constitutes a “farmer-like manner” can be shown by custom, practices of area farmers, or county extension agents.

trade-offs. For example, Allen and Lueck (2002) show that cash rent contracts provide perfect labor incentives for farmer effort, but because the farmer's rent is based on the amount of land farmed and not the amount of soil exploited, the cash rent contract provides an incentive to overuse the land. In particular, farmers have an incentive to exploit the valuable water and nutrient attributes, which will enhance their crops. Cropshare contracts, on the other hand, generate moral hazard costs for farmer effort, but this implies that the farmer has less incentive to overwork the land. This reduction in soil exploitation is the benefit of the cropshare contract. Cropshare contracts do not always dominate cash rent contracts because they also contain an incentive for the farmer to underreport the crop. For example, when a farmer must pay the landowner every other bale of hay in a 50–50 contract, every bale he underreports is an implicit increase in his share.²⁶

This framework also explains much of the contract choice found in other parts of the world. For example, a study of historical vineyard contracts in Spain (Carmona & Simpson, 1999) shows that share contracts dominate because of the incentives they provide for proper long-term care of the vines.²⁷ Similarly, Dubois (2002) finds that in the Philippines, corn—a crop that is hard on the land—is more likely to be cropshared than are the main alternatives of rice and sugar. Indeed, this view explains Adam Smith's long-standing observation that sharing was more common in France and southern Europe than in England. In northern Europe, the most common type of farming would have been of small grains and grass crops. In Southern France, Italy, and Spain, grapes, olives, and fruit were much more important. With fruit trees, the share contract prevents the farmer from exploiting the tree or vine asset over the short term. With grapes and fruit, pruning can be done in certain ways to increase the short-run volume of fruit, but which, over time, kill the tree or drastically reduce the long-term productivity of the tree. Sharing lowers the incentive of the farmer to do this in the same manner that it lowers the incentive to exploit soil attributes. The observations about European land leasing, first made by Adam Smith and John Stuart Mill (1871), thus suggest that land contracts were chosen in response to the costs of enforcing contracts over assets with many attributes. The dominance of cash rent contracts in England and northern Europe can be explained as the result of the relative dominance of small grain and grass farming. Similarly, the dominance of cropsharing in the Romance countries can be explained as the result of the relative dominance of orchard crops.²⁸

²⁶ Other transaction cost models, for example, Allen and Barzel (2023), are based on moral hazard on both sides of the transaction.

²⁷ Galassi (1992) and Hoffman (1984) are studies of historical European contracts that are also part of this literature. Allen and Lueck (2019) also show how transaction costs explain the different organization found in vineyards versus wineries.

²⁸ As we note in Sect. 27.2, just over 10% of all farmland leases in the United States are a cash-share combination. Our data are not detailed enough to examine these contracts and we are not aware of any study that attempts to explain the use and structure of such contracts. It thus remains an important open area for study.

27.4.3 Input Sharing in Share Contracts

An important feature of share contracts is whether or not the inputs are shared. Beyond land and labor, seed, fertilizer, machines, water, pesticides, and a host of other inputs are also required, and these other inputs are often shared.²⁹ In one of the few early studies of input sharing, Heady (1947) argued that sharing input costs in the same proportion as the output share offsets the moral hazard effect of the share on that input.³⁰ Recall that a share on output lowers the marginal product of every input, which is the “tax” effect introduced by Adam Smith. When an input cost is shared, this lowers the cost of the input and acts as an input subsidy. When the two shares are the same, the tax and the subsidy completely offset each other, and the input is used at its first-best level. Heady’s model predicted that *all inputs should be shared* in the same proportion as the output. But this is refuted: inputs are either shared in the same proportion as the output, or they are not shared at all.³¹

When the costs of monitoring shared inputs are incorporated into the analysis, this dichotomy can be explained. Whenever inputs are shared, there is an incentive for the farmer to *overreport* the actual cost of the inputs. Thus, input sharing requires monitoring effort on the part of the landowner. When these costs are high, the landowner opts to have the farmer pay for all input costs, and the share of output to the farmer is adjusted upward. When these monitoring costs are low, the inputs are shared in the same proportion as the output. For example, when a farmer has more than one plot of land, his cropshare contracts tend not to share input costs, so the landowner can avoid the potential for input cost shifting away from the leased plot (Allen & Lueck, 1993, 2002, Chap. 5).

27.4.4 Equipment Contracting and Custom Operation

For equipment, the main question is not contract choice per se but, rather, the choice between ownership and contracting for the use of equipment. Both ownership and contracting have costs: ownership costs include the foregone gains from specialization and the costs of borrowing if there are wealth constraints; contracting costs include moral hazard and timelines—that is, the reduction in output when tasks are not undertaken at the optimal time.³² Since equipment is often specialized and used

²⁹The cost of the farmer’s labor and the cost of the land’s attributes are never shared

³⁰Other studies include Bardhan and Singh (1987), Berry (1962), and Braverman and Stiglitz (1986).

³¹Braverman and Stiglitz (1982, 1986) reach similar conclusions in a principal-agent framework.

³²Timeliness costs, though not often given much attention by economists (an exception is Masten et al., 1991), are considered very important in the agricultural literature. See Edwards and Boehle (1980), Hopkin (1971), or Short and Gifu (1991).

for specific stages of production, it is extremely costly to use rental markets.³³ As a result, the farmer often owns even seldom-used equipment.³⁴

For example, trucks and tractors, used for many general farming tasks, are more likely to be contracted than owned, whereas specialized equipment, such as a combine, are more likely to be owned than rented (Irwin and Smith 1972). Farmers who use more than one piece of a particular type of equipment are more likely to contract for one piece than farmers who use just one piece. This evidence is consistent with the importance of timeliness costs in determining the ratio of contracted to owned assets.³⁵

Custom harvesting provides an example of the trade-offs between specialization gains, moral hazard costs, and timeliness costs. The benefit from contracting for custom combining is, of course, derived from the intensive use of highly specialized equipment with skilled operators. Typical wheat farmers may use their own combines for just 20 days each year. A custom cutter (or wheat harvester) could, however, by moving north with the ripening wheat, use his combines for 100–150 days each year. Williams (1953, p. 53) notes these tremendous gains to specialized equipment:

The economic keynote of the cutters' activities is the ability to get maximum utilization of their expensive and complicated specialized machinery, as they follow the progressive ripening pattern of the south-to-north contour of the wheat belt.

The gains from specialized labor are also important. Combine operators and truck drivers repeatedly working long seasons are more specialized than any farmer and his short-term help. These specialized operators are extremely knowledgeable in the use and maintenance of this expensive equipment.

The gains from specialized machines and labor would seem to imply that contracting for harvesting would dominate wheat and grain farming. Timeliness costs for a wheat harvest, however, are severe and limiting. For any given wheat farmer, the window for optimal harvest is just a few weeks, and standing grain is extremely vulnerable to hail, rain, and wind. How do migratory custom cutters and the farmers who hire them reduce these timeliness costs enough to sustain a viable industry? The answer is found in the geographic and climatic contiguousness of the Great Plains. Custom cutters know they can reliably follow the ripening wheat north in order to make intensive use of their expensive equipment. The exceptional circumstances of the contiguous wheat harvest of the United States point to how rare yearlong custom contracting is. For instance, wheat is also an important crop in California and Washington, but

³³Studies of ownership and contracting in trucking show similar empirical findings Baker & Hubbard, 2003, Nickerson & Silverman, 2002).

³⁴Although this has started to change, and some of the largest equipment can now be rented by the day.

³⁵Allen and Lueck (2002) also find that wealthier farmers are more likely to own equipment.

the lack of a long and continuous harvest has limited the emergence of large custom harvest industry.

27.4.5 The Role of Customs and Traditions in Contract Design

As noted in Sect. 27.2, cropshare contracts typically use a rather small set of fractional sharing ratios that are seemingly resilient to changes in underlying economic forces, such as differences in land and labor quality. Allen and Lueck (2009) show that cropshare contracts automatically adjust for changes in many fundamentals, and therefore, shares are less sensitive to the changes. In this sense, it “appears” that customs play a role, even though they do not.

Allen and Lueck focus on the distinction between the use of simple discrete fraction terms in cropshare contracts and the nearly continuous payment terms used in cash rent contracts. They show that the pattern of shares is best explained as a response to moral hazard problems spread over large numbers of inputs; that is, as the number of inputs increases, the optimal share becomes a simple fraction. Their transaction cost model explains the pattern of shares, the difference in flexibility with cash rent contracts, and the lower bound on shares. Empirical analysis that tests the model implications finds a wide range of support for the model based on moral hazard and measurement costs.³⁶

27.5 Transaction Costs and/or Risk Aversion

The transaction cost approach assumes risk-neutral parties and relies on a trade-off between different incentive margins to explain contractual terms. Since Cheung (1969), Stiglitz (1974) and Newberry and Stiglitz (1979) the main alternative approach has been to assume contracts are designed to balance risk against moral hazard incentives.³⁷ Huffman and Just (2004) have the most general model which allows for landowner and farmer heterogeneity in productivity, cost of effort, reservation levels of utility, and risk preferences. With this large set of parameters, they explain a wide range of stylized facts. The observation that cropshare contracts are more common in developing rather than developed economies is explained in their model by arguing that in the developing world farmers are more risk averse than landowners but that in the developed world farmers are less risk averse than landowners. Huffman and Just similarly offer an explanation for why cash renting

³⁶ Alternatively, Young and Burke (2001) explain the custom of common shares using evolutionary focal points as a method of reducing bargaining costs.

³⁷ In the most canonical form, this approach is called the principal-agent model (Garen (1994), Sappington (1991)). Risk-based models of contract choice, however, have developed far beyond the original simple models.

became more common in North America over the past 20 years, arguing that landowners have become more risk averse relative to farmers.³⁸

A long literature has showed evidence *inconsistent* with risk as a factor to explain contracts.³⁹ In an effort to reconcile this, Chiappori and Salanie (2000) and Akerberg and Botticini (2002) have argued that standard regression estimates of contract choice determinants using various covariates, including proxies for risk aversion, may be biased if there is *endogenous matching* or selection between landowners and farmers. If such matching is based on risk preferences of the contracting parties, then it could be the case that more risky situations are governed by high powered cash rent contracts—the opposite of the standard prediction. The intuition is straightforward: higher powered contracts may be necessary to induce stronger effort in cases where there is much variability; however, such contracts may attract only risk-neutral or risk-taking farmers. Risk-averse farmers will opt for low-risk contracts in areas where variability is limited.

A small number of studies have found evidence for matching between firms and workers, including landowners and farmers.⁴⁰ In their historical agricultural application, Akerberg and Botticini (2002) control for matching using geographical location as an instrument and find that this makes their risk proxy variable significant in their contract choice regression. They take this as evidence that risk aversion is an important factor in contract design through a selection effect.

Endogenous matching can help explain why the empirical evidence for the effect of risk preferences is mixed. If all farmers were risk neutral, then contracts would be driven only by transaction cost incentives. In this context, the presence of risk-averse farmers or landowners becomes a *cost* of contracting because it necessarily lowers overall surplus. Risk-averse participants must have lower powered contracts to avoid some risk, and this lowers the total wealth created by distorting incentives.

Consider a situation with a group of landowners potentially contracting with an *even larger* group of potential farmers who also have many alternative forms of employment, and assume the farmers have various risk preferences distributed from risk loving to risk hating. In such a case, landowners will design contracts solely in

³⁸ Within the agricultural economics literature, it is common to find risk aversion as the go-to explanation of various contract features. See Mishra et al. (2020), Hayami and Otsuka (1993), Dubois and Vukina (2004), and Fukunaga and Huffman (2009).

³⁹ See Rao (1971), Hobbs (1997), Allen and Lueck (1995, 1999, 2002), Masten and Saussier (2000), Akerberg and Botticini (2002), Vassalos and Li (2016), and Prendergast (1999, 2000) for examples and summaries of the evidence. Some work finds evidence that risk preferences do matter for contract design (e.g., Fukunaga and Huffman 2009). Outside the area of agriculture, a series of papers have found similar results. See, for instance, Hallagan (1978), Martin (1988), Laffontaine (1992), Leffler and Rucker (1991), Lyon and Hackett (1993), and the summary in Prendergast (2002).

⁴⁰ Prasada and Salmon (2013), for example, examined self-selection based on risk profiles and market power on either side of the market and, within the context of their experiments, found evidence for matching and risk aversion affecting contract design under some market power scenarios.

terms of mitigating transaction costs: some will be high powered, others low powered. Farmers will match with contracts they prefer. Risk-averse “farmers” might take wage jobs in town if none of the farming contracts appeal to them. Risk-neutral or risk-preferring farmers will accept cash rent contracts in risky environments. In the presence of many landowners and farmers, *transaction costs still best explain the contract form*.⁴¹

On the other hand, when there are not enough risk-neutral/loving farmers, the marginal contract will be made with a risk-averse farmer, and risk will become a factor in the contracting terms. In this case, controlling for endogenous matching, evidence for risk in contract design will show up. It is possible that Akerberg and Botticinni’s finding of evidence for risk aversion in the *small villages* of fourteenth-century Italian villages and Allen and Lueck’s (1999) failure to find such evidence in *modern North American agriculture* simply reflect the size of each market and the effect this had on the risk preferences of the marginal contractor.

Beyond the potential for matching over risk preferences, arguments based on risk aversion remain problematic. Measuring exogenous risk and risk preferences are notoriously difficult. This often means risk-based arguments are often ad hoc and hinder the incentive to dig deeper and find incentive-based determinants of contract choice.⁴²

Consider the enormous growth of cash renting in North America over the past quarter century. Huffman and Just (2004) claim this is because landowners became more risk averse compared to farmers over this period. Allen and Borchers (2016), using annual cross-sectional, field-level data from the Agricultural and Resource Management Survey from 1996 to 2011 on nine different crops, show that the rise of no-till conservation farming is the source of the change.⁴³ The finding of Allen and Borchers is consistent with Allen and Lueck (2002) in that cropsharing is used to mitigate soil exploitation problems caused by tillage and other soil manipulation practices, and if farmers adopt methods of planting that *do not* require tillage, then the cash rent contract dominates because it avoids the costs of sharing. Exploiting field-level panel data on crop insurance, Allen and Borchers find no evidence that risk aversion played any role in the switch.⁴⁴

⁴¹ Many farmers and landowners mean the market is “thick,” and there is a sufficient supply of any given risk type, such that risk is never a binding constraint on contract choice.

⁴² Given that (i) landowners are often simultaneously lessors and lessees, (ii) farmers often own their own land, and both parties simultaneously engage in cash rent and cropshare contracts, risk-based arguments can stretch credulity.

⁴³ “No-till” farming is the practice of planting seeds directly into the previous year’s stubble and not tilling or cultivating the soil. It often requires genetically modified seeds and the use of herbicides but leads to a regenerated soil.

⁴⁴ Allen and Borchers use several methods to identify the causal relationship between no-till practices and the adoption of cash rent contracts. The magnitude of their finding is very large and dominates all other explanatory variables. Furthermore, they have data on the presence of crop insurance for every crop at the plot level, and therefore can identify when crop risk is eliminated.

27.6 Conservation and Contract Design

Of all the extensions done with respect to the matter of contract choice, no topic has received more attention than the role played by the form of agricultural contracts in soil conservation. This literature can generally be divided into economic research and environmental science research. Economics research on soil conservation and contracts includes Dubois (2002), Ray (2005), Lichtenberg (2007), and Fukunaga and Huffman (2009). This literature is a response to the transaction cost literature's innovation of recognizing the exploitation of soil as a contracting problem.

In Dubois's (2002) model, cropshare contracts lower the incentive to exploit the land variable, and when the land variable becomes very important, leasing is abandoned for owner operation.⁴⁵ Lichtenberg (2007) uses a multi-task framework that allows for landowners to make decisions that might reduce the moral hazard of land exploitation and notes that choices of rental contracts and conservation practices will be made simultaneously. This often includes not just the choice of cropshare contracts but more tenure security. For example, Deaton et al. (2018) find that even with sharing, farmers with more secure tenure are more likely to engage in conservation practices around tillage and cover crops.

Huffman and Fukunaga (2008) present the economics of contracts to the environmental science literature. It tests both risk aversion and transaction costs factors, including landlord and land variables, and concludes:

... that under diverse economic conditions crop-share contracts are long-run sustainable contracts relative to cash leases. Moreover, they create stronger incentives for long-term sustainability of land quality for agricultural purposes because of the direct involvement of landlords in production and management decisions. (2008, p. 394)

The environmental science literature has adopted the idea that positive transaction costs lead to incomplete contracts, both in terms of a farmer's ability to exploit soil and other land attributes because of high measurement costs but also because contracts are necessarily defined over shorter time horizons than ownership. Since cropshare contracts slow the farmer down in terms of exploiting the assets of the landowner, this literature argues for long-term cropshare contracts to mitigate conservation issues (Kassie et al. 2012; Eder et al. 2021; Zhanga et al. 2018; Stevens 2022).

The environmental science literature related to contract choice tends to take the form of the contract as given. As a result, it is full of policy recommendations to discourage cash renting over cropsharing, impose regulations on contracting parties, subsidize conservation improvements, force longer-term agreements, and more generally discourage short-term leasing, and develop methods of measurement of erosion.⁴⁶ This literature does not acknowledge that contract form and conservation are

⁴⁵This issue of ownership vs. leasing is more generally explored in Chap. 8 of Allen and Lueck (2002).

⁴⁶An irony here is that the agricultural economic policy literature has long discouraged the use of cropshare contracts because of its supposed inefficient use of labor!.

endogenous choices made simultaneously. A cropshare contract is likely chosen in situations where soil conservation is a problem, and the landowner is unable to prevent farmer-induced erosion at a reasonable cost. On the other hand, a cash rent contract may be chosen *because* the soil erosion problem has been solved or reduced. Calling for a blanket reduction in cash renting would therefore be wealth reducing.

Allen and Borchers (2016) address the simultaneous choice between conservation practice in the form of no-till agriculture (which has come to dominate many dryland crops over the past 30 years) and contract choice. This form of tillage almost eliminates soil manipulation and therefore provides fewer opportunities for soil exploitation and damage. The reduced tillage also encourages the regeneration of the soil and reduces the need for fertilizers. All of these factors reduce the incentives for cropshare contracts and explain the rise in cash renting across North America. Farmers and landowners have a joint incentive to mitigate transaction costs, including those related to soil and other asset overuse. As a result, contract design and the distribution of ownership across all assets will reflect this joint wealth maximizing interest. It is possible that in equilibrium, the actual amount of soil erosion or conservation practices would be similar across the different forms of ownership. Indeed, this is what was recently found by Deaton et al. (2018), and a USDA study (Burnett et al. 2022), which concluded:

... we examined conservation practice adoption and behavior across eight separate survey years of the USDA's Agricultural Resource Management survey from 2010 to 2019 ... The trend analyses across these survey years suggest that renters behave almost identically to owners with conservation tillage adoption, as demonstrated by similar soil tillage intensity rating values. (p. 44)

27.7 Conclusion

The transaction cost approach has an empirical focus on real contracts, emphasizing testable hypotheses and recognizing the complexity of assets in the market and within the firm. When this approach is brought to bear in the study of agricultural contracts, it yields robust and empirically supported explanations for the structure of contracts, in both historical and modern agriculture. The transaction cost approach abstracts from risk aversion and risk avoidance and, instead, focuses on pure incentive trade-offs. As we have mentioned, the inclusion of risk sharing adds analytical complications without adding empirical tractability.

We have argued the transaction cost approach explains the basic facts regarding agricultural contracts, and this includes several findings. First, when farmers and landowners contract for land, the contracts are simple in the sense that they are mostly oral and short term. They tend to be enforced through the market via reputation and through the common law via its default rules that simplify the structure of contracts. Second, contract structures are used to police behavior that is difficult to verify by a third party. This costly to observe behavior is present when individuals are not full residual claimants, and this behavior is strongly influenced by

enforcement and measurement costs. For example, the choice between cash rent and cropshare is a trade-off between soil exploitation and crop underreporting incentives.

Third, and perhaps the most important, the classic trade-off between risk and incentives does not explain the choice of contracts or organizations in agriculture, nor elsewhere, as the recent agency literature has shown. Fourth, farming is dominated by family production when the random and systematic effects of nature cannot be controlled. Nature not only provides an opportunity for moral hazard, but it limits the possibilities of specialization. Farm production provides many opportunities for moral hazard, and few for exploiting economies of size. Thus, the dominant labor contract in agriculture is the residual claim farmer.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute

References

- Akerberg, D., & Botticini, M. (2002). Endogenous matching and the empirical determinants of contract form. *Journal of Political Economy*, 110: 564–591.
- Alchian, A. A. (1950). Uncertainty, evolution and economic theory. *Journal of Political Economy*, 58, 211–221.
- Allen, D. W. (1991). What are transaction costs? *Research in Law and Economics*, 14, 1–18.
- Allen, D. W. (2000). Transaction costs. In B. Bouckaert & D. Geest (Eds.), *Encyclopedia of law and economics* (Vol. 1, pp. 893–926). Edward Elgar Press.
- Allen, D., & Barzel, Y. (2023). *Economic analysis of property rights* (3rd ed.). Cambridge University Press.
- Allen, D. W., & Borchers, A. (2016). Conservation practices and the growth of U.S. cash rent leases. *Journal of Agricultural Economics*, 67(2), 491–509.
- Allen, D. W., & Lueck, D. (1992a). Contract choice in modern agriculture: Cash rent vs. Cropshare. *Journal of Law and Economics*, 35, 397–426.
- Allen, D. W., & Lueck, D. (1992b). The Back-forty on a handshake: Specific assets, reputation, and the structure of farmland contracts. *Journal of Law, Economics, and Organization*, 8, 366–377.
- Allen, D. W., & Lueck, D. (1993). Transaction costs and the design of cropshare contracts. *RAND Journal of Economics*, 24, 78–100.
- Allen, D. W., & Lueck, D. (1995). Risk preferences and the economics of contracts. *American Economic Review*, 85, 447–451.
- Allen, D. W., & Lueck, D. (1998). The nature of the farm. *Journal of Law and Economics*, 41, 343–386.
- Allen, D. W., & Lueck, D. (1999). The role of risk in contract choice. *Journal of Law, Economics, and Organization*, 15, 704–736.
- Allen, D. W., & Lueck, D. (2002). *The nature of the farm: Contracts, risk, and organization in agriculture*. The MIT Press.
- Allen, D. W. & Lueck, D. (2005). Agricultural contracts. In C. Ménard & M. M. Shirley (Eds.), *Handbook of new institutional economics*. Springer.
- Allen, D. W., & Lueck, D. (2009). Custom and incentives in contracts. *American Journal of Agricultural Economics*, 91, 880–894.
- Allen, D. W., & Lueck, D. (2019). “The Organization of Vineyards and Wineries” in A. A. Ugaglia et al (eds.), *The Palgrave Handbook of Wine Industry Economics* (London: Palgrave McMillan).

- Anderson, T. L., & Lueck, D. (1992). Land tenure and agricultural productivity on Indian reservations. *Journal of Law and Economics*, 35, 427–454.
- Baker, G. P., & Hubbard, T. N. (2003). Make versus buy in trucking: Asset ownership, job design and information. *American Economic Review*, 93(3), 551–572.
- Bardhan, P., & Singh, N. (1987). On moral hazard and cost sharing under sharecropping. *American Journal of Agricultural Economics*, 69, 382–383.
- Berry, R. L. (1962). Cost sharing as a means of improving the share rent lease. *Journal of Farm Economics*, 44, 796–807.
- Braverman, A., & Stiglitz, J. E. (1982). Sharecropping and the interlinking of agrarian markets. *American Economic Review*, 72, 695–715.
- Braverman, A., & Stiglitz, J. E. (1986). Cost-sharing arrangements under sharecropping: Moral Hazard, incentive flexibility, and risk. *American Journal of Agricultural Economics*, 68, 642–652.
- Burkhart, B. G. (1991). Leases: Farmland lease provisions in Oklahoma. *Oklahoma Law Review*, 44, 461–491.
- Burnett, J., Wesley, D. S., & Scott, C. (2022). Farmland tenancy and conservation practice adoption: are owners and renters any different?. Selected paper prepared for presentation at the 2022 Agricultural & Applied Economics Association Annual Meeting, Anaheim, CA.
- Carmona, J., & Simpson, J. (1999). Rabassa Morta in Catalan viticulture: The rise and decline of a long-term sharecropping contract, 1670s–1920s. *Journal of Economic History*, 59, 290–315.
- Cheung, S. N. S. (1969). *The theory of share tenancy*. University of Chicago Press.
- Chiappori, P. A., & Salanie, B. (August 2000). Testing contract theory: A survey of some recent work. In *Invited lecture, world congress of the econometric society*.
- Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 3, 1–41.
- Crocker, K. J., & Masten, S. E. (1988). Mitigating contractual hazards: Unilateral options and contract length. *RAND Journal of Economics*, 19, 327–343.
- Deaton, J., Lawley, C., & Nadella, D. (2018). Renters, landlords, and farmland stewardship. *Agricultural Economics*, 49(4), 521–531.
- Drache, H. (1964). *The day of the bonanza*. North Dakota Institute for Regional Studies.
- Dubois, P. (2002). Moral Hazard, land fertility and sharecropping in a rural area of the Philippines. *Journal of Development Economics*, 68(1), 35–64.
- Dubois, P., & Tomislav, V. (2004). Grower risk aversion and the cost of moral hazard in livestock production contracts. *American Journal of Agricultural Economics*, 86(3), 835–841.
- DuBow, S. (1999, August). Wheaties: Chasing the ripening harvest across America's Great Plains. *Harper's Magazine* (pp. 33–44).
- Eaton, C., & Shepherd, A. (2001). *Contract farming: Partnerships for growth* (Vol. 145). Food & Agriculture Org.
- Eder, A., Klaus, S., & Eva, S. (2021). Land tenure, soil conservation, and farm performance: An eco-efficiency analysis of Austrian crop farms. *Ecological Economics* 180, 106861.
- Edwards, W., & Boehle, M. (1980). Machinery selection considering timeliness losses. *Transactions of the American Society of Agricultural Engineers*, 23(4), 810–821.
- Edwards, W., & Meyer, V. (1986). *Acquiring farm machinery services: Ownership, custom hire, rental, lease*. Cooperative Extension Service of Iowa State University.
- Ellickson, R. C. (1991). *Order without law*. Harvard University Press.
- Eswaran, M., & Kotwal, A. (1985). A theory of contractual structure in agriculture. *American Economic Review*, 75, 352–367.
- FCC. (2022). Farm credit Canada, 2022 FCC farmland values report.
- Frank, S. D., & Henderson, D. R. (1992). Transaction cost as determinants of vertical coordination in the U.S. food industries. *American Journal of Agricultural Economics*, 74(November), 941–950.
- Fukunaga, K., & Huffman, W. E. (2009). The role of risk and transaction costs in contract design: evidence from farmland lease contracts in U.S. agriculture. *American Journal Agricultural Economics*, 91(1), 237–249.

- Fulton, M. (1995). The future of Canadian agriculture cooperatives: A property rights approach. *American Journal of Agricultural Economics*, 77, 1144–1152.
- Galassi, F. (1992). Tuscans and their farms: The economics of share tenancy in fifteenth century Florence. *Rivista di Storia Economica*, 9, 77–94.
- Garen, J. E. (1994). Executive compensation and principal-agent theory. *Journal of Political Economy*, 102, 1175–1199.
- Gaynor, M., & Gertler, P. J. (1995). Moral Hazard and risk spreading in partnership. *RAND Journal of Economics*, 26, 591–613.
- Goodhue, R., Heien, D., Lee, H., & Sumner, D. (2003). Contracts and quality in the California Winegrape industry. *Review of Industrial Organization*, 23, 267–282.
- Gray, L. C., Stewart, C., Turner, H., Sanders, J. T., & Spillman, W. J. (1924). Farm ownership and tenancy. In *Yearbook of the Department of Agriculture, 1923*. Printing Office.
- Hallagan, W. (1978). Self-selection by contractual choice and the theory of sharecropping. *Bell Journal of Economics*, 9, 344–354.
- Hayami, Y., & Otsuka, K. (1993). *The economics of contract choice: An agrarian perspective*. Oxford University Press.
- Heady, E. O. (1947). Economics of farm leasing systems. *Journal of Farm Economics*, 29, 659–678.
- Hobbs, Jill (1997). Measuring the importance of transaction costs in cattle marketing, *American Journal of Agricultural Economics*, 79(4), 1083–1095.
- Hoffman, P. T. (1984). The economic theory of sharecropping in early modern France. *Journal of Economic History*, 44, 309–319.
- Holmström, B., & Milgrom, P. (1991). Multi-task principal-agent analyses: Incentive contracts, asset ownership, and job design. *Journal of Law, Economics, and Organization*, 7, 24–25.
- Hopkin, J. (1971). Leasing versus buying of machinery. *Journal of American Society of Farm Managers and Rural Appraisers*, MF-2953, 17–23.
- Huffman, W. E., & Just, R. E. (2004). Implications of agency theory for optimal land tenure contracts. *Economic Development and Cultural Change*, 52(3), 617–643.
- Huffman, W. E., & Fukunaga, K. (2008). “Sustainable land use: landlord-tenant contracting in the United States of America,” *NJAS - Wageningen Journal of Life Sciences*. 55(4): 379–396.
- Irwin, G. D., & Smith, L. N. (1972). Machinery leasing: Perspectives and prospects. *Agricultural Finance Review*, 33, 42–47.
- Isern, T. D. (1981). *Custom combining on the Great Plains*. University of Oklahoma Press.
- Johnson, D. G. (1950). Resource allocation under share contracts. *Journal of Political Economy*, 58, 111–123.
- Joskow, P. L. (1987). Contract duration and relationship-specific investments: Empirical evidence from coal markets. *American Economic Review*, 77, 168–185.
- Leffler, L., & Rucker, R. (1991). Transactions costs and the efficient organization of production: A study of timber-harvesting contracts. *Journal of Political Economy*, 99, 1060–1087.
- Kassie, M., Moti, J., Bekele, S., Frank, M., & Geoffrey, M. (2012). Plot and household-level determinants of sustainable agricultural practices in rural tanzania. *Environment for development discussion paper series*, 12(2), 1–41.
- Klein, B., & Leffler, K. B. (1981). The role of market forces in assuring contractual performance. *Journal of Political Economy*, 89, 615–641.
- Klein, B., Crawford, R. G., & Alchian, A. A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21, 297–326.
- Kliebenstein, J. B., & Lawrence, J. D. (1995). Contracting and vertical coordination in the United States pork industry. *American Journal of Agricultural Economics*, 77, 1213–1218.
- Knoeber, C. R. (1989). A real game of chicken: Contracts, tournaments, and the production of broilers. *Journal of Law, Economics, and Organization*, 5, 271–292.
- Knoeber, C. R. (2000). Land and livestock contracting in agriculture: A principal-agent perspective. In B. Bouckaert & G. De Geest (Eds.), *The encyclopedia of law and economics* (Vol. III, pp. 1133–1153). Edward Elgar.

- Knoeber, C. R., & Thurman, W. N. (1994). Testing the theory of tournaments: An empirical analysis of broiler production. *Journal of Labor Economics*, 12, 155–179.
- Laffontaine, F. (1992). Agency theory and franchising: Some empirical results. *RAND Journal of Economics*, 26, 263–283.
- Lagrone, W. F., & Gavett, E. E. (1975). *Interstate custom combining in the Great Plains in 1971* (Economic research service paper no. 563). U.S. Department of Agriculture.
- Lazzarini, S. G., Chaddad, F. R., & Cook, M. L. (2001). Integrating supply chain and network analysis: The study of Netchains. *Journal of Chain and Network Science*, 1:7–22.
- Lichtenberg, E. (2007). Tenants, landlords, and soil conservation. *American Journal of Agricultural Economics* 89(2), 294–307.
- Luporini, A., & Parigi, B. (1996). Multi-task sharecropping contracts: The Italian Mezzadria. *Economica*, 63, 445–457.
- Lyon, T. P., & Hackett, S. C. (1993). Bottlenecks and governance structures: Open-access and long-term contracting in natural gas. *Journal of Law, Economics, and Organization*, 9, 380–398.
- Macaulay, S. (1963). Non-contractual relations in business: A preliminary study. *American Sociological Review*, 28, 55–67.
- Martin, R. (1988). Risk sharing and franchising. *American Economic Review*, 78, 954–968.
- Martinez, S.W. and Reed. “From farmers to consumers: Vertical coordination in the food industry.” USDA Report 720, (1996).
- Masten, S. E., & Saussier, S. (2000). Econometrics of Contracts: An Assessment of Developments in the Empirical Literature on Contracting. *Revue d'Économie Industrielle*, 92(2nd-3rd Trimesters), 215–236.
- Masten, S. E., Meehan, J., & Snyder, E. (1991). The costs of organization. *Journal of Law, Economics, and Organization*, 7, 1–26.
- Ménard, C. (1996). On clusters, hybrids, and other strange forms: The case of the French poultry industry. *Journal of Institutional and Theoretical Economics*, 152, 154–183.
- Ménard, C., & Klein, P. (2004). Organizational issues in the Agrifood sector: Toward a comparative approach. *American Journal of Agricultural Economics*, 86(3), 750–755.
- Mill, John S. ([1871]1965). *Principles of political economy*, Ashley (ed.) (New York, Kelley).
- Mishra, A. K., Anthony, N. R., & Mike, G. T. (2020). Production under input endogeneity and farm-specific risk aversion: evidence from contract farming and Bayesian method. *European Review of Agricultural Economics*, 47(2), 591–618.
- Newberry, D., & Stiglitz, J. (1979). Sharecropping, risk-sharing, and the importance of imperfect information. In J. Roumasset, J.-M. Boussard, & I. Singh (Eds.), *Risk, uncertainty and agricultural development*. University of California Press.
- Nickerson, J. A., & Silverman, B. S. (2002). Why aren't all truck drivers owner-operators? *Journal of Economics and Management Strategy*, 12(1), 91–118.
- Onofri, L., Trestini, S., Mammine, F., & Loughrey, J. (2023). Understanding agricultural land leasing in Ireland: A transaction cost approach. *Agricultural and Food Economics*, 11(17), 1–20.
- Prasad, K., & Timothy c. Salmon. (2013). Self Selection and market power in risk sharing contracts. *Journal of Economic Behavior & Organization*, 90, 71–86.
- Perry, J., Morehart, M., Banker, D., & Johnson, J. (1997). Contracting: A business option for many farmers. *Agricultural Outlook*, May, 2–5.
- Plueger, B. W. (1994). *Farm machinery costs: Own, lease, or custom hire*. Cooperative Extension Service of South Dakota State University.
- Prendergast, C. (1999). The provision of incentives in firms. *Journal of Economic Literature*, 37, 7–63.
- Prendergast, C. (2000). What trade-off of risk and incentives? *American Economic Review*, 90, 421–425.
- Prendergast, C. (2002). The tenuous trade-off between risk and incentives. *Journal of Political Economy*, 110, 1071–1102.
- Rao, C. H. H. (1971). Uncertainty, entrepreneurship, and sharecropping in India. *Journal of Political Economy*, 79, 578–595.

- Ray, T. (2005). Sharecropping, Land Exploitation and Land-Improving Investments. *Japanese Economic Review*, 56(2): 27–143.
- Roy, E. P. (1963). *Contract farming, U.S.A.* Interstate Printers and Publishers.
- Sappington, D. E. M. (1991). Incentives in principal-agent relationships. *Journal of Economic Perspectives*, 5, 45–66.
- Sauvee, L. (2002). Efficiency, effectiveness and the design of network governance. In *5th conference on chain management in agribusiness and the food industry*, The Netherlands.
- Short, C., & Gitu, K. W. (1991). Timeliness costs for machinery selection. *Canadian Journal of Agricultural Economics*, 39, 457–462.
- Statistics Canada. Census of Agriculture, 1976–2011.
- Stevens, A. W (2022). The economics of land tenure and soil health Soil Security 100047
- Stiglitz, J. (1974). Incentives and risk-sharing in sharecropping. *Review of Economic Studies*, 41, 219–255.
- Swinnen, J., Van Herck, K., & Vranken, L. (2016). The diversity of land markets and regulations in Europe, and (some of) its causes. *The Journal of Development Studies*, 52(2), 186–205.
- USDA, (1999, 2014) Census of Agriculture (Washington, DC: USDA).
- Vassalos, M., & Li, Y. (2016). Assessing the impact of fresh vegetable growers risk aversion levels and risk preferences on the probability of adopting marketing contracts: a Bayesian approach. *International Food and Agribusiness Management Review*, 19(1), 25–42.
- Williams, C. M. (1953). Enterprise on the prairies. *Harvard Business Review*, 31, 97–102.
- Williamson, O. E. (1979). Transaction-cost economics: The governance of contractual relations. *Journal of Law and Economics*, 22, 233–261.
- Young, H. P., & Burke, M. A. (2001). Competition and custom in economic contracts: A case study of Illinois agriculture. *American Economic Review*, 91, 559–573.
- Zhang, H., Jianhong E. M., & Bruce A. M. (2018). Adaptation to climate change via adjustment in land leasing: Evidence from dryland wheat farms in the U.S. Pacific Northwest. *Land Use Policy* 79, 424–432

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part VI

Institutional Change

The Economics of Institutional Change: Achievements and Challenges

28

Jean-Philippe Platteau

Contents

28.1	Introduction	702
28.2	The Optimistic Perspective of Postwar Development Economists and Their Successors	703
28.2.1	The Crude Evolutionary View of Many Postwar Development Economists	703
28.2.2	The Induced Institutional Innovation View	706
28.2.3	Caveats and Limitations	707
28.3	The Pessimistic Perspectives of Two Currently Dominant Approaches	708
28.3.1	The Equilibrium-of-the-Game Approach	708
28.3.2	The Political Economy Approach	710
28.3.3	Empirical Works and Caveats	712
28.4	Qualified Optimism: The Gradual Change Approach	714
28.4.1	Institutional Change as a Gradual Process	714
28.4.2	The Role of Quasi-parameters	714
28.4.3	Cascading Effects	715
28.4.4	Institutional Change as an Incremental Process Following a Shock	717
28.4.5	Strengths and Limitations	720
28.5	A Broad Assessment of Achievements and Challenges	721
	References	723

Abstract

How do economists approach the issue of institutional change, and how have their approaches evolved since the Second World War? What are the strengths and limitations of each of them, and how can we define the challenges lying ahead? These are the questions at the center of this chapter, and the attempted answers are based on an assessment that takes into account not only the theoretical soundness of the approaches considered but also their productivity in terms of

J.-P. Platteau (✉)

Department of Economics and Centre for Research in the Economics of Development (CRED),
University of Namur, Namur, Belgium
e-mail: jean-philippe.platteau@unamur.be

© The Author(s) 2025

C. Ménard, M. M. Shirley (eds.), *Handbook of New Institutional Economics*,
https://doi.org/10.1007/978-3-031-50810-3_28

701

the empirical body of knowledge generated. We can see that the field of institutional economics has dramatically evolved during the last 70 years and that the perspectives adopted can be schematically described as having shifted from attitudes of initial optimism to dominant pessimism before ending in qualified optimism. We thus travel through the territories of crude evolutionism and induced institutional innovations, through those of the equilibrium-of-the-game and political economy approaches, to eventually land in the domain of the most recent, so-called gradual change approach. The latter rests on rather sophisticated evolutionary and coevolutionary models. At the end of the journey, the main conclusion is that all the existing approaches have their own merits and application domains so that they should be seen as essentially complementary and an eclectic attitude towards them is fully warranted.

Keywords

Institutional change · Evolutionary change · Gradual change · Induced institutional innovation · Political economics · Game-theoretic approach · Natural historical experiments

28.1 Introduction

Looking back at the way the subdiscipline of institutional economics has broadly evolved since the Second World War, it is possible to distinguish between three main periods depending on the degree of optimism of the perspective adopted for the analysis of institutional change. Schematically, there would then be periods of optimism, pessimism, and qualified optimism. In the following, we propose to briefly characterize these three major periods and the associated schools of thought, and at each step, we stress their strengths and weaknesses. This is done in Sects. 28.2, 28.3, and 28.4 where the historical order of appearance of the different theories is respected. We will then be ready to assess the challenges that lie ahead, a task earmarked for Sect. 28.5, the final section of the paper.

The above undertaking will make us travel across the territories of crude evolutionism and induced institutional innovations, through those of the equilibrium-of-the-game and political economy approaches, to eventually land in the domain of the most recent, so-called gradual change approach. The latter rests on rather sophisticated evolutionary and coevolutionary models. At the end of the journey, the main conclusion is that all the existing approaches have their own merits and application domains so that they should be seen as essentially complementary and an eclectic attitude towards them is fully warranted.

It bears emphasis that the perspective adopted in this paper is primarily conceptual, in the sense that its central purpose is to scrutinize the main approaches to institutional change (rather than institutions) available, seek the main insights they have delivered or are able to deliver, look rather summarily at their empirical contributions, and highlight their limitations. However, owing to a lack of space,

we bypass a thorough discussion about the precise definitions and meanings of key concepts, such as institutions, formal and informal institutions, and culture.

28.2 The Optimistic Perspective of Postwar Development Economists and Their Successors

28.2.1 The Crude Evolutionary View of Many Postwar Development Economists

Many early development economists implicitly held a sort of rough and optimistic (co-) evolutionary view of institutional and cultural change. Thus, Arthur Lewis (1955) thought that religious beliefs, for example, may evolve and be reinterpreted depending on the economic environment confronting societies. In other words, traditional values and attitudes, whenever they are hostile to economic advancement, will eventually adapt themselves to new economic opportunities (p. 106). As for Alfred Hirshman (1958), he pointed out that traditional images of change will remain a critical bottleneck for constructive development actions, yet their obstructive effects will only last until experience modifies them in the appropriate direction (see also Bauer & Yamey, 1957; Meier & Baldwin, 1957). Combined with the idea of a Big Push deemed necessary to extract poor countries from low-level equilibrium traps (Leibenstein, 1957; Bruton, 1965; Higgins, 1968; Myrdal, 1968), the dominant postwar vision suggests a conditional scenario: if a shock of sufficient magnitude can be produced on the economic and technical fronts (through massive aid efforts by advanced countries, in particular), forces operating at social, cultural and political levels will somehow adjust so that the beneficial effects of the shock can materialize, at least in the medium term.

A possible theoretical underpinning of such an adaptive view is a sort of crude evolutionary mechanism: the emergence, diffusion, and demise of rules or institutions would be the outcome of an organic process of Darwinian natural selection epitomizing the competitive pressures of the market and the invisible hand. Institutions or rules are then seen as evolving unconsciously and gradually as a result of the pursuit of individual interests as agents repeatedly face the same types of social problems or situations. In the simplest version, inefficient institutions are expected to have a low evolutionary fitness, and therefore they tend to be displaced in the long run by more efficient institutions (see, e.g., Schotter, 1981; Axelrod, 1984; Sugden, 1986, 1989). Looked at in this way, institutions emerge not as a result of rational, purposeful design by any individual or organization of individuals, but as the result of spontaneous evolution (Hayek, 1960, 1973).

The above-described Darwinian process is grounded in the assumption of limited or bounded rationality: individuals follow trial-and-error behaviors (what works well for an individual is more likely to be used again, whereas what turns out poorly is more likely to be discarded), they imitate other people if successful, and they may even learn from experience that following a given constraint or custom can actually serve their own individual interests. In short, they take their decisions on the basis of fragmentary

information, and this is translated into the principle that the probability of taking an action in the present increases with the payoff that resulted from taking that action in the past (Axelrod, 1997, pp. 47–48; Young, 1998, p. 28; Aoki, 2001, p. 40).

Before turning to the next optimistic approach, we want to devote some space to an institutional reinterpretation of Lewis' seminal paper of 1954. In that work, the author proposes a highly optimistic view of structural transformation and economic development whereby informal institutions and rules, distinct from cultural traits, are bound to be replaced by formal ones.¹ To be more precise, Lewis conceptualizes economic growth as a process the first phase of which involves continuous transfers of labor from a traditional informal sector (called agriculture) to a modern urban-based sector (say, the industrial sector). At the start of the process is a “kernel of capitalists” (or, if it does not exist, the national state or a group of foreign investors) eager to accumulate capital and develop the modern sector. Once they operate and plough back their profits, growth is sparked and maintained. As long as an unlimited supply of labor exists in the traditional sector, which is ready to be tapped by the modern capitalist sector at a constant (reservation) wage in the way first suggested by Nurkse (1953), capital-widening growth proceeds with the properties predicted by the golden accumulation rule.

The important point from the institutional standpoint is that rural-urban migration (or, equivalently, traditional-modern or informal-formal migration) articulates together two sectors endowed with different institutions, themselves reflected in different remuneration rules. The rules prevailing in the modern sector consist of the institutional prerequisites of market-based economic growth, and they are exogenously given: production factors are separately owned (so that wage employment is possible, and capital can receive a separate remuneration that constitutes an incentive to accumulate), contracts are enforced, and property rights are well established. Given that competition exists inside this sector, production factors are remunerated at their marginal productivity. By contrast, the customary way of sharing effort and income is embedded in the rules governing the traditional sector. Implicitly, they involve a community acting as an authority in charge of devising and implementing the custom. Finally, there exist rules operating at the intersection between the two sectors, and their function is to make economic transactions and trade relationships possible between them. Like the rules at play inside the traditional sector, these rules are endogenous to the expansion of the modern sector. They are market rules, each sector selling its surplus to the other at a price that represents the terms of trade between the two sectors and adjusts over time.

Rules governing the traditional sector are bound to gradually disappear and leave room for the “modern” capitalist rules, causing the enforcement authorities in this sector to progressively lose their influence and authority. There is thus a (largely exogenous) engine of change, namely the expansion of the modern sector, and institutional change is a process of induced and gradual transformation, or

¹This is in contrast to the current practice among economists of amalgamating cultural traits and informal institutions (see Alesina & Giuliano, 2015).

absorption, of customary rules into modern rules. Yet, during a transition period, they will subsist and coexist with modern rules, thus creating a situation of institutional dualism, which entails efficiency costs. Inside the traditional sector, indeed, labor is paid above its marginal productivity, which is zero or close to zero. The institutional transformation occurs when the so-called “turning point” is reached, implying that labor has become scarce (its supply is “limited”) throughout the economy: as a result of outmigration the marginal productivity of labor in the traditional sector has risen up to the reservation wage level. Wages are then equal to the marginal productivity of labor (the capitalistic or neo-classical rule), which has become identical across the two sectors. Whether in the modern or the traditional sector, growth follows the path of a capital-deepening process reflected in rapidly rising labor productivity and wages. Not only is the market integrated over the whole national territory, but the society is now institutionally unified thanks to the establishment of a single rule for input remuneration.

Combined with the adaptive view of social customs and values, and the associated individual attitudes, the Lewisian theory of overall development depicts a process of gradual transformation anchored in a well-described economic mechanism but poorly elucidated evolutionary processes. The main upshot of such a transformation is clear: all informal arrangements and the underpinning values and attitudes will ultimately and irretrievably give way to modern institutions and cultural features (individualism, working habits, innovative spirit, and the like). The rules governing the economy will be homogenized once a tipping point is reached, whereas cultural practices and attitudes will take some more time to get adjusted to the new economic institutions that have emerged.

It is noteworthy that the mode of interaction between the formal and informal sectors and their respective institutions in the Lewisian setup does not quite fit into the well-known typology proposed by Helmke and Levitsky (2004), two sociologists. The difficulty arises from the static characterization implicit in their typology, where four types of formal-informal interactions—complementarity, substitution, accommodation, and competition—are conceived as though they are stable and self-sustained states. In addition, the typology is geared toward explaining the persistent role of informal institutions in the face of formal ones, thus making it impossible to account for the demise of informal rules as envisaged by Lewis. Reflecting the static approach is the implicit assumption that, unlike formal institutions, informal arrangements are always effective in the sense that they are well enforced.² In contrast, in the Lewisian

²Helme and Levitsky’s scheme is based on pairwise relationships between two binary variables: convergence or divergence in the outcomes sought by formal and informal institutions, and strength or weakness of their respective enforcement capacities. Interactions are complementary when effective formal institutions coexist with (effective) informal institutions, and both pursue convergent objectives. Substitution between the two types is observed when they have convergent purposes but formal institutions have weak enforcement capacity. For the sake of completeness, accommodation and competition obtain when the two types of institutions have divergent objectives, the former in the case formal institutions are effective and the latter in the opposite case (For a presentation and discussion of the typology, see Baland et al., 2020).

process of dualistic development that unfolds during the phase of unlimited labor supply, the formal and informal arrangements appear to be complementary while, beyond the “turning point,” the latter are replaced by the former so that (complete) substitution becomes the relevant interaction type.

28.2.2 The Induced Institutional Innovation View

Equally optimistic, yet squarely rooted in a rationality framework, is the induced institutional innovation (III) view that emerged in the 1980s (Hayami & Ruttan, 1985; Hayami & Kikuchi, 1981). The central idea is that changes in the environment—in particular, long-term changes in the relative scarcity of production factors as a result of changes in market demand, demography, or technology—are translated in the price realm and rational agents respond to changes in relative prices by modifying not only the technology they use but also the institutions that regulate their lives.³ It is noteworthy that the extension from technical to institutional innovations went hand in hand with an enlargement of the cost-benefit apparatus so as to take into account all sorts of transaction costs as well as the benefits of internalizing externalities.

Allen and Lueck (2002) explicitly invoke the forces of competition (competition among farmers for land, among landowners for renters, and between on—and off-farm opportunities generally) to explain why the most valuable contract or arrangement, in the sense of maximizing the expected value of the joint wealth of all the parties concerned, net of transaction costs, is eventually chosen by all the parties. Indeed, it is because competition is strong enough that, in their own words, “it seems only reasonable to assume that contracts and organizations are fundamentally driven to maximize wealth” (p. 6).

The transaction-costs approach, often equated to the New Institutional Economics, thus appears to belong to the III approach with variations in transaction costs being among the possible parametric changes susceptible of inducing institutional innovations. When such costs cannot be made reasonably explicit, however, the modern theory of contracts suggests an alternative way of defining optimal institutional arrangements. Grounded in the assumption of strategic rationality—an even more demanding assumption than parametric rationality—it is based on the Principal-Agent model, which posits the presence of an information gap that the agent can fully exploit for their own benefit if they are not constrained through appropriate contract terms set by the principal. When a change occurs in some parameter of the model, say as a result of the emergence of new outside opportunities accompanying market development, population growth, or technological innovations (or a fall in transaction costs), the optimal contract or arrangement is modified, perhaps in a quasi-continuous manner if the change in parameters occurs with high frequency.

³ When applied to technological choices, the theory is known as the Induced Technological Innovation hypothesis, initially propounded by John Hicks (1932). See Hayami (1997, pp. 16–19), for a short presentation.

Recognizing this modification, rational actors are somehow able to bring about the new optimal arrangement.

Together, the above two lines of approach have proven quite rich in the sense of providing plausible explanations behind a wide array of institutional changes, contract forms, and micro-institutional arrangements in particular. These range from the individualization of land property rights, and the emergence of family farms and nuclear households to the disappearance of serfdom, the growing incidence of late marriage, and the breaking of kinship structures in Western Europe through the shift from matrilineal to patrilineal rules of inheritance and from relation-based to rule-based governance (Binswanger & Rosenzweig, 1986; Greif, 1989, 2006; Baland & Platteau, 1996; Platteau, 2000; La Ferrara, 2007; Voigtländer & Voth, 2013; Guirkinger & Platteau, 2015, 2020; Li, 2003; Schulz, 2017).

28.2.3 Caveats and Limitations

Both the crude evolutionary and the III approaches present an obvious drawback: they fail to account for the possible persistence of inefficient institutions in the face of changes in the economic, demographic, and technological environment. Regarding the first approach, it is worth noting that, as will come out of evolutionary models developed later, and contrary to what immediate intuition suggests, these models do not unambiguously point to efficient adaptation of human institutions. Evolutionary equilibria have the properties of strategic equilibria: for a large class of evolutionary games, if the dynamics converge, they converge towards a steady state in which the limiting distributions are in equilibrium in the same sense as in classical game theory. In other words, even though behavior of the players is not rational, the population seems to learn the rational equilibrium as its distribution evolves, and nothing guarantees that this equilibrium is efficient (Montet & Serra, 2003, pp. 8–9).

As for the second strand of literature, its emphasis on the demand for institutional change eschews the problem of its supply. As Hayami later admitted (1997), the III model is quite naïve, especially because institutional change typically requires collective action which cannot be assumed to be automatically forthcoming whenever the aggregate social benefit resulting from such a change exceeds the social cost. Its pertinence is therefore limited to situations in which the individual responsible for introducing an institutional innovation is able to capture all the resulting benefits. In such cases, “institutional change is like producing a private good which can be carried out by one or more individuals without the need for collective action,” so that changes in institutions understood as contractual forms may only require “the leadership of one party to the contract (or consensus among the various parties) to bring them about” (Lin & Nugent, 1995, pp. 2326, 2348). The approach is also pertinent when (a limited number of) agents are in symmetrical positions so that their interests are identical. On the contrary, it is inappropriate to analyze situations in which numerous actors interact simultaneously, say to produce public goods.

28.3 The Pessimistic Perspectives of Two Currently Dominant Approaches

Presently dominant are two economic theories that do not yield optimistic conclusions of the sort attained by the two approaches discussed in Sect. 28.2. The two new strands of literature are the equilibrium-of-the-game (or the self-enforcing equilibrium) approach and the political economy approach. Below, we discuss them in turn.

28.3.1 The Equilibrium-of-the-Game Approach

The approach we are concerned with here has been dominant in the economic profession during the last decades. Stressing the conventional nature of institutions, it conceives of them as self-enforcing (equilibrium) arrangements. Its success partly stems from the fact that it lends itself to rigorous modelling on the basis of classical game theory and, relatedly, it enables economists to propose a theory of institutions whose realm includes social norms and informal coordination mechanisms.

Conventions are defined as “patterns of behavior that are customary, expected and self-enforcing: everyone conforms, everyone expects others to conform, and everyone has good reason to conform because conforming is in each person’s best interest when everyone else plans to conform” (Young, 1996, p. 105, as inspired by Lewis, 1969). On the basis of this understanding, institutions are conceptualized as Nash equilibria on the ground that, since beliefs have converged, actions are coordinated between individuals. The self-enforcing property of institutions, which is the cause of their inherent stability, makes them appear as driven by cultural inertia (Basu et al., 1987). It arises from the fact that at equilibrium the actors’ expectations about each other’s behavior in a particular situation turn out to be consistent with the experience generated by the resulting actions. Agents have then no incentive to deviate from their own action-choice rule and institutions represent stable outcomes (Aoki, 2001, pp. 10–11). As succinctly put by Aoki (2001), an institution is “*a self-sustaining system of shared beliefs*” about a salient way in which a game is repeatedly played. This way of playing can be viewed as the rules of the game that are “endogenously created through the strategic interactions of agents, held in the minds of agents, and thus self-sustaining” (p. 10).

A number of important implications follow from the strict definition of institutions as Nash equilibria in repeated games. First, there are typically multiple possible institutional arrangements that meet the requirements of a Nash equilibrium: they correspond to all those which meet the two following requirements: the actors’ beliefs converge, and their actions are coordinated. Second, there is absolutely no assurance that their outcomes are all socially efficient.⁴ Because actors get stuck to

⁴It is only in particular situations, namely in pure coordination problems, that people are indifferent between possible stable solutions: the only thing that matters to them is that everybody follows the same rule or convention.

the outcome that their expectations have led them to, they may well end up coordinating on a rule that is not socially efficient. Following the same logic, nothing precludes the emergence of paradoxical situations in which individuals choose to support rules or other institutional arrangements which they dislike or even find repugnant (such as the caste system in India). As shown in the pioneering work of Akerlof (1976), individuals may then comply with unpleasant rules (e.g., female excision rules, or customs forbidding the remarriage of widows), or obey a totalitarian regime, if there exist an effective network of mutually reinforcing social sanctions against disobedience and a system of converging pessimistic expectations that sustain the existing arrangement. The intuition is powerful: a harmful rule may persist because the actors mutually suspect each other of supporting it, and a web of self-reinforcing sanctions makes them wary of violating it. In such instances, everyone is both a victim and a supporter of a possibly oppressive rule or arrangement and a power-wielding central authority needs not even exist to enforce it. It is noteworthy that, for this sort of effective sanctions to prevail, meta-punishment must be applied, that is, a person is considered disloyal to a regime or a rule if either s/he does not cooperate or s/he maintains relations with someone who is disloyal (Kuran, 1995, pp. 118–136; Basu, 2000, pp. 136–147).

Clearly, the self-enforcing approach with its emphasis on persisting institutions, is generally pessimistic insofar as change is precluded in the absence of exogenous shocks, and even harmful norms and practices can be sustained over time. For changes to happen, they typically need to be sparked by exogenous shocks that modify some key parameters of the situation produced by their history and culture. Since a self-enforcing institution is one in which each player's behavior is a best response, the inescapable conclusion is that a change in self-enforcing institutions must have an exogenous origin (Greif & Laitin, 2004, p. 633). Shocks may be caused by accidental circumstances (say, a change of political regime in a neighboring country, a military defeat, exposition to a new foreign ideology, religion, foreign markets, or new technologies), or by deliberate attempts of the government to bring radical changes in the economic, social or political system (for instance, by opening the country to foreign trade). In game-theoretic terms, exogenous shocks are translated into a modification of the payoffs accruing to at least some actors in given interactions with others, the entry of new actors or the withdrawal of past players, an enlargement or a narrowing of the repertoire of available action-choices, a revision of individual expectations regarding others' actions, and/or a change in the informational characteristics of the game.

The abusive use of the word "rules" when referring to conventions needs to be stressed. As it is plainly evident from the foregoing presentation, a convention is an outcome of individual strategies, which, in a repeated interaction framework, happen to converge to an equilibrium. It follows that such an equilibrium could be disrupted even in the absence of an exogenous shock, say because some agents make mistakes and inadvertently deviate from their established pattern of behavior, or erroneously punish other players, or because their expectations have suddenly changed although the game environment has not. There is no way we may therefore conceive of a convention as a rule or an institution in the original sense defined by North (1990): a humanly elaborated device that constrains individual behavior.

28.3.2 The Political Economy Approach

To understand radical changes in the policies of a government, we obviously need to consider the state as a full-fledged actor with its own preferences and payoffs. This is the task addressed by the new political economy theory. An extension of the reasoning behind the Coase theorem to the political sphere means that political and economic transactions ought to push towards efficient policies and institutions (i.e., those that achieve the best outcomes for the society), regardless of who, or which social group, holds political power (Acemoglu, 2003, p. 621). In reality, however, we do not often observe politicians and powerful social groups making deals with the rest of society for the purpose of adopting policies and establishing institutions that maximize aggregate production or social welfare while ensuring the redistribution of part of the gains to themselves. In short, they do not “predate efficiently.” At the root of this recurrent state failure or political failure are “inherent commitment problems associated with political power” (Acemoglu, 2003, p. 622; see also Besley & Coate, 1998). More precisely, while the validity of the so-called political Coase theorem presupposes the ability of the state to write enforceable contracts, the state, or the social groups controlling it, cannot actually commit to abstaining from using their power to renege on their promises, or to refrain from modifying the terms of the contract, if their selfish interests dictate that they should do so.

Because contracts written by the state are by definition unenforceable—there is no state above the state—a serious commitment problem arises that prevents efficient outcomes from materializing. On the other hand, if a ruler relinquishes their power, the citizens cannot commit to making side payments in the future since this ruler will no longer possess the political power to enforce the agreed-on promises. This is the political form of the so-called double commitment problem, which makes credible commitment failures ultimately appear as problems about the future exercise of political power. A general conclusion of the political economy literature is that the degree of inequality in the distribution of the costs of reforms, or of the benefits from the status quo, plays a critical role in determining whether reforms will be feasible. More specifically, when the inequality of power, or the asymmetry of the benefits from existing institutions (or status quo policies) are quite large, elites are prompted to oppose reforms in order to be able to hold on to their power and the associated rents in the future. Since the initial inequality in wealth and power is high, it is then difficult to commit in future to making everyone better off without making the elites worse off (Khemani, 2020, p. 5). Only through revolution or through a change in the composition of the elites and their preferences can key reforms become possible.

Other reasons than unequal initial endowments can explain political opposition to reforms, however. In particular, uncertainty may exist over the distribution of the costs and benefits from reforms, due to a lack of solid technical evidence, to heterogeneous preferences leading to different evaluations of the pro and contra of a change, or to ideological resistance to available evidence. Regarding the latter source of uncertainty, the propagation of ideological prejudices against a reform may

actually be driven by powerful groups whose interests would be harmed by its implementation. An additional bias towards the political status quo arises when some individual winners and losers from the reform cannot be identified beforehand.

Finally, an important strand of the political economy literature looks into the question as to how rulers manage to stay in power by using seduction and co-option of elites, or direct transfers to people or subgroups of people. When the regime is autocratic, coercion is an additional instrument in the hands of the ruler. Co-optive tactics, direct transfers, and repression can thus be deemed by rulers to be more convenient ways to maintain themselves in power than genuine bargaining with the population. Inasmuch as maximization of growth or other components of development is not the primary objective of the ruler, tactics of political survival are expected to cause costs and inefficiencies, both static and dynamic. Thinking in political economy terms then means that the analysis is centered on trade-offs between political and economic objectives, such as is explicitly done, for example, in Acemoglu and Robinson (2006). As for Auriol and Platteau (2017a, b), and Auriol, Platteau and Verdier (2023), they explicitly model the trade-off between political stability and the intensity of reforms when strong groups, possibly backed by religious or tribal leaders, are antagonistic to them.

Hard trade-offs between economics and politics do not stem only from the nagging lust of ruling elites for power and its associated economic privileges. They may also arise from conflicts between objectives of economic growth and social stability, such as those that must be typically resolved in any process of nation-building. This problem is especially relevant for societies fraught with horizontal cleavages along religious, ethnic, or tribal lines. Thus, African societies tend to be highly segregated in the sense that their population is homogeneous within regions and different between regions, thus predisposing them to develop regionalist identities. Moreover, artificial borders drawn by colonial powers divided the territories of many ethnic groups into different states. As a result, these countries face a serious risk of instability, separatist movements, and civil conflicts (see Alesina et al., 2011; Rohner & Zuravskaya, 2023, p. 4). The same predicament is observed in many parts of Asia: think of the north-eastern states of India, the Western Frontier Province and the Baluchistan in Pakistan, the northern region in Myanmar, the Chittagong Hill tracts in Bangladesh, the south-eastern and northern parts of Uzbekistan (Fergana Valley and Karakalpakistan, respectively), and so on. For national authorities, the challenge therefore consists of confronting “a dense assemblage of cultural, racial, local, and linguistic categories of self-identification and social loyalty that centuries of uninstructed history had produced,” and then finding an acceptable compromise between the need to modernize the social, political, and economic structures of the country, on the one hand, and the need to win a mass-based legitimacy and to involve traditional interests that play upon primordial attachments, on the other hand (Geertz, 1973, pp. 239, 277; see also Iliffe, 2007).

Efficiency losses resulting from the pursuit of social cohesion and national integration may take on different forms: incentive distortions caused by policies of positive discrimination aimed at a better integration of marginal social groups; lack of legal integration and rule uniformization accepted in order to reduce tensions

between central authority and local informal loci of power; and costly co-option of top army officers or religious leaders intended to prevent them from destabilizing the country.

28.3.3 Empirical Works and Caveats

The equilibrium-of-the-game approach has stimulated a proliferating body of empirical knowledge that exploits the occurrence of (past) exogenous shocks. The focus then tends to be on the long-term effects on current outcomes generated by institutions that emerged in the distant past. The underlying idea is that such outcomes are likely to reflect the influence of critical institutions and their persistence to this day. The methodology, which gives primary importance to the issue of causality, requires the researchers to uncover and analyze historical natural experiments, defined as situations where institutions are varied in geographical units that possess common environmental, cultural, and all sorts of socio-economic conditions susceptible of affecting outcomes.

An oft-made claim is that, although the mechanisms behind institutional (or cultural) persistence are typically speculative, the new economics of history is strong on identification so that causality is well established. Variation in some key determinants, whose presence or absence was observed a long time ago, is considered as exogenous, and is therefore used as a valid instrument to explain today's different outcomes. Since many geographical characteristics have the appearance of being deterministic, they are often seen as the deep causes behind the institutional divergence deemed to be responsible for different growth and development trajectories (for a survey of this literature, see Cantoni & Yuchtman, 2021; Voth, 2021; and for a well-known illustration, see Engerman & Sokoloff, 1997).

Upon careful look, however, things do not appear so clear. The assumption that the distinct institutions considered in the natural experiment arose in otherwise common cultural, geographical, and environments (as typically claimed in Spatial Regression Discontinuity Designs) is far from evident. Equally disputable are all attempts to identify culture as the cause of prosperity, by isolating cultural variation in historical environments with a common institutional setup. In point of fact, it is generally hard to disentangle institutional features of a society from its long-term, intergenerationally transmitted cultural traits. Regarding the operating mechanisms, it seems important to bear in mind Dixit's warning (2007) that "the notion that geographic and historical variables are merely instruments for institutional determinants of economic success is supported more by the intuitive appeal of the stories told than by the statistical significance of the tests performed" (p. 137). As underlined by Rodrik (2004) and others, it is not because a researcher finds an appropriate econometric instrument that they are able to offer an adequate explanation.

Keeping these cautionary remarks in mind, some rich and challenging insights have been provided by empirical research based on historical natural experiments. Several of them are worth singling out. Before taking that step, however, we need to

precise that, according to us, the concept of culture ought not to be amalgamed with the concept of institutions. We define institutions in the sense of North (1990), that is, as man-made rules and procedures that constrain individual behavior in such a way that human actions become coordinated in a rather stable manner. This implies that, whether formal or informal, they are effectively enforced. Culture, by contrast, is as essentially made of values, preferences, and norms or codes of behavior (e.g., a dress code). These elements are typically transmitted across generations through various socialization practices, or they are acquired through interactions between peers. Such understanding is at odds with the common definition of culture, which encompasses both values (and preferences) and beliefs, as found in Alesina and Giuliano (2015) for example. As we conceive of them, beliefs are synonymous with informal institutions when they refer to expectations about, or interactions with, others whereas they are akin to values or preferences when this is not the case. In particular, trust, civic-mindedness, legitimacy, cooperation spirit, and other forms of social capital which might fall under the heading of “belief,” belong to the realm of culture and are therefore not informal institutions.

This being clarified, we can now list the most valuable insights gained from the empirics born of the self-enforcing approach. First, various determinants of development interact, hence the same institutional change may have differential effects depending on different cultural environments (see, e.g., Putnam, 1993; Padro-i-Miguel et al., 2014). Second, culture may play a more important role than institutions in determining development outcomes (e.g., the role of ethnicity in Sub-Saharan Africa as found in Michalopoulos & Papaioannou, 2014). Third, when an important role is demonstrated for geography and biogeography in the onset and diffusion of economic development over the past millennia, careful examination shows that the effects are indirect rather than direct, and they are plausibly channeled through institutions (see the discussion in Spolaore & Wacziarg, 2013, pp. 327–336).

Fourth, political institutions seem to be a key factor for development. This is attested by evidence from in Sub-Saharan Africa that, compared to societies lacking political hierarchy beyond the local level, those possessing a historical legacy of political centralization tend to display better (contemporary) economic performance (Michalopoulos & Papaioannou, 2020, pp. 106–114). Finally, institutional change, and even cultural change, can sometimes be rather rapid. It can even be doubted that cultural attitudes always change (much) more slowly than institutions, as witnessed by the impact of international migration to more advanced countries (Docquier et al., 2016; Barsbai et al., 2017).

Turning to the political economy approach, which essentially rests on theoretical constructs, empirical support is generally limited to (historical) narratives. First, they are rather ad hoc, especially when they are short and superficial in the sense that they do not give a proper account of the contextual elements of the institutional situations at stake. Second, they may be plausible yet offer only one explanation among others. A useful lesson from the empirical literature born of the EG approach is that political narratives may take advantage of a shock (say, a war or a natural disaster) whose effects predicted by the theory are then tested against the observed changes (see, e.g.,

Auriol et al., 2023). This is admittedly no rigorous test in the sense required today by the economics profession in matters of quantitative empirical research. However, dealing with political economy issues is hard and sound narratives may be instructive to the extent that they demonstrate the possibility of imparting structure on complex realities and thereby facilitate their understanding. Like political scientists and historians, economists would then resort to narratives instead of rigorous empirical tests, yet these narratives would be grounded in analytically deduced propositions acting as guides to the available evidence.

28.4 Qualified Optimism: The Gradual Change Approach

28.4.1 Institutional Change as a Gradual Process

Heralded by sociologists, the gradual change approach starts from the idea that change does not proceed abruptly but gradually and in a hardly perceptible manner. The piecemeal changes are born of endogenous developments that often unfold incrementally and only “show up” or “register” as change if a somewhat long-time frame is considered (Mahoney & Thelen, 2010, p. 2; see also Portes, 2006, p. 251). In this perspective, the sources of institutional change are not simply exogenous shocks or shifts in the game environment, and path-dependent lock-in effects are not considered to be a common occurrence in real societies. Understandably, causal relationships cannot be easily established, yet it would be wrong to sideline the gradual change approach on that ground alone. The analytical insights it provides help shed light on important aspects of the development process. In particular, gradual and imperceptible change can be seen as preparing the ground for an institutional breakthrough, and/or representing the ripple effects of an exogenous shock.

The gradual approach is best described as evincing qualified optimism. It is not altogether pessimistic insofar as its focus is on change processes and not on equilibrium traps of institutional status quo. And it is not squarely optimistic insofar as it does not show that institutions necessarily and ultimately move in a desirable direction as in the crude evolutionary approach. Yet, gradual change theorists strive to explore pathways that can potentially lead to a socially efficient institution (or institutional configuration) following a shock to an existing and inefficient setup, and they highlight the conditions under which a positive transformation can be obtained.

Economists have addressed the issue of gradual change through varied approaches. In contrast to the pessimistic approaches in which institutional change is the outcome of an exogenous shock, all of them are attempts to endogenize this change. Three of them retain our attention here.

28.4.2 The Role of Quasi-parameters

Greif and Laitin (2004) have proposed a theory based on the idea that “surely all institutions endogenously unleash processes of stability and change simultaneously”

(p. 636). Their theory incorporates two new concepts: quasi-parameters and institutional reinforcement. The basic intuition is that some aspects of a situation should be considered as parametric in studying self-enforceability but as endogenously determined, and therefore variable in the long run, in studying institutional dynamics (pp. 633–634). The authors call “quasi-parameters” these parameters of the model which support the observed equilibrium, but whose values are themselves affected by the self-enforcing outcomes in a manner potentially leading to long-term behavioral change.

Given that a particular equilibrium can generically be sustained over a broad range of parameters, two possibilities arise. A (self-enforcing) institution is “reinforcing” if “the behavior and processes it entails, through their impact on quasi-parameters, increase the range of parameter values (and thus “situations”) in which the institution is self-enforcing” (Greif & Laitin, 2004, p. 634). On the contrary, an institution is “self-undermining” if this range is gradually narrowed down, implying that the entailed changes in the quasi-parameters make the associated behavior self-enforcing in a smaller set of situations. A direct consequence is that the endogenous institutional change will only occur when the self-undermining process proceeds continuously so as to reach a critical level beyond which past patterns of behavior are no longer self-enforcing. Another consequence is that institutional change resembles a punctuated equilibrium: change is evolutionary but appears to be abrupt, typically associated with a “crisis” (p. 639).

It is noteworthy that to this date and to our knowledge, there is a shortage of empirical works explicitly using the above framework for the purpose of understanding institutional change or stagnation. This is less true of the next approach considered below.

28.4.3 Cascading Effects

Like Greif and Laitin’s approach, the one we are concerned with in this subsection is rooted in classical game theory. Another similarity is that it postulates the existence of a threshold beyond which institutional change comes into the open. The underpinning analytical structure lies in cascade-based models where gradual and radical transformations coexist. In these models, parameters can change gradually and over a long period without upsetting the equilibrium. Once a tipping point is reached, however, a mass transformation in behaviors occurs, so that gradual change and revolutionary change appear as both belonging to a single process of transformation. While radical change is only made evident “by the moment of social explosion itself” (Portes, 2006, p. 251), it has been long prepared by a series of gradual and imperceptible steps. These steps take the form of a cascade, or a bandwagon effect, in which successive behavior adjustments are responses to the information obtained by looking at how others behave.

If we accept the contribution of Arthur Lewis as reinterpreted in Sect. 28.2, a pioneer attempt along the aforementioned line is owed to Kuran (1987, 1988, 1995).

His theory of so-called “collective conservatism” is based on the observation that because of group pressures, the policy preferences people express in public (“public lies”) often differ from those they hold privately (“private truths”). Under many realistic conditions, the possibility of such “preference falsification” yields multiple equilibrium distributions of publicly exposed individual preferences, each associated with a different policy, and each attainable, given the right initial expectations, through a bandwagon process. To obtain this result, it must be the case that the distribution of publicly exposed preferences in one period affects the actors’ expected payoffs in the next. They then rely on the observation of others’ behaviors to decide about their own. Their preferences are thus sensitive to the expected size of the publicly manifested support or opposition to an institutional reform, implying that publicly expressed preferences are interdependent.

To make the matter clearer, assume that all individuals dislike a harmful institution (say, the dowry practice, or female genital mutilation) but they differ in the intensity of their repugnance. If their willingness to express their dislike publicly depends on both this intensity and the proportion of people who have already openly expressed their opposition, it is possible to have a status quo equilibrium in which there are simply not enough outspoken opponents to spark an institutional change. What is then obtained is a situation of the type referred to by Akerlof (see above, Sect. 28.3). In some settings, however, a slight change of preference of a pivotal agent (or a few pivotal agents) may trigger off a bandwagon or cascading effect that will eventually lead to massive opposition and the subsequent demise of the noxious institution.

Other applications of possible cascading effects have been offered in relation to gender-biased norms and practices, like female genital mutilation, foot binding, early marriage, etc. We have thus learned from Mackie (1996, 2000) that, if all agents play sequentially, the movement leading to the demise of a noxious custom can start from a group of individuals who have a strong aversion to it. This group will gradually grow in size until a critical mass is reached and everybody is incited to reject the custom. In Platteau, Camiloti and Auriol (2018), agents play simultaneously, and the authors show that the distribution of individual preferences plays a major role in determining the type of effect that can be expected from changes occurring in (some) people’s expectations. In particular, if preferences are polarized (i.e., there are two masses of individuals with either very high or very low aversion to the custom), an interior stable equilibrium may exist. This is not the case if most people have a moderate aversion: it is then possible that either nobody or everybody follows the harmful practice. Assuming that the latter equilibrium is established, a rather slight change in the preferences of people with medium-high aversion to the practice may well spark a “revolution” whereby the former equilibrium becomes the only possibility: everybody now shuns the practice (pp. 330–331, 340–342).

Still another variant allows for a possible interim abortion of the cascading effect leading to radical institutional change. Called “stepping stone approach” by their authors (Gulesci et al., 2021), the new theory uses a dynamic, game-theoretic model that focuses on interim dynamics. However, unlike in the above-discussed models, all agents are assumed to have identical preferences. This new theory yields the

prediction that the abandonment of a harmful arrangement may be facilitated by the presence of a “mildly harmful” alternative acting as a stepping stone in the medium run. However, the possibility also exists that the intermediate alternative arrangement becomes absorbing, meaning that the move to a radical change is blocked at some intermediary point (illustrated by the adoption of a less invasive form of female genital cutting in the article referenced above).

28.4.4 Institutional Change as an Incremental Process Following a Shock

Finally, we want to draw attention to a recent literature which emphasizes the incremental nature of many institutional changes and is concerned with the dynamics of gradual changes that an exogenous shock may generate, when it happens. The works falling into this category have a distinct theoretical flavor, but they are consistent with many development experiences or episodes in which major institutional transformations have proceeded almost out of sight. For example, the franchise in nineteenth-century England was extended progressively at the parliamentary level (Reform Acts were enacted in 1832, 1867, and 1884) to be roughly replicated at the level of local government. In stark contrast to the interpretation of Acemoglu and Robinson (2000), who hypothesized that the extension of franchise was decided under the threat of revolution, Lizzeri and Persico (2004) proposed an alternative explanation that better fits the gradual character of this extension process. According to them, indeed, the elites willingly extended the franchise because elections with a broader franchise were more susceptible of satisfying the rising demand for local public goods that accompanied rapid urbanization. The underlying mechanism is the shift from clientelism and patronage (pork-barrel politics) to a more public-oriented legislative activity.⁵

Another illustration that comes to mind is the deep and gradual transformation of family structures and gender relations in Western and Central Europe, following the Gregorian reforms of the Catholic Church (Goody, 1973, 1983; Schulz, 2017; Guirking & Platteau, 2020, pp. 493–494; Henrich, 2020, pp. 180–182). These reforms consisted of prohibitions against practices aimed at maintaining the control of kinship groups over the transmission of property across generations, namely marriages arranged between close kin, levirate, adoption of children, concubinage, divorce, and remarriage. Their effects stretched over several centuries and reached all the countries under Rome’s influence, in particular.

⁵More precisely, special-interest politics prompts politicians to favor members of the elite who are swing voters, thus leaving other voters from the elite to suffer from this political system. Expansion of the franchise, with the resulting shift in policy toward public goods provision, is therefore predicted to be favored by those who are not swing voters. When the number of the latter voters exceeds 50%, a majority of the electorate will prefer to expand the franchise (Lizzeri & Persico, 2004, p. 713).

Bisin and Verdier (2023) have made an ambitious attempt to analyze how culture and institutions simultaneously evolve, interact, and jointly determine socio-economic outcomes. In their coevolutionary model, institutions are understood in a specific way that relates to the distribution of power in the society. One of their applications concerns the effectiveness of property rights protection. They assume the presence of two groups unequally ready to accept violence and, moreover, they are randomly matched in contests for wealth. They show that societies where the conflict-prone group is in strong control of institutions and property rights policy, but represents only a small fraction of the population (relative to the other, violence-averse group), will have limited rights protection and display institutional persistence. When, on the contrary, the same group is comparatively large so that its members have a high probability to interact and use violent means, there is a great social need for a reduction of violent effort and, hence, for protection of property rights. The interest of the violent-prone group is therefore to devolve institutional power to the other group so as to bring about that outcome. The limited rights protection equilibrium will then be gradually undone by the forces of cultural dynamics. More precisely, the induced diffusion of the conflict-prone culture (when violent-prone people are numerous) will reach a threshold beyond which endogenous institutional dynamics are triggered, ultimately resulting in greater devolution of power and better legal protection of property rights. In other words, a tipping point is reached at which violent-prone people begin to change their attitudes as they understand that low protection of property rights works against their own interests. A temporary institutional shock, such as an event that gives more formal power to the peaceful group, may therefore unleash persistent (long run) effects on the dynamics of culture and institutions, implying that the system will be irreversibly driven into a new long run trajectory of the institutional and cultural dynamics.

In the same vein of coevolutionary models, Bidner and François (2010) study the coevolution of behavioral norms and institutions, with institutions defined as trade-facilitating formal institutions (contract-enforcing institutions and legal protection of property rights). They seek to address the following challenge raised by Greif (1994): as new trading opportunities arise (an external shock), institutional changes are necessary because the initial vulnerability of institutions to opportunistic behavior will inescapably lead to the deterioration of norms.

In their framework, firms seek trade partners who may cheat and confiscate profit or be honest. A contract-enforcing institution detects cheaters with some probability, punishes them, and returns part of the embezzled profit to the cheated firm. The authors attempt to determine the impact of new trading opportunities on the probability that the economy will converge toward fully functional institutions—understood as the absence of cheaters—and toward institutions of good quality—as measured by a high probability of detecting cheaters.

For a given quality of institutions, a threshold for honesty norms is shown to exist below which opportunistic behavior by trading partners leads to the continuous deterioration of norms. This ominous trend can be counteracted by an improvement

of institutions, however. A “race” between institutions and honesty norms unfolds, the final outcome of which depends on the pace of institutional change, itself dependent on the additional trade that it generates. If the pace is not fast enough, honesty norms eventually vanish, leading to a breakdown in trade and a stifling of further institutional improvement. If, on the contrary, the pace of institutional change is fast enough, institutions become functional and, since honesty becomes incentive compatible, honesty norms stand reinforced.

Moreover, when the value of additional trade is high, the incentives for institutional change are strong and institutions improve quickly. They are then capable of making up for the decline of honesty norms and directing the economy towards a steady state with no cheater, that is, where institutions are fully functional. When (additional) trade is of low value, institutions improve more slowly, and they cannot prevent the economy from converging toward an equilibrium with fully dysfunctional institutions and where honesty norms have declined. Trade is then maintained only because of the existence of high-quality institutions, which are able to retrieve part of the firms’ losses when they have been cheated.

Within this overall framework, any exogenous increase in trade opportunities—or the size of the economy—is predicted to expand the attraction area of the fully functional institutional equilibrium: an economy is more likely to have high trust levels the larger its population size. The authors verify that this is actually the case with the help of a cross-section regression performed in a sample of 72 countries observed in 2005.

Another striking result is that there is no strict equivalence between “functional” and high-quality institutions. The reason is that low- or middle-quality institutions can still be “functional” if they are associated with a high level of trust and if they stimulate trade. The authors cite the example of China, a country where protection of individual property rights and the rule of law are rather poorly developed and, yet, trade has not been inhibited presumably, according to their theory, because of an (unusually) high level of trust. Bidner and François’ conclusion thus runs counter to the well-known argument that good institutions are needed to create trustworthy people by detecting and punishing opportunistic acts (Greif, 1994; Dixit, 2004).

In another important paper, the same authors propose a dynamic political game between leaders and citizens which also relies on cultural transmission (see Bidner & François, 2013). They explain how democratic accountability can be gradually consolidated following a shock consisting of the appearance of a leader who acts accountably (that is, they do not abuse her/his office for personal gain). A special feature of the model, which distinguishes it from standard political agency models, is the assumed presence of a few unconditional cooperators, called “gripped” voters, in the population: these voters are viscerally opposed to corrupt political leaders and, therefore, they demand their removal from office even if such punishment will not improve the performance of elected leaders in the future. In this framework, it is possible that voters become increasingly intolerant of corrupt leaders and political accountability comes to be eventually established at equilibrium.

28.4.5 Strengths and Limitations

The ability to explore long-term trajectories of societies is obviously an advantage over models that only explain how an exogenous shock may cause a shift from one equilibrium outcome to another. Yet, the dynamic efficiency of the trajectories found in gradual change models is typically not discussed. Moreover, an important limitation of these models is that they often assume the irreversibility of the processes under study. This is a serious problem given the frequent occurrence of institutional detours, stalemates, or turnarounds and regressions. Too often, we have indeed witnessed unexpected backsliding of reforms, lapses following breakthroughs, or returns to prior stages of institutional development and progress.

On the other hand, it bears emphasis that, for reasons of analytical tractability, dynamic considerations are generally introduced in economic theories at the cost of relaxing some usual behavioral assumptions, typically those regarding rationality. It is therefore not surprising that in most if not all gradual change models, some form of limited rationality is assumed. More precisely, agents act strategically, and their strategies take the evolutionary dimension of the problem into account, yet they can do so only myopically. Two remarks need to be made here. First, the assumptions limiting the forward-lookingness of agents are less restrictive in these models than in the crude evolutionary theory which we discussed in Sect. 28.2. In the latter, agents do not indulge in any strategic calculus or, if they do, they ignore that they are part of an evolving environment. This ignorance means that the Nash equilibrium played by them is an element of the evolutionary process, but the evolution itself is never an argument of their optimizing function. Typically, the theory then consists of superimposing on a static game some dynamics meant to describe the flow of the distribution of strategies among the population (Blume, 2002, p. 2).

Second, assuming that individuals make decisions somehow unrelated to considerations of the future should not necessarily be considered as a factor limiting the plausibility of the models under concern. It can even be argued that in some circumstances behavioral assumptions of limited rationality are more realistic than the assumption of perfect (strategic) rationality. There is solid ground to take the following result seriously: it is hard to demonstrate how, after the disruption of a Nash equilibrium following a shock, fully rational agents converge to a new equilibrium through the dual role of action-choices and expectations (Aoki, 2001). This is plausibly a reflection of the limited ability of rationality to solve dynamic problems that arise in reality (Young, 1998).

To be able to explore dynamic trajectories along the line indicated above, economists tend not only to limit the rationality of (some of) their actors, but they also introduce mechanisms that are rather abstract and may even look like analytically ingenious but convenient artefacts or stratagems. Thus, in Bisin and Verdier (2023), the institutions involved are not really described but are subsumed under an endogenous distribution of power reflected in Pareto weights. And in Bidner and François (2010), a central role is played by the so-called “institutional designer,” a fictional player standing for a benevolent political agency. Such appeals to abstract

mechanisms seem to be the price to pay for avoiding recourse to black box simulation models that fail to yield intuitive explanations.

28.5 A Broad Assessment of Achievements and Challenges

From our bird's-eye view of the economics of institutional change from the Second World War to this date, it is evident that the field has been far from stagnant as new perspectives have been developed in different directions and in close relation to the progress of general economic theory. This is even manifest when we look back at the excellent survey that Lin and Nugent (1995) wrote toward the middle of the period covered in the present paper. Although they paid significant attention to the role of politics, there is no doubt that the political economy of development has received a great impetus at the start of the twenty-first century under the influence of the works of Acemoglu and Robinson, Besley and Persson, and others. Equally impressive has been the proliferation of empirical works inspired by the equilibrium-of-the-game approach, and the rising influence of the evolutionary theory. Regarding the latter, perhaps the most striking novelty lies in the recent emergence of genuinely dynamic models of gradual change, including attempts to understand the coevolution of institutions and culture.

One of the lessons that seems to come out of this quick survey is that, despite the aforementioned recent developments, none of the old approaches deserves to be discarded. Thus, why should we dispense with the idea of induced institutional innovation and the related approaches of transaction-cost economics and the principal-agent model when so many insightful applications have been proposed to important issues of institutional change? What our discussion nevertheless shows, and what Lin and Nugent actually stressed in their 1995 survey, is that the domain of application of these frameworks is necessarily limited to certain forms of institutional change, in particular those that arise at a micro-level and do not involve complex coordination problems.

On the other hand, the equilibrium-of-the-game approach, born of classical game theory, has greatly contributed to our understanding of institutional stagnation even when pressing or latent demand for change exists. By bringing into light the critical role of individual expectations and the possibility of multiple self-enforcing equilibria, it has allowed economists to analyze social norms, customs, and informal arrangements, although it is not clear that they qualify as institutions. The realm of institutional economics was thus suddenly enlarged so as to include subjects that were considered until then as the preserve of sociological studies. Furthermore, the approach prompted a spate of empirical works based on the idea that the impact of institutions, or their cultural determinants, can be properly identified through the detection of exogenous shocks that occurred in a somehow distant past. An oft-heard criticism against the works born of this new economic history is that although they are strong on identification they tend to be weak in terms of explanatory mechanisms. Yet even the claim to proper identification of causality calls for serious caveats in contexts where institutional, cultural, and geographical characteristics are often deeply entangled.

Caution is likewise needed when status quo situations are explained by the existence of strong social norms or customs. Alternative explanations may indeed be available that require to be considered. In particular, institutional stagnation may be caused by political factors, such as opposition to reforms by powerful interest groups linked to the political elite. Clearly, the political economy approach has marked a significant progress over all previous approaches where the state was absent or (implicitly) assumed to be benevolent. It nevertheless suffers from two main limitations. First, many models of political economy are quite stylized, particularly in the sense that only general categories of actors are considered, or some important types of actors are ignored, or else their action spaces are unduly restrained. Moreover, the objective functions of the political leaders are typically simplified, and the extent to which they are endowed with rationality is questionable. Second, the empirical support of political economic theories tends to be weak, as it often consists of short and cursory narratives, which are vulnerable to criticisms from political scientists and historians.

The latter limitation also applies to the most recent approach to institutional change, which we may label the gradual change school. This is a greatly appealing way to address institutional issues since in reality what appears as an abrupt change may often be the outcome of a succession of gradual alterations or decaying of prevailing institutions. On the other hand, an exogenous shock, instead of instantaneously leading to a new institutional configuration, most plausibly gives rise to successive adaptations that may end up in different institutional situations depending on the transition path. The rising availability of genuinely dynamic analyses adopting the foregoing perspective is a real bonus for development economists eager to gain a better grasp of actual change processes. But it must be reckoned that the same quantitative rigor as demanded in studies based on historical experiments cannot be attained when complex evolutionary or coevolutionary mechanisms are at the center of attention. The construction of narratives is the way to proceed empirically when political economy and evolutionary models provide the theoretical scaffolding to guide thinking about institutional change. What needs to be stressed is that the narratives presented by economists in support of their theories, whether inspired by political or evolutionary economics, should be more elaborate than they typically are. They should not appear as historical material conveniently chosen and superficially glanced at, but as summaries anchored in a deep knowledge of the existing literature.

This also implies that the period covered should be long enough to ensure that the account of institutional change is not too dependent on peculiar circumstances. Just to take one example, there is no such thing as a rapid and smooth transition from autocracy to democracy.⁶ Regimes trying to carry out this transition face all sorts of difficulties, which partly arise because while the associated costs are immediate the benefits often take a longer time to materialize. Citizens may therefore be

⁶Another example can be found in the survey of Guirking and Platteau (2020), in which the authors put great emphasis on the non-monotonous transformation of family structures in Europe during the last centuries.

discouraged and turn their back on political reforms, especially if they were sold grand promises by popular leaders. In short, “the path to democracy is fraught,” and this is why “history is littered with failed experiments” (Economist, 2023, p. 66). A straightforward implication is that a comparison between countries that eventually succeeded to become democratic after, say, several decades, and countries entrapped in autocracy, is somehow biased: it bypasses all the countries which are more or less in-between and struggle to avoid the pitfalls of democratization or even got stuck on the way (Campos & Coricelli, 2002).

Relatedly, a strong and consistent relationship has been found between higher income and both democratization and democratic survival in the medium term (10–20 years), but not necessarily in shorter time (Treisman, 2015, 2020). Hence the need for a “conditional modernization theory,” which can account for the lags. In Treisman’s view, the effect of development on democracy is triggered by disruptive events such as economic crises, military defeats or, most generally, leader change. Political outcomes then depend on both the development level and how citizens coordinate (at intermediate income ranges).

These examples testify that processes based on unidirectional causality and monotonous transformation tend to appear less pertinent than those allowing for various kinds of feedbacks, spillovers, tipping points, detours, stalemate, and “fortune reversals.” Reversals and setbacks are much more important than the first institutional and development economists imagined: they were used to think in terms of smooth and monotonous institutional changes sparked by some Big Push or massive shock evoking the postwar Marshall Plan. Theories that predict the possibility of setbacks and reversals are still in their infancy, and this is certainly an area of research that deserves a lot of attention in the future.

Two central messages emerge from the foregoing discussion. The first message is that an eclectic approach to the various methods available to study institutional change is advisable. They all have their strengths and weaknesses, and they tend to offer complementary insights into different dimensions of the problems at stake. Second, the deepening of narratives adduced in support of theory must necessarily accompany further advances in political economics and evolutionary approaches in the years to come.

Acknowledgement of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D. (2003). Why not a political Coase theorem? Social conflict, commitment, and politics. *Journal of Comparative Economics*, 31(4), 620–652.
- Acemoglu, D., & Robinson, J. A. (2000). Political losers as a barrier to economic development. *American Economic Review*, 90(2), 126–130.
- Acemoglu, D., & Robinson, J. A. (2006). Economic backwardness in political perspective. *American Political Science Review*, 100(1), 115–131.

- Akerlof, G. (1976). The economics of caste and of the rat race and other woeful tales. *Quarterly Journal of Economics*, 90(4), 599–617.
- Alesina, A., & Giuliano, P. (2015). Culture and institutions. *Journal of Economic Literature*, 53(4), 898–944.
- Alesina, A., Matuszewski, J., & Easterly, W. (2011). Artificial states. *Journal of the European Economic Association*, 9(2), 246–277.
- Allen, D. W., & Lueck, D. (2002). *The nature of the farm: Contracts, risk, and organization in agriculture*. The MIT Press.
- Aoki, M. (2001). *Towards a comparative institutional analysis*. MIT Press.
- Auriol, E., & Platteau, J.-P. (2017a). The explosive combination of religious decentralisation and autocracy: The case of Islam. *The Economics of Transition*, 25(2), 313–350.
- Auriol, E., & Platteau, J.-P. (2017b). Religious co-option under autocracy: A theory inspired by history. *Journal of Development Economics*, 127, 395–412.
- Auriol, E., Platteau, J.-P., & Verdier, T. (2023). The Quran and the Sword. *Journal of the European Economic Association*, 21(5), 1772.
- Axelrod, R. (1984). *The evolution of cooperation*. Basic Books.
- Axelrod, R. (1997). *The complexity of cooperation: Agent-based models of competition and collaboration*. Princeton University Press.
- Baland, J.-M., & Platteau, J.-P. (1996). *Halting degradation of natural resources: Is there a role for rural communities?* Oxford Clarendon Press.
- Baland, J.-M., Bourguignon, F., Platteau, J.-P., & Verdier, T. (2020). Economic development and institutions: An introduction, Chapter 1. In J.-M. Baland, F. Bourguignon, J.-P. Platteau, & T. Verdier (Eds.), *The handbook of economic development and institutions* (pp. 1–20). Princeton University Press.
- Barsbai, T., Rapoport, H., Steinmayr, A., & Trebesch, C. (2017). The effect of labor migration on the diffusion of democracy: Evidence from a former Soviet Republic. *American Economic Journal: Applied Economics*, 9(3), 36–69.
- Basu, K. (2000). *Prelude to political economy: A study of the social and political foundations of economics*. Oxford: Oxford University Press.
- Basu, K., Jones, E., & Schlicht, E. (1987). The growth and decay of custom: The role of the new institutional economics in economic history. *Explorations in Economic History*, 24(1), 1–21.
- Bauer, P. T., & Yamey, B. S. (1957). *The economics of under-developed countries*. Cambridge University Press.
- Besley, T., & Coate, S. (1998). Sources of inefficiency in a representative democracy: A dynamic analysis. *American Economic Review*, 88(1), 139–156.
- Bidner, C., & François, P. (2010). Cultivating trust: Norms, institutions and the implications of scale. *The Economic Journal*, 121, 1097–1129.
- Bidner, C., & François, P. (2013). The emergence of political accountability. *Quarterly Journal of Economics*, 128(3), 1397–1448.
- Binswanger, H. P., & Rosenzweig, M. (1986). Behavioral and material determinants of production relations in agriculture. *Journal of Development Studies*, 22(3), 503–539.
- Bisin, A., & Verdier, T. (2023). On the joint evolution of culture and political institutions: Elites and civil society. *Journal of Political Economy*, 132, 1485–1564.
- Blume, L. (2002). *Stigma and social control* (HS economic series, working paper 119). Institute for Advanced Studies.
- Bruton, H. J. (1965). *Principles of development economics*. Prentice-Hall.
- Campos, N. F., & Coricelli, F. (2002). Growth in transition: What we know, what we don't, and what we should. *Journal of Economic Literature*, 40, 793–836.
- Cantoni, D., & Yuchtman, N. (2021). Historical natural experiments: Bridging economics and economic history, Chapter 8. In A. Bisin & G. Federico (Eds.), *Handbook of historical economics* (pp. 213–241). Elsevier & North Holland.
- Dixit, A. (2004). *Lawlessness and economics*. Princeton University Press.
- Dixit, A. (2007). Evaluating recipes for development success. *The World Bank Research Observer*, 22(2), 131–157.

- Docquier, F., Lodigiani, E., Rapoport, H., & Schiff, M. (2016). Emigration and democracy. *Journal of Development Economics*, 120, 209–223.
- Economist. (2023). The price of a vote. *Economist*, 19–25 Aug, Finance & Economics, Free Exchange, p. 66.
- Engerman, S. L., & Sokoloff, K. L. (1997). Factor endowments, institutions, and differential paths of growth among new world economies: A view from economic historians of the United States. In S. Haber (Ed.), *How Latin America fell behind: Essays on the economic histories of Brazil and Mexico, 1800–1914* (pp. 260–304). Stanford University Press.
- Geertz, C. (1973). *The interpretation of cultures*. Basic Books.
- Goody, J. (1973). Strategies of heirship. *Comparative Studies in Society and History*, 15(1), 3–20.
- Goody, J. (1983). *The development of the family and marriage in Europe*. Cambridge University Press.
- Greif, A. (1989). Reputations and coalitions in medieval trade: Evidence on the Maghribi traders. *Journal of Economic History*, 49(4), 857–882.
- Greif, A. (1994). Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy*, 102(5), 912–950.
- Greif, A. (2006). *Genoa and the Maghribi traders: Historical and comparative institutional analysis*. Cambridge University Press.
- Greif, A., & Laitin, D. (2004). A theory of endogenous institutional change. *American Political Science Review*, 98(4), 633–652.
- Guirking, C., & Platteau, J.-P. (2015). Transformation of the family farm under rising land pressure: A theoretical essay. *Journal of Comparative Economics*, 43(1), 112–137.
- Guirking, C., & Platteau, J.-P. (2020). The dynamics of family systems: Lessons from past and present times, Chapter 13. In J.-M. Baland, F. Bourguignon, J.-P. P. Platteau, & T. Verdier (Eds.), *The handbook of economic development and institutions* (pp. 449–519). Princeton University Press.
- Gulesci, S., Jindani, S., La Ferrara, E., Smerdon, D., Sulaiman, M., & Young, H. (2021). A stepping stone approach to understanding harmful norms. CEPR discussion paper no. 15776, London.
- Hayami, Y. (1997). *Development economics: From the poverty to the wealth of nations*. Clarendon Press.
- Hayami, Y., & Kikuchi, M. (1981). *Asian Village economy at the crossroads*. University of Tokyo Press/Johns Hopkins University Press.
- Hayami, Y., & Ruttan, V. (1985). *Agricultural development: An international perspective*. The Johns Hopkins University Press.
- Hayek, F. (1960). *The constitution of liberty*. Routledge & Kegan Paul.
- Hayek, F. (1973). *Law, legislation, and liberty. Vol. 1: Rules and order*. University of Chicago Press.
- Helmke, G., & Levitsky, S. (2004). Informal institutions and comparative politics. *Perspectives on Politics*, 2(4), 725–740.
- Henrich, J. (2020). *The weirdest people in the world: How the west became psychologically peculiar and particularly prosperous*. Allen Lane.
- Hicks, J. R. (1932). *The theory of wages*. Macmillan.
- Higgins, B. (1968). *Economic development: Problems, principles and policies*. W.W. Norton & Co. [first edition 1959].
- Hirshman, A. O. (1958). *The strategy of economic development*. Yale University Press.
- Iliffe, J. (2007). *Africans: The history of a continent*. Cambridge University Press.
- Khemani, S. (2020). Political economy of reform. In A. Dixit & J. Hamilton (Eds.), *Oxford research encyclopedia of economics and finance* (pp. 1–28). Oxford University Press.
- Kuran, T. (1987). Preference falsification, policy continuity, and collective conservatism. *Economic Journal*, 97(387), 642–665.
- Kuran, T. (1988). The tenacious past: Theories of personal and collective conservatism. *Journal of Economic Behavior and Organization*, 10(2), 143–171.
- Kuran, T. (1995). *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press.

- La Ferrara, E. (2007). Descent rules and strategic transfers: Evidence from matrilineal groups in Ghana. *Journal of Development Economics*, 83(2), 280–301.
- Leibenstein, H. (1957). *Economic backwardness and economic growth: Studies in the theory of economic development*. Wiley.
- Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *The Manchester School*. Reprinted in A. N. Agarwala & S. P. Singh (Eds.), *The economics of underdevelopment* (pp. 400–449). Oxford University Press (1958).
- Lewis, W. A. (1955). *Theory of economic growth*. Harper & Row.
- Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press.
- Li, S. (2003). The benefits and costs of relation-based governance: An explanation of the east Asian miracle and crisis. *Review of International Economics*, 11(3), 651–673.
- Lin, J. Y., & Nugent, J. B. (1995). Institutions and economic development, Chapter 38. In J. Behrman & T. N. Srinivasan (Eds.), *Handbook of development economics* (Vol. 3A, pp. 2301–2370). Elsevier.
- Lizzeri, A., & Persico, N. (2004). Why did the elites extend the suffrage? Democracy and the scope of government, with an application to Britain's age of reform. *Quarterly Journal of Economics*, 119(2), 707–765.
- Mackie, G. (1996). Ending Footbinding and infibulation: A convention account. *American Sociological Review*, 61(6), 999–1017.
- Mackie, G. (2000). Female genital cutting: The beginning of the end. In B. Shell-Duncan & Y. Hernlund (Eds.), *Female circumcision in Africa: Culture, controversy, and change* (pp. 253–282). Lynne Rienner Publishers.
- Mahoney, J., & Thelen, K. (Eds.). (2010). *Explaining institutional change: Ambiguity, agency, and power*. Cambridge University Press.
- Meier, G. M., & Baldwin, R. E. (1957). *Economic development*. John Wiley & Sons.
- Michalopoulos, S., & Papaioannou, E. (2014). National Institutions and subnational development in Africa. *Quarterly Journal of Economics*, 129(1), 151–213.
- Michalopoulos, S., & Papaioannou, E. (2020). Historical legacies and african development. *Journal of Economic Literature*, 58(1), 53–128.
- Montet, C., & Serra, D. (2003). *Game theory and economics*. Palgrave Macmillan.
- Myrdal, G. (1968). *Asian Drama: An inquiry into the poverty of nations*. Penguin Books.
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- Nurkse, R. (1953). *Problems of capital formation in underdeveloped countries*. Oxford University Press.
- Padro-i-Miguel, G., Qian, N., Yao, Y., & Xu, Y. (2014). Making democracy work: The effects of social capital and elections on public goods in China. *Society for Economic Dynamics*, WP No. 1414.
- Platteau, J.-P. (2000). *Institutions, social norms, and economic development*. Routledge.
- Platteau, J.-P., Camilotti, G., & Auriol, E. (2018). Eradicating women-hurting customs: What role for social engineering? In S. Anderson, L. Beaman, & J.-P. Platteau (Eds.), *Towards gender equity and development* (pp. 319–356). Oxford University Press.
- Portes, A. (2006). Institutions and development: A conceptual reanalysis. *Population and Development Review*, 32(2), 233–262.
- Putnam, R. D. (1993). *Making democracy work: Civic traditions in modern Italy*. Princeton University Press.
- Rodrik, D. (2004). Institutions and economic performance: Getting institutions right. CESifo DICE Report, Institut für Wirtschaftsforschung, München Universität, 2(2), 10–15.
- Rohner, D., & Zuravskaya, E. (2023). Nation building: What could possibly go wrong?, Chapter 1. In D. Rohner & E. Zuravskaya (Eds.), *Nation building: Big lessons from successes and failures* (pp. 1–23). CEPR Press.
- Schotter, A. (1981). *The economic theory of social institutions*. Cambridge University Press.
- Schulz, J. F. (2017). *The churches' bans on consanguinity, kin-networks and democracy*. Yale University.

- Spolaore, E., & Wacziarg, R. (2013). How deep are the roots of economic development? *Journal of Economic Literature*, 51(2), 325–369.
- Sugden, R. (1986). *The economics of rights, co-operation and welfare*. Basil Blackwell.
- Sugden, R. (1989). Spontaneous order. *Journal of Economic Perspectives*, 3(4), 85–97.
- Treisman, D. (2015). Income, democracy, and leader turnover. *American Journal of Political Science*, 59(4), 927–942.
- Treisman, D. (2020). Economic development and democracy: Predispositions and triggers. *Annual Review of Political Science*, 23, 241–257.
- Voigtländer, N., & Voth, H.-J. (2013). How the west ‘invented’ fertility restriction. *The American Economic Review*, 103(6), 2227–2264.
- Voth, H.-J. (2021). Persistence: Myth and mystery, Chapter 9. In A. Bisin & G. Federico (Eds.), *Handbook of historical economics* (pp. 243–267). Elsevier North-Holland.
- Young, H. P. (1996). The economics of convention. *Journal of Economic Perspectives*, 10(2), 105–122.
- Young, H. P. (1998). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Institutions and Development: Persistent Puzzles

29

Mary M. Shirley 

Contents

29.1	The Challenge of Development	730
29.2	Which Institutions Must Function Effectively for Countries to Prosper and How Have They Evolved?	732
29.2.1	Institutions that Foster Exchange	732
29.2.2	Institutions that Protect Property and Persons	735
29.3	What Explains Why Countries Fail to Create the Rules and Norms that Foster Growth and Social Progress?	737
29.3.1	Elite Dominance	738
29.3.2	Colonial Heritage	739
29.3.3	Colonial Heritage Plus Factor Endowments	740
29.3.4	Beliefs and Culture	742
29.3.5	All of the Above	744
29.4	How Can Developing Countries Improve Their Institutions?	745
29.5	Can Outsiders Promote Institutional Development?	748
29.6	What Next?	750
	References	752

Abstract

How does institutional economics explain development and underdevelopment? Poverty has declined drastically over recent decades, but the proportion of the global population living in high-income countries with strong institutions (14%) has barely budged. This chapter summarizes how institutionalists address four questions: Which institutions must function effectively for countries to converge with high-income countries? What are the obstacles to developing such well-functioning institutions? How can countries improve their institutions? And can outsiders, including foreign aid agencies, help them develop? While research has made much progress in answering these questions, persistent puzzles remain.

M. M. Shirley (✉)
The Ronald Coase Institute, Philomont, VA, USA
e-mail: mshirley@coase.org

Keywords

Institutional change · Development · Underdevelopment · Poverty reduction · Foreign aid · Path dependency · Elite dominance · Culture

29.1 The Challenge of Development

Rapid economic growth has drastically reduced the proportion of the world's population living in poverty, even as global population has increased. The percentage of people living below the World Bank's extreme poverty line (US\$2.15 a day in 2017 dollars adjusted to reflect purchasing power) fell from 44% in 1981 to only 8.5% in 2019. This means that the number of people in extreme poverty fell from almost 2 billion in 1981 to 659 million in 2019 (World Bank, 2023a), a period when the world's population grew by 2.4 billion.¹ This welcome change was largely thanks to growth in China and India, which lifted millions above the poverty line.² As a result, by 2019, the vast majority of people—75%—lived in middle-income countries, compared to only 9% in low-income countries (Prydz & Wadhwa, 2019).³ If, however, we also include measures of the standard of living, education, and health, there are still over 1.1 billion people living in impoverished circumstances, half of them in Sub-Saharan Africa.⁴ Despite the decline in poverty, the proportion of the population living in wealthy developed countries barely increased from 15% in 1993 to 16% in 2019 (Prydz & Wadhwa, 2019). Only 14% of the world's population lives in high-income countries with strong institutions.⁵ What is the explanation for these circumstances?

New institutional economics (NIE) emphasizes institutions as the root cause of economic development. Yet most explanations for why so many countries moved from lower-income to middle-income do not focus on institutional improvements but on policy changes, especially greater openness to global trade and freer markets that propelled a shift of resources from lower- to higher-productivity sectors, typically from agriculture into industry (Diao et al., 2017; Rodrik, 2022). Nevertheless, there were

¹ World population during this period grew from 5.3 billion to 7.7 billion (World Bank, 2023b).

² China's GNI per capita grew from \$330 in 1990 to \$9470 in 2018 and India's grew from \$380 to \$2020 in 2017 purchasing power parity dollars (Prydz & Wadhwa, 2019).

³ Middle-income countries are defined by the World Bank as those with a per capita gross national income between \$1026 and \$12,375 in 2019 (calculated using the World Bank Atlas method). Those below the \$1026 threshold are defined as "low-income" and those above \$12,375 are "high-income" (Prydz & Wadhwa, 2019).

⁴ Based on the United Nations Development Program multidimensional poverty index, which includes measures of nutrition, child mortality, years of schooling, school attendance, and standard of living indicators such as drinking water, electricity, housing, etc. (UNDP, 2023).

⁵ Most, but not all high-income countries have well-functioning institutions. Excluding (mostly oil-rich) countries with low scores on the World Bank governance indicators (rule of law, voice and accountability, control of corruption, political stability and regulatory quality), 14% of the world's population in 2021 was in high-income countries with strong institutions (author's estimate based on World Bank, 2023b).

important institutional improvements underlying these structural transformations, which become more apparent when we look in depth at individual cases. For example, improvements in China's contract enforcement institutions underpinned the structural shift of labor into manufacturing, as I describe later. Institutions also explain why so few countries have converged with the top 14% of countries. Middle-income countries may have improved their institutions for enforcing contracts, but their institutions protecting investors from state or elite expropriation are still weak (Acemoglu & Johnson, 2005). As discussed below, the institutions that protect against expropriation include impersonal rule of law, civic engagement, checks on arbitrary state actions (such as balance of powers), and systems of accountability, including democracy and a free press. Most middle-income countries (excluding oil rich countries) that have converged with the prosperous few developed such institutions. For example, almost two thirds of the countries that moved into the World Bank's high-income category in the early 2000s were Central European countries that strengthened their institutions to protect against state or elite expropriation as part of their commitments to the European Union.⁶ But most middle-income countries lack these protections and that has hampered them from responding to rising manufacturing wages and tightening global trade by investing in new technology and shifting more resources into domestic demand-driven manufacturing and services.

Most of the literature on institutions and development uses North's definition of institutions as the "humanly devised constraints that structure human interaction" (1990, p. 3). North's definition encompasses formal rules, such as written constitutions and laws enforced by third-parties, and informal constraints, such as norms, conventions, and self-imposed codes of conduct that are unwritten and self-enforced or enforced by communities. To North, institutions are the "rules of the game in a society." Organizations, "groups of individuals bound together by some common purpose to achieve certain objectives" (such as legislatures, firms, trade unions, churches, clubs, schools, etc.), are the players of the game. While many simple rules and norms change constantly, the institutions of interest to NIE are a society's durable and fundamental constitutions, laws and norms that form what North termed its institutional framework.

At one time mainstream economics did not include institutions in growth models, focusing instead on investment, technological change, education, and similar variables as the causes of development. Yet investment, technological innovation, or years of schooling, are not correlated with higher rates of economic growth (Easterly, 2002), while cross-country regressions persistently demonstrate large and statistically significant correlations between institutional variables and growth.⁷ These regressions have their flaws (discussed in Shirley, 2008, Chap. 5), but the significant

⁶Hagemejer et al. (2021) find large gains from EU membership compared to counterfactuals. Most of the other converging countries were Latin American countries that had slowly developed such institutions over many years, such as Panama, Chile, and Uruguay.

⁷For cites of cross-country regressions using institutions as an explanatory variable for growth and development see Shirley (2008), Table 5.1 and Durlauf (2018), Table 1.

correlation between institutions and growth has endured over numerous studies even as data sets and methodologies have improved. According to North and Thomas (1973, p. 2), non-institutional factors such as innovation, economies of scale, education, capital accumulation, and so on “are not causes of growth; they *are* growth” (italics in original). The fundamental explanation for differences in growth is differences in institutions. Not all scholars accept this argument. Some (e.g., Sachs, 2003) argue that geography is the ultimate cause of development because of its effects on agriculture, diseases, and transport. But institutionalists counter that geography’s effect operates through institutions (see the discussion of colonial heritage and factor endowments below, also Koyama and Rubin (2022)). Thus, Rodrik et al. (2002) find only weak evidence that geography has a direct effect on growth but strong evidence that geography’s effect is indirect, through institutions.

Which institutions must function effectively if countries are to develop to the point where lower-income countries move to middle-income and the middle-income converge with high-income countries? What are the obstacles to establishing these well-functioning institutions? How can developing countries improve their institutions? Can outsiders promote institutional development? New institutional economics has made progress towards answering these questions, but many puzzles remain. The rest of this chapter takes stock of how the NIE has addressed these questions and proposes research to address the gaps in our understanding. I begin by exploring the two types of institutions that must function well for countries to prosper.

29.2 Which Institutions Must Function Effectively for Countries to Prosper and How Have They Evolved?

New institutional economics holds that to meet the challenge of development countries need an institutional framework that supports a market economy, which includes two distinct and not necessarily complementary sets of institutions: (i) those that foster exchange by supporting private contracts, lowering transaction costs, and encouraging trust, and (ii) those that influence the state and other powerful actors to protect private property and persons rather than expropriate and subjugate them (see Acemoglu & Johnson, 2005; North, 1981; Shirley, 2008).⁸

29.2.1 Institutions that Foster Exchange

The current literature on the importance of institutions for exchange is rooted in Ronald Coase’s theory of transaction costs. As Coase explained, what are traded on

⁸I use the term development to mean countries which have achieved a level of per capita income that puts them in the World Bank’s high-income category (above \$12,375 in 2019), as well as high scores on selected social indicators (e.g., life expectancy at birth of over 70 years, infant mortality rate of less than 10 per 1000, adult literacy rates of 100%).

the market are not, as is often supposed, commodities, but rather the rights to take certain actions. Thus, for NIE, transactions involve exchanging the rights to use a resource, and transaction costs are the costs to establishing, transferring, and policing these rights (see Alchian, 1965; Allen, 2000; Ménard & Shirley, 2022, Chap. 2).⁹ Coase maintained that the effects of transaction costs “are pervasive in the economy. Businessmen, in deciding on their ways of doing business and on what to produce, have to take into account transaction costs. If the costs of making an exchange are greater than the gains that exchange would bring, that exchange would not take place and the greater flow of production from specialization would not be realized” (Coase, 1992, p. 716). The costs of transacting, as Coase explained, “depend on the institutions of a country: its legal system, its political system, its social system, its educational system, its culture, and so on” (1998, p. 73). When information is costly and rights are poorly protected, contracts become hard to specify and enforce and transaction costs are high. This results in less economic development, since societies with persistently higher transaction costs have less trade, fewer firms, less specialization, less investment, and lower productivity.

According to North (1990, 2005), the evolution of institutions that support exchange begins with small communities producing at low levels of specialization and relying largely on face-to-face barter trade between individuals who know one another and who typically share kinship, ethnic, religious, or similar ties. Agreements are enforced by informal mechanisms such as family loyalties, ostracism, or coercion by private groups.¹⁰ Over time the groups engaging in exchange tend to grow through natural population growth, urbanization, migration, and as more and more people begin to see advantages in trade. Improvements in agricultural and industrial technology, medicine, and education combine to increase human control over the environment, improve nutrition, and reduce disease, contributing to population growth, urbanization, and the expansion of markets. With the expansion in the size and geographic area of the trading group and the rise of urban centers, traders envision lucrative opportunities to do business with people who live even further away and do not belong to their networks. Merchants and investors seek more information about these unknown trading partners and demand better mechanisms to enforce agreements between strangers.

Up to a point, parties to contracts may be able to rely largely on norms and networks to enforce agreements between strangers. Greif (1993, 2006) describes how the Maghribi traders used an extensive network of communication, social ties, a common language, and a common religion (Judaism), to share information on the behavior of their agents and to assure that dishonest agents were collectively

⁹Transaction costs also include what Dahlman described as “search and information costs, bargaining and decision costs, policing and enforcement costs” (Dahlman, 1979, p. 148). These are the costs of finding a trading partner, deciding on the terms of the trade, drawing up a contract, monitoring and enforcing a contract, and the like.

¹⁰Private coercion can increase as well as decrease uncertainty. Bates (2001) describes how clashes between private security forces among kinship groups can lead to retaliation and a cycle of violence that spans generations.

punished. This framework motivated agents to develop a reputation for honest dealing, allowing the Maghribi to safely rely on agents who were not part of their family or community.

To take advantage of opportunities for profits and to respond to increased competition in home markets, trading parties begin to devise additional contractual safeguards; for example, one party might pledge an asset as a “hostage” that is forfeit if they renege on the agreement, much as people once enforced treaties by sending a family member to act as a hostage in a show of good faith (Williamson, 1985, 1996). In the twelfth and thirtieth centuries, towns, cities, and communes created “community responsibility systems” to guarantee the trustworthiness of their members when they transacted in distant markets, such as the annual Champagne fairs (Greif, 2006). The community was obligated to pay compensation if a member defaulted, assuring enforcement of contracts with unknown merchants who were not part of any local social network but known in their home communities.

Even today, traders rely on private means to enforce contracts (Williamson, 1985), which were vital to the successful development of private enterprises in China, for example (see ► Chap. 35 by Shirley and Xu in this Handbook). Norms of trust and cooperation reduce transaction costs and foster exchange (see for example, Knack & Keefer, 1997, and ► Chap. 37 by Keefer in this Handbook). However, enforcing bargains through networks and norms has drawbacks. Norms rooted in one group’s history and culture are not easily transferred and enforcement via networks requires information on group membership. As a result, opportunities for lucrative trades between those not able to use the enforcement mechanism are lost.¹¹

For impersonal exchange between strangers, private ordering and norms of conduct need the support of third-party rules and enforcement. This happened in Europe, for example, when the codes of conduct of guild merchants evolved into merchant law and were gradually integrated into the body of laws enforced by states (North & Thomas, 1973; Milgrom et al., 1990). Written contracts and rules with third-party enforcement were important to expanding trade in Western Europe during the Middle Ages (even though they do not always reduce transaction costs, since litigation and legalism can hinder trade as well as help it).

The experience of countries like China and India suggest that informal institutions can support exchange among far flung networks and locations well beyond what the European experience might suggest. Yet these countries also have improved their formal legal rules to support contract enforcement. For example, Chemin (2007, p. 1) found that India’s 2002 Code of Civil Procedure Amendment Act simplified and shortened judicial procedures and led to “fewer breaches of contract, encouraged investment, facilitated access to finance, and expanded rental markets.”

Contract enforcement institutions, formal or informal, can only take the economy so far. People will be more willing to specialize and invest the more their personal and property rights are protected from expropriation, and I turn to that subject next.

¹¹ The pros and cons of informal versus formal institutions are reviewed in Keefer and Shirley (2000).

29.2.2 Institutions that Protect Property and Persons

Acemoglu and Johnson (2005) argue that the contracting institutions described above are less important for long-run growth, investment and financial development than institutions protecting property from expropriation by the state or powerful elites. They maintain that entrepreneurs can change their businesses or the terms of their contracts to adapt to weak contract enforcement, but cannot protect themselves against predatory elites or governments by writing new contracts. A state capable of protecting property rights, requires sufficient, consolidated power over the means of violence to prevent disorder and instability and to defend against external threats (North et al., 2009), yet it must be constrained from using its powers arbitrarily to confiscate property and infringe on rights.

NIE theories about the development of institutions protecting property against expropriation draw on European history, particularly the characteristics that made the Netherlands and England the loci of the Industrial Revolution (see, for example, Acemoglu & Robinson, 2012; Koyama & Rubin, 2022; North, 1990; North & Thomas, 1973; North et al., 2009). Acemoglu and Robinson (2019) and Greif (2006) stress state institutions inherited from the Roman Empire combined with participatory norms inherited from Germanic tribes. Koyama and Rubin (2022) note that the political fragmentation and conflict arising from Europe's fractured geography fed the separation of sovereignty between rulers, the Church, and the nobility, which served as a "precondition for the emergence of parliaments, independent cities, and other representative institutions" (p. 130) and contributed to "more interstate competition, which gave impetus to economic, political, scientific, and technological breakthroughs" (p. 134).

In the Netherlands and England, ruling groups were weaker compared to other European states vis-à-vis rising economic groups, such as the merchant classes. The Netherlands developed representative institutions early, thanks to its central location and excellent ports that facilitated flourishing trade and capital markets (Shirley, 2008), while development along its many waterways and ports contributed to a highly decentralized urbanization (de Vries & van der Woude, 1997). The Protestant Reformation also played a role by motivating and legitimizing the Dutch rebellion against the Catholic Habsburgs (Koyama & Rubin, 2022, p. 141) and the many clergy and nobles who remained loyal to the Spanish Crown. These traditional elites lost much of their influence after the revolt against Spain, opening the way for an increasingly powerful merchant elite. The Dutch also developed norms of greater local independence and private property early on to entice farmers to colonize its waterlogged areas, motivate local communities to invest in meeting their hydrologic needs, and convince private investors to sink large amounts of capital into reclaiming land from the sea (North & Thomas, 1973, p. 143; Schama, 1988, pp. 39–40). But the Netherlands' early progress faltered. Despite a flourishing commercial center with low transaction costs and a strong capital market, the Netherlands began to decline in the late seventeenth century as its commercial rivals such as England and France erected tariff barriers and waged wars. Simultaneously, the powerful Dutch

guilds locked in high wages and opposed labor saving innovations, while wealthy elites revived long-standing restrictions on internal commerce and pushed for inefficient restrictions on competition (such as a monopoly on all navigational charts) (Shirley, 2008, pp. 29–30).

Koyama and Rubin (2022, pp. 164–165) argue that there were many reasons why England, unlike the Netherlands, was able to sustain the Industrial Revolution, including its size (England was large enough to resist its commercial rivals, such as France), its larger skilled workforce with an apprentice program not tied to guilds, and the fact that its political elites who, although mercantilist, did not repress competition and trade. As in the Netherlands, English rulers were weaker than the wealthy monarchs in France or Spain, tipping the balance of power toward a rising merchant class, and the Protestant Reformation reduced the powers of the Catholic Church. The English monarchy increasingly turned to Parliament as a source of legitimacy (Koyama & Rubin, 2022, p. 143–144). North and Weingast (1989) credit the Glorious Revolution in 1688 with ushering in a change in England's institutions by giving Parliament a larger role in fiscal decisions, making it harder for the king to call or dissolve Parliament at will, and curtailing royal prerogatives and subordinating them to common law. These concessions gradually evolved into new institutions, such as common law courts, that further constrained state coercion and made property rights more secure.

Ultimately England evolved an open access society, which is characteristic of most highly developed countries today. According to North, Wallis and Weingast (2009), open access institutions (what Acemoglu & Robinson, 2012, 2019 call “inclusive institutions”) allow entry to a large and varied set of organizations (unions, business associations, political parties) that coordinate and articulate the interests of diverse groups of people and protect their rights against the state or powerful elites. Such institutions include rule of law; a free press; constitutional guarantees of freedoms of expression and rights to assemble; and judicial and bureaucratic mechanisms to enforce these rights. Among the mechanisms that work to curb state abuses, hold politicians accountable and promote policies to benefit broad social interests are elections, federalism, and constitutional curbs on executive power such as separation of powers with independent parliaments and judiciaries (Ménard & Shirley, 2022, pp. 95–96).

The most developed countries today are those that evolved such mechanisms, while simultaneously growing states strong enough to enforce contracts, protect property rights and assure stability and peace. They also advanced the contractual institutions discussed earlier: mechanisms and cooperative norms to support expanding exchange among strangers and bargains among competing interest groups and growing investment in ever more specialized skills and assets. Their institutions keep down transaction costs and curb the ascendancy of any single interest group, creating widespread opportunities for employment and consumption and encouraging investment in human capital. Increasing returns reinforce incentives in these societies to refine and strengthen the institutions that made these developments possible, except where “unanticipated consequences of

choices, external effects, and sometimes [exogenous] forces” alter the path (North, 1990).¹²

As mentioned, only a few countries have been able to develop and sustain both types of beneficial institutions described above. There has been improvement in some countries regarding contractual institutions, but it has proved far more difficult to create or sustain strong institutions to protect persons and property from expropriation. The next section surveys the literature that tries to explain why.

29.3 What Explains Why Countries Fail to Create the Rules and Norms that Foster Growth and Social Progress?

New Institutional Economics has made progress in identifying plausible explanations for weak institutions, but does not yet offer a satisfactory general explanation. One common theme is path dependency, the idea that historical choices and network effects make change very costly. As North (1990) described it, when a new institution is formed, organizations are created to take advantage of it, leading to informal constraints and modifications to formal rules that eventually produce an interconnected web resistant to change (see also Ménard & Shirley, 2022, p. 133).¹³ While path dependency can explain persistence, it cannot explain how countries get locked into good or bad paths at the outset. There are four prominent explanations in the NIE literature:

- (i) Elite dominance—elites use their power to maintain the institutions that endowed them with power in the first place;
- (ii) Colonial heritage—countries inherited poor institutions from their colonial masters;
- (iii) Colonial heritage plus factor endowments—countries had valuable resources, people that could be enslaved, or land suitable to plantation agriculture, enticing colonizers to design institutions to exploit these endowments, while other factors—dangerous diseases or hostile inhabitants—deterred settlers who might have imported better institutions;
- (iv) Beliefs and culture—shared societal beliefs and values lead people to resist changing damaging institutions, which then become embodied in norms that are inhospitable to markets or engender mistrust, preventing societies from building more open and inclusive institutions.

¹²England and the Low Countries were early examples of these developments in Western Europe, for example.

¹³The typical example of path dependency is the QWERTY keyboard, which “persisted in the face of more efficient alternatives because of large setup costs, learning effects, coordination effects (advantages to cooperation with other economic agents taking similar action) and adaptive expectations (when prevalence on the market enhances beliefs of further prevalence)” (Ménard & Shirley, 2022, p. 133). See David, 1985; Arthur, 1988, cited in North, 1990, p. 93.

Table 29.1 Summary of explanations for underdeveloped institutions

Explanations	Summary	Authors	Drawbacks
Elite Dominance	Elites created and sustained institutions to empower them further and prevent non-elites from gaining power	Acemoglu et al., 2020; Alston et al., 2018; North, 1981, 1990; North et al., 2009	Proximate, not ultimate cause? Many counterfactuals
Colonial Heritage	Countries inherited poor institutions from colonial masters.	North, 1990	Countries with same colonial master but different outcomes
	Common or civil law origins affect contemporary institutions.	La Porta et al., 1997, 1998	Common law has no growth effect. Historical anomalies. Large differences in enforcement
Colonial Heritage Plus Factor Endowments	Countries had valuable resources/people and colonizers designed institutions to exploit them. Interaction institutions & initial conditions created institutions that led to persistent inequality	Acemoglu et al. 2001, 2002; Acemoglu & Robinson, 2012; Engerman & Sokoloff 2002	Ignores differences among colonial powers. No test of casual effects of institutions. Many counterfactuals
Beliefs and Culture	Countries had beliefs and/or cultures inhospitable to markets or trust that prevented them from building institutions to encourage trade & investment	Alesina & Giuliano, 2015; Aoki, 2007; Greif, 1994, 2006; Greif & Laitin, 2004; Greif & Mokyr, 2017; North, 2005; Nunn & Wantchekon, 2011	Hard to falsify. Leaves little room for reform. Hard to distinguish between informal institutions (norms) and culture

Table 29.1 gives an overview of some of the studies in these categories. Despite the apparent distinctions between these explanations, it should be stressed that most authors do not describe them as discrete causes of poverty, but as combined obstacles to development.

Below, I consider how each of these arguments is plausible and supported by evidence documenting their importance at certain places and times. However, an important and open question for future research is whether and how much these institutions can account for the differences between countries that remain mired in poverty; countries that reduce poverty but do not become rich; and countries that are rich with high standards of living, education, and health.

29.3.1 Elite Dominance

The idea that elites impose institutions that benefit them and exclude non-elites from power is an enduring explanation for underdevelopment in NIE (for example,

Acemoglu & Robinson, 2012; Alston et al., 2018; Engerman & Sokolof, 2002; North, 1990; North et al., 2009). Indeed, North, Wallis and Weingast (2009) suggest that the “limited access” order that sustains elite dominance is the “natural state” of human societies. The “open access” orders that emerged in the nineteenth century and are now hallmarks of development still govern only 14% of the world’s population. Acemoglu et al. (2020, p. 3) summarize why elite dominance persists: “groups that are empowered by current institutions benefit from these institutions and thus use their power in order to maintain them, in the process reproducing their own power over the future institutions.” Although the identity of the elites may change, these “extractive” institutions persist. Even when the old order is overthrown by rebellion, the new rulers often choose to continue to use repressive institutions to exploit non-elites, rather than risk opening access to power more broadly. The persistence of autocratic institutions in former colonies after they are freed from their colonial masters (discussed below) is a case in point.

29.3.2 Colonial Heritage

Sometime during the last 600 years most of the countries that we call underdeveloped today were colonies, prompting some institutional economists to conclude that poor institutions are a colonial legacy. Since some richer countries were also colonized (USA, Australia, Canada, and New Zealand, for example), a stint as a colony is not necessarily inimical to institutional development. What features of colonial heritage might cause institutional failings?

North (1990) suggests that colonial powers created institutions that mirrored their own. Spain transplanted its centralized government, a large and interventionist bureaucracy and the hegemonic property rights of favored nobility to Latin America, while England brought its decentralized, limited government to its colonies in the New World. As a result, the United States and Canada were better positioned to curb state power, create more competitive markets, and industrialize faster than Latin America. But this explanation fails to explain why the English heritage failed to benefit its former colonies in Africa, the Caribbean, or South Asia or why Spain and England converged over time to a greater extent than their former colonies.¹⁴

La Porta and co-authors (1997, 1998) argue that a specific aspect of colonial heritage—a common law versus civil law system—has a profound effect on a country’s current institutions. Simplifying, common law combines laws passed by legislators with customary rules and judicial precedents while civil law requires judges to uphold laws received from the legislature and comply with jurisprudence created by higher courts (see Ménard & Shirley, 2022, Box 6.2). In the view of these authors, colonies that inherited common law systems from England developed institutions that were better at limiting the state’s power, enforcing contracts and

¹⁴Location may be one reason for convergence. Stimulus from the rest of Europe played a key role in the development of England and the Netherlands in the Middle Ages (North, 2005), and benefited Spain in the twentieth century.

protecting property rights, especially the rights of minority investors. Countries with civil law origins, particularly French civil law, developed a state more prone to threatening property rights, establishing monopolies, squelching innovation, and providing less protection for minority shareholders. Beck and Levine argue in ► [Chap. 16](#) of this Handbook that civil law legal origins are also correlated with underdeveloped financial systems, which is important since financial system development is highly correlated with growth. Subsequently, La Porta, Lopez-de-Silanes, and Shleifer (2008) broadened the meaning of legal origin to encompass the “style of social control of economic life,” where “common law stands for the strategy of social control that seeks to support private market outcomes, whereas civil law seeks to replace such outcomes with state-desired allocations” (p. 287).

Many studies contest the idea that historical legal origins determine current institutions such as those protecting property rights (Arrunada & Andonova, 2005; Hadfield and Bernier’s ► [Chap. 12](#) in this Handbook; Shirley, 2008). Pistor et al. (2000) finds large differences in current commercial laws and enforcement in countries with similar legal traditions. Siems (2022) notes that the simple taxonomy of legal families as largely common or civil law ignores the variety of influences on legal systems, including indigenous legal systems in, for example, Africa and Latin America.¹⁵ Importantly for the issues of development, the legal origin’s theories do not explain why the villains in this story, the colonial powers that brought civil laws to their colonies, were able to overcome their own legal origins and develop when many of their former colonies could not. Berkowitz et al. (2003) suggest that how a legal system was initially received—whether through conquest, colonization, or imitation—and how well it was adapted to local conditions may have more influence on how it functions today than whether it is French, German, British, or Scandinavian.

29.3.3 Colonial Heritage Plus Factor Endowments

Some scholars argue that the kinds of institutions Europeans imposed in their colonies depended on the conditions they found there. Acemoglu and co-authors (2001, 2002) contend that where Europeans colonizers faced high mortality risks because of dangerous diseases (such as malaria or yellow fever in Africa) and where the land was suited for plantation agriculture and there were large populations to exploit (such as Latin America and India), they created or adapted oppressive production methods and tax and tribute systems designed to extract resources from the rest of the population. In safer places where the population was relatively sparse and the land was less suited for plantations, such as the Northern US, Canada, or

¹⁵The legal origins literature ignores other aspects of history as well. Grajzl and Murrell’s computations (2022) of topics in preindustrial English case law shows that English financial law did not develop in the common law courts, but in the equity “whose formal procedures were perhaps closer to the civil law systems than to those of the common law courts. . .” (p. 224).

Hong Kong, Europeans settled in larger numbers, bringing beneficial institutions that supported private property and wider participation. When industrialization began in the nineteenth century, their reasoning goes that formerly rich colonies burdened with “extractive institutions” lacked secure property rights and provided few opportunities for technological and entrepreneurial innovation. The better institutions inherited by formerly poor colonies allowed them to industrialize and grow more rapidly.¹⁶

Acemoglu and Robinson’s explanation ignores the differences between the colonial powers’ home institutions that North describes. In their model, factor endowments and diseases determine the institutions that the colonial powers imposed, rather than the institutions prevalent in their home countries.¹⁷ In effect, given its factor endowments, the US would have developed as it did even if it had been colonized by Spain instead of England. Moreover, the authors explain underdevelopment as the result of “extractive colonial institutions,” but they do not directly test the causal relationship between growth and colonial institutions. Nor do they detail which “extractive colonial institutions” cripple societies so much that they stay poor for centuries. Most of Africa, for example, shares a low level of institutional and economic development despite differences in factor endowments, generally low population density, and other variations in colonial legacy. Dangerous diseases that curbed European settlement are the reason Acemoglu and co-authors give for Africa’s general state of underdevelopment. They point to the differences between the rest of Africa and those safer colonies where more Europeans settled because the risk of diseases was less: Rhodesia (now Zimbabwe), South Africa, and Kenya. But Europeans intensively settled some places such as the West Indies despite high mortality rates from dangerous diseases (Engerman & Sokoloff, 2002). Also, Kenya has not done as well as the US or much of East Asia, and Zimbabwe and South Africa have experienced recent economic declines despite their presumably superior institutional heritage.

Engerman and Sokoloff (2002) also argue that factor endowments explain large differences in contemporary institutions between countries settled by the same colonial power: for example, the United States and Jamaica or the northern and southern United States. In their view, institutions are endogenous to the conditions the colonists found when they arrived. Where soils, climate, and the size or density of the native population encouraged plantation agriculture with slaves, elites were able to establish institutions that ensured their ascendancy, contributing to persistently high degrees of inequality (Engerman and Sokoloff, 2002). The pernicious

¹⁶ Among former colonies, a 10% higher population density in 1500 is associated with a 4% lower per capita income today (Acemoglu et al., 2002). Countries which were never colonized by Western Europe did not experience this “reversal of fortune” according to Acemoglu and co-authors (2002). Rodrik et al. (2002), however, find that institutional quality today among never colonized countries is as widely dispersed as among former colonies.

¹⁷ The authors found that adding the identity of the colonial power to the estimate has little effect on the results, perhaps because their instrument (mortality rates of colonial settlers) captures the exogenous sources of variation in institutional quality.

influence of slavery—slaves from Africa made up 60% of the more than six million people who migrated to the New World from 1500 to the end of the eighteenth century—contributed to persistent disparities in wealth, human capital, and political power. Engerman and Sokoloff (2002) specify the institutions that evolve where initial conditions favored equality and homogeneity, including rules that encouraged immigration, expanded the franchise, promoted secure, and cheap land acquisition and increased access to schooling and banking. Engerman and Sokoloff's explanation seems, superficially at least, to be contradicted by experience outside the Americas, especially the relatively higher income equality in some underdeveloped former colonies in Asia, and African countries' similar institutional failures across different factor endowments.

29.3.4 Beliefs and Culture

For many institutionalists, shared societal beliefs and/or cultures explain why some countries developed better institutions than others. For example, North (2005, p. 2) argues that over time a society's "dominant beliefs—those of political and economic entrepreneurs in a position to make policies—result in the accretion of an elaborate structure of institutions that determine economic and political performance." In his view changes in beliefs do not drive institutional changes; rather, some beliefs allow a learning process that permits beliefs to evolve as institutions change and influence which sorts of institutional changes will be accepted (see also Alston et al. 2018, ► Chap. 8). For example, North argues that Christian dogma evolved in ways that supported economic growth and technological innovation in Western Europe during the Middle Ages. Thus, the Christian view that nature should serve mankind supported technological improvements in agriculture, while the Protestant Reformation changed beliefs in support of greater individual freedom and economic expansion in the Netherlands and England.

Avner Greif (1994) illustrates how beliefs affect institutional development by contrasting the individualistic cultural beliefs of the Christian Genoese and the collectivist cultural beliefs of the Jewish Maghribi traders mentioned earlier. As we have seen, the Maghribis' horizontal social structure relied on partnerships, community ties, and "formal friendships" among cooperating traders to enforce bargains. Genoa, however, had a vertical social structure, and its merchants evolved bills of lading, written contracts, laws, and permanent courts to support bargains among traders. The Maghribis did not need written documents and courts to enforce bargains; fears of losing reputation and ostracism worked just as well within their collectivist system. Ultimately, however, the Maghribis' failure to develop formal contracts and laws enforced by courts confined their trade to their network, where their collectivist enforcement was effective, while the Genoese grew rich through extensive and expanding trade.

Some institutionalists view beliefs as the underlying basis for institutions (see for example, Aoki, 2007; Greif, 2006; Greif & Laitin, 2004; Greif & Mokyr, 2017). As Aoki (2007, p. 6) explains, "agents' shared beliefs about how the game is played and

to be played” result in “self-sustaining patterns of social interactions,” in effect, institutions. Thus, people’s shared beliefs about the rules of the game lead them to shared expectations of how others will behave and how others will expect them to behave, thereby generating the regularity of behavior that equates to an institution. This pattern is self-reinforcing; the observed behavior confirms each individual’s belief that others will behave in a certain manner, and “given those beliefs it is optional for each individual to do so” (Greif, 2006, pp. 158–159). Beliefs change when a buildup of contrary behaviors and circumstances destabilizes shared beliefs and upsets the equilibrium. For example, in early modern Europe according to Greif and Mokyr (2017, p. 39), “the new discoveries and instruments emerging after 1500 showed the many errors of classical learning and raised skepticism and contestability about age-old accepted beliefs. . . . Evidence and logic replaced unassailable authority.” Cumulatively, these changes eventually contributed to the Enlightenment’s “belief in progress, in the capability of economic agents to work successfully towards improving their lives” (see also Mokyr, 2016; Mokyr ► Chap. 34 in this Handbook).

Increasingly, culture has become another explanation for institutional persistence (see Ménard & Shirley, 2022, pp. 137–142). Culture is usually defined in this literature as “those customary beliefs and values that ethnic, religious, and social groups transmit fairly unchanged from generation to generation” (Guiso et al., 2006).¹⁸ According to Alesina and Giuliano (2015, p. 905), the most studied aspect of culture is trust. Studies find that trust in people who are not necessarily one’s family or friends correlates strongly with growth and development (for example, Algan & Cahuc, 2010, 2014; Knack & Keefer, 1997). Higher levels of trust are associated with stronger financial market development, greater innovation, more efficient labor relations, and other beneficial measures (Algan & Cahuc, 2014). Knack and Keefer (1997) shows that trust is also correlated with private investment, perhaps because it reduces the transaction costs of securing agreements.

Anthropological research suggests that culture is a consequence of evolution: humans learn skills, tools, and behaviors from one another that enable them to survive better as a group than as individuals (Henrich, 2016). As humans agglomerated into ever larger societies, members became more cooperative, less distrustful of strangers, and more individualistic; they also incorporated these traits into informal and eventually formal institutions (Henrich, 2020). Over time, individuals in such societies internalized “market norms” which “establish the standards for judging oneself and others in impersonal transactions and lead to the internalization of motivations for trust, fairness, and cooperation with strangers and anonymous others” (Ibid, p. 293).

¹⁸For some scholars’ culture includes norms and conventions, which North considered institutions. Mokyr’s (2016) distinction is helpful: culture is “something *entirely of the mind*, which can differ from individual to individual and is, to an extent, a matter of individual *choice*. Institutions are socially determined conditional incentives and consequences to actions. These incentives are parametrically given to every individual and are beyond their control” (p. 9, italics in original).

Beliefs and cultures seem to be deeply ingrained and the product of intractable factors, such as a society's ethnic, religious, or linguistic heterogeneity or events in its history. For example, Nunn (2008) finds that slavery had a significant negative effect on African development, which Nunn and Wantchekon (2011) attribute largely to a culture of mistrust that persists today, especially among those whose ancestors were most heavily enslaved. While some slaves were captured through raids or in warfare, many were kidnapped or tricked into slavery by members of their own ethnic group, community, even family, and friends. The slave trade also weakened the social and political institutions that might have constrained individuals to act in a more trustworthy manner, reinforcing a sense of insecurity and mistrust.

29.3.5 All of the Above

Despite disagreement on the ultimate determinants of institutional development, the explanations summarized above broadly agree on the proximate causes: (i) greater equality combined with (ii) sufficient political competition to limit the ability of rulers to expropriate. Authoritarian regimes, where a consolidated, wealthy, and despotic ruling group exploited a poor or enslaved workforce might have been successful in the past, when organizing plantation agriculture or mining was key to economic success, but their oppressive institutions were inimical to competition, specialization, and industrialization. Acemoglu and Robinson (2012, p. 150) point out that the growth experienced under extractive institutions is ultimately unsustainable since, by their nature, they generate limited amounts of technological progress and create strong incentives for infighting and instability. They cite the experience of the Soviet Union, where, despite rapid catch up growth at its beginnings, distorted incentives ultimately stifled innovation and productivity.

Another example of the damage caused by extractive institutions is China. Huang (2023) argues that China's "deferential culture" was entrenched by its long legacy of competitive exams (Keju) that inculcated the statist values of Confucianism (p. 111) and diverted the best human capital into the services of the state while "decimating or preempting collective action—the essential feature of a civil society" (p. xi). The "marginalization of society, ideological homogeneity, and stifling of intellectuals by the state" (p. 328) inherited from China's imperial past were coupled with modern authoritarian institutions imposed by the Chinese Communist Party. While these repressive institutions were briefly eased during China's reformist era in the 1980s, China never developed strong protections against state expropriation. Instead, there has been in Huang's view, a "regression to the autocratic mean," especially since the elimination of presidential term limits in 2018. As of this writing, the abrupt expropriation of private firms and jailing of prominent business leaders, the recentralization of power in the central government, the increasing stress on ideological conformity and the unpredictable risk of government-mandated shut downs have contributed to a decline in private investment and a rise of interlocking ownership between private firms and (on average less productive) state-owned enterprises (see Shirley and Xi ► [Chap. 35](#) in this Handbook).

The institutions that protect property rights and support strong market economies in Western Europe emerged gradually from a long and disorderly process of adaptation and experimentation spurred by competition and wars. Perhaps this organic progress toward efficiency would have happened more widely but was interrupted by colonialism's transplant of institutions that were less well-adapted to local norms, beliefs, and environments (Djankov et al., 2002). Or it could be that the fortuitous circumstances that produced a supportive institutional environment in today's developed countries were simply missing in many other places. More research is needed to sort out the role of elites, colonialism, factor endowments, knowledge, historical accidents, and the evolution of supportive beliefs and cultures.

Nevertheless, thanks to the literature surveyed above, we are closer to understanding underdevelopment than ever before. Studies that look for distant determinants of institutional quality, however, tell us little about how to foster the institutions necessary for a country to develop today, the subject of the next section.

29.4 How Can Developing Countries Improve Their Institutions?

The NIE has had less to say about institutional reform, except that it is hard to accomplish. North's work suggests that a great deal of change occurs constantly at the margin, but the underlying institutional framework is typically stable, except when change is imposed by force or revolution. This stability is the product of path dependency and raises the obvious question: How do institutions ever change?

One possibility is through conflict among the powerful elites who usually resist change. As North et al. (2009) point out, elites are not unified, "but are composed of disparate groups that compete and cooperate and sometimes go to war against each other. (p. 149). They are usually not unified enough to decide to share power, but their disunity can open an opportunity for change. Rival groups can sometimes push for more representative institutions that gradually or abruptly open the way for a more open social order. Change is more likely when a rival coalition of elites is empowered by a "critical juncture" or shock, especially if these circumstances coincide with changing beliefs and the accumulation of many smaller institutional innovations.¹⁹ In England, according to Acemoglu and Robinson (2012, p. 209), there was a "slow incremental drift of power away from elites to citizens more generally." In their view the rise of Atlantic trade was a critical juncture that empowered and enriched a new class of merchants and businessmen in England.

¹⁹According to Acemoglu and Robinson (2012, p. 101), a critical juncture is a "major event or confluence of factors disrupting the existing economic or political balance in society" that can open the way to more inclusive institutions or intensify the emergence of extractive institutions. They give the example of the Black Death, the plague in the mid-fourteenth century. Plague led to a scarcity of labor that encouraged peasants in England to successfully free themselves from compulsory labor services, but in Eastern Europe landlords took over large tracts of land and increased their control over the labor force, ushering in a Second Serfdom (pp. 97–101).

Atlantic trade did not become a critical juncture in France, Spain, or Portugal because their kings were better able to use trading profits to solidify the power of the monarchy. But in England the crown was weaker and the Atlantic traders formed a broad coalition with the gentry, commercial farmers, and some manufacturers, such as wool producers, to oppose the king's efforts at absolutism and monopoly.

In some circumstances elite control is challenged by the “disempowered majority” that occasionally overcomes their collective action problems, take power, and impose major institutional change (Acemoglu et al., 2020). For example, during the French Revolution of 1789, the peasant and urban poor united with the bourgeoisie to destroy the old institutions. Despite subsequent turmoil, the Napoleonic dictatorship and the Restoration, the French Revolution still partly destroyed important feudal institutions and aristocratic privileges (Vahabi et al., 2020).

Elites may also respond to pressure from non-elites by undertaking institutional reform to forestall worse outcomes, which Acemoglu and Robinson (2000) argue was the reason why some countries (e.g., England, Germany, and Sweden) extended the franchise in the late nineteenth and early twentieth century. However, in other cases, elites respond to non-elite pressure by increasing repression, such as the French response to the Haitian Revolution in the late eighteenth century. History is replete with many examples where the institutions that empower elites and exclude non-elites persisted unchanged for centuries, and the explanations for why these institutions sometimes change seem hard to generalize. Moreover, even if the formal, written laws and rules are suddenly altered by conflict, revolution, or invasion, unless norms, beliefs, and cultures also change, the new status quo will be overturned after the conflict abates, the revolution ends, or the invaders leave.

What can a society do if its informal norms, beliefs, or culture are inhospitable? Keefer and Knack (2005) suggest two factors that increase development-promoting values: income equality and education. They cite studies of East Asia suggesting that the relatively high levels of education and income equality help explain why East Asian countries have grown faster and produced better social welfare measures than other less developed regions. However, if North is correct that beliefs filter the information that people derive from experience then it will be hard for schooling alone to change beliefs. Moreover, schooling may sometimes have the opposite effect. Huang (2023) argues that centuries of lengthy training for China's centralized exams internalized an attitude of unquestioning subordination to the emperor among the elite that still influences Chinese society today. Some of the more effective schools in developing countries are those structured to inculcate current dogma (for example, religious schools that teach fanaticism). Public schools are themselves captives of their institutional environments and are as effective or ineffective as the institutional environment would lead you to expect. For example, World Bank studies (cited in 2018, Fig. 3.10) reported significant losses of instruction time due to teacher absences. “Unannounced visits to primary schools in six countries found that in public schools, on average, about one teacher in five was absent on a typical school day” (Ibid, p. 181). Moreover, “globally, 125 million children are not acquiring functional literacy or numeracy, even after spending at least four years in school” (Ibid, p. 71). When and how education and income distribution interact to

alter beliefs and cultures in ways that foster development is not well-understood and deserves further study.

Yet education and new ideas have been a powerful impetus for institutional reform when the setting is receptive, as with the intellectual movement known as Enlightenment in Western Europe. Mokyr (2016) has argued that the set of cultural beliefs in human progress and especially in the power of useful knowledge to transform the economy that characterized the Enlightenment “uniquely marked Europe to become the locus of economic modernity” (p. 267). There are more recent examples of the transformative power of ideas as well, although in a narrower sense than the Enlightenment. For example, Shirley (2008, Chap. 7) emphasizes the pivotal role of a critical mass of well-trained scholars working in universities and think tanks in structural changes in Chile, South Korea, and Taiwan. These scholars were able to influence changes in policies, paradigms, and institutions when a crisis made change seem necessary, even essential, to the policy makers and when the local scholars were perceived as experts, had access to powerful actors and presented a consensus plan based on an alternative paradigm. She also provides counterexamples—Argentina and Indonesia—where the preconditions were not met, and the reforms failed.

A large literature describes the historical transformations in institutions that fostered the emergence of modern, developed economies in Europe, North America, and Asia, but there have been fewer studies explaining institutional changes in modern economies. Of particular interest is China. China’s spectacular transformation is described by Qian (2002) as “puzzling” given the weakness of rule of law and secure property rights, elements that the literature I just described views as necessary for development through private initiatives. As mentioned, China since the death of Mao in 1976 developed a variety of institutions supporting private contracts and reducing transaction costs, including: personal relations, clan, and family ties; firm level networks, trust, and reputation; business associations and industrial clusters; and, increasingly, formal legal protections enforced by progressively more independent and fairer courts. But China has been less successful in evolving institutions to protect private entrepreneurs against expropriation by government or elites. Personal relations with local governments have been the most important safeguard, but Shirley and Xu report in ► Chap. 35 that the current anti-corruption campaign has had a chilling effect on these relationships. To protect their investments, some private entrepreneurs have joined the Communist Party or political bodies, and many private firms have initiated or acquiesced to equity investments by state-owned enterprises as a form of security against expropriation and also to access credit from the state-controlled banking system. A recent slowdown in private entry and investment may signal the limitations of these forms of protection compared to the robust checks on state expropriation possible in an open access system. Acemoglu and Robinson (2012, p. 443) argue that China’s dramatic growth so far has been “catch up” growth made possible thanks to a “radical shift away from extractive economic institutions.” They predict that China’s “authoritarian led growth” will “run out of steam unless extractive political institutions make way for inclusive institutions” (p. 441).

Given how quickly NIE has evolved from a time when institutions were not even included in most development models, the gap in our understanding of development is not surprising. Foreign assistance agencies have entered this gap under the assumption that institutions can be changed by outside advice and funds with mixed results as described in the next section.

29.5 Can Outsiders Promote Institutional Development?

Outsiders have usually changed deeply rooted institutions by fomenting revolutions or invading, sometimes in consort with powerful local reformers.²⁰ For example, Napoleon's armies imposed radical institutional changes in occupied states, including the introduction of a civil legal code and equality before the law, the abolition of feudalism and guilds, and the elimination of aristocratic privileges (Acemoglu et al., 2011). In some cases, these interventions brought enduring changes to Europe's legal, political, and social institutions in relatively short periods of occupation. Force alone cannot explain Napoleon's enduring impact. Some intellectuals and merchants were receptive to Napoleon's innovations. Dissatisfied with domestic institutions and inspired by the Enlightenment, they saw his reforms as progressive, the heritage of the French Revolution, or as advantageous, opening the way for new entrants and businesses. Acemoglu et al. (2011) assert that places where this new elite was able to hang on to power after Napoleon's defeat industrialized more rapidly than where the *ancien régime* was restored.

Outsiders have also contributed to enduring institutional change in countries where powerful elites welcomed foreign ideas, such as Tsar Peter in Russia or Mustafa Kemal Atatürk in Turkey. Absent such powerful local support, can aid or advice foster enduring improvements in another country's embedded institutions? Some observers believe that aid cannot spur and may even slow institutional change by preventing political competition, propping up local elites who might otherwise have been removed, promoting rent seeking or reducing the need for governments to appeal to their citizens for tax revenues and to account for their spending (for example, Bates, 2001; Bräutigam & Knack, 2004; Dijkstra, 2018; Svensson, 2000).

Some have argued that aid can actually undermine the sustainability of the institutional reforms it is designed to support by promoting rent seeking and shirking. Buchanan termed this problem the Samaritan's dilemma (Buchanan, 1977). The payoff is highest to the Samaritan if the Samaritan provides aid and the beneficiary responds by exerting high effort. But the payoff is highest to the beneficiary if he or she can receive the aid without increasing effort. The weaker a country's institutional framework, the more likely it is that this is the game being

²⁰This section addresses whether outsiders have been able to promote sustainable improvements in the institutional environment by changing, e.g., constitutions and laws, norms and cultures, enforcement mechanisms, etc. It does not address outsiders' arguably more successful record in influencing changes in less embedded institutions, such as the regulations governing electricity.

played. “When the recipient country is governed by officials who are primarily interested in seeking out opportunities for private gain, and few institutions are in place to keep these motivations in check, moral hazard problems can become substantial” (Ostrom et al., 2002, p. 11). Moral hazard problems are exacerbated when the goal is institutional change because projects directed at changing institutions lack tangible outputs, making impact “more diffuse and hard to verify” (Martens et al., 2002, p. 17).

Surveys of the empirical evidence on aid and institutions (Dietrich & Winters, 2021; Dijkstra, 2018) found that aid had negative effects on the quality of government in earlier studies, but more recent research suggests that these negative effects are overblown. In studies of aid since the end of the Cold War, the effects on measures such as corruption, democracy, constraints on the executive, checks, and balances, etc., are sometimes positive. However, most such studies suffer serious identification problems. Not only can it be difficult to establish causal effects, since aid is one of many factors affecting institutions (and a small one at that), there are often omitted variables and measures of institutions suffer well-known flaws (Dijkstra, 2018). Dietrich and Winters (2021) also suggest that some studies of the effects of aid on quality of government suffer from reverse causality and rely on instrumental variables that cannot be excluded as independent causes of changes in the quality of government.

More fundamentally, there are reasons—summarized in Table 29.2—to question the very premise that aid can positively alter the underlying institutional framework in a meaningful way (see Shirley, 2008). First, the foreign aid community generally assumes that institutions are malleable and can be changed through aid within the 3–5 year life span of a development project, or at most the 15–20 year span of several projects. Aid projects try to reform institutions through conditionality, which are a

Table 29.2 Comparison of the characteristics of institutional frameworks and foreign aid

Institutional frameworks	Foreign aid
Deeply rooted & usually durable	Addresses short run changes in policies, organizations, formal rules Three year projects Staff rotates every 3–5 years Staff rewarded for approval, not sustainability No long run accountability
Status quo institutions supported by powerful elites & shared beliefs	Projects require support or permission of incumbent governments Staff incentive is to cooperate to win more project approvals Advocates of radical change would be asked to leave Beliefs are not changed by advice or money
Changes often arise locally from heterodox experiments	Focus on Western best practices Incentive is to push for changes that are Rapid Measurable against benchmarks

Based on Shirley (2008), Table 4.3. Used with permission

list of specific changes that the country must enact before funds will be disbursed. But projects and conditionality do not fit well with what is known about fundamental institutional change. The NIE literature described above suggests that institutional frameworks usually change as the result of a long and often painful process of competition and adaptation, extending far beyond the time frame of even a series of aid projects. Moreover, the incentives in aid agencies (described in Shirley, 2008) motivate staff to focus on approval and short-term accomplishments, not sustainability, and deeply rooted changes.

A second premise is that well-intentioned outsiders can discover needed institutional changes and persuade governments to implement reforms and sustain them. Aid agencies, however, cannot act in a country without the cooperation, or at least the acquiescence of powerful elites who control the government and will oppose any reforms likely to undermine their grip on power. Ruling elites often prefer pro forma changes so they can obtain funds without politically costly changes in deep-seated constitutional rules, norms, and beliefs—the Samaritan’s dilemma. Andrews et al. (2017) called this “isomorphic mimicry”: developing countries comply with aid conditionality by adopting the form of best practice but not the function, for example, by adopting a labor law but never enforcing it.

Moreover, the NIE literature suggests that changes are only sustained if shared belief systems change as well. Aid as presently constructed is a poor tool for changing the deep-seated beliefs or the norms and conventions that underlie many formal institutions. Sustained improvements in education, for example, depend on curriculum choice; rules governing teacher selection, salaries, and accountability; beliefs and norms about schooling (of girls, for example); and the like. These often politically sensitive and culturally bound elements are not likely to change because of conditionality and advice directed at central government ministries or incentives tied to financing for construction of schools, purchase of textbooks, or technical assistance.

Third, successful institutional adaptations have been engineered by insiders and sometimes work quite contrary to conventional wisdom or best practice touted by the aid community. It can even be hard for outsiders to recognize that a fundamental institutional change is taking place, at least in its early stages when it may not resemble reforms in other countries. As Shirley and Xu discuss in ► Chap. 35 in this handbook, the extraordinary transformation of China’s institutions that fostered its rapid, market-driven growth stemmed from heterodox, idiosyncratic reforms. The early changes after the death of Mao in 1976 were largely bottom-up, driven by a surge of private entrepreneurship unleashed with the removal of the more draconian restrictions on private activity and the opportunity to hide private activity in nominally publicly owned township and village enterprises (TVEs). These were difficult for outsiders to identify and support, especially multilateral agencies supporting Western best practice.

29.6 What Next?

Although progress has been remarkable, new institutional economics has not fully answered the questions I posed at the outset: Which institutions must function effectively if poorer countries are to converge with rich economies with high living

standards? Why have so few countries been able to create and sustain the rules and norms that foster growth and social progress? How can poorer countries attain well-functioning institutions? Can outsiders help them in this effort? Historical analyses have produced intriguing findings, but no single explanation is fully satisfactory; there are glaring gaps and some research ignores disturbing counterfactuals. Cross-country studies have put institutional variables into mainstream models and produced some consistent regularities, but the devil is in the details and the details can be numerous. While much is known about how institutions developed in Western Europe, and research on institutional development in third-world countries has increased considerably, there are still important gaps.

What can be done to fill the gaps in our understanding? Thanks to a new generation of cross-country studies coupled with increasingly detailed databases and advances in methodology, we are progressing in understanding how institutions affect development. For example, Acemoglu et al., 2019 used a new, consolidated measure of democracy and a dynamic panel strategy controlling for country fixed effects to show that, contrary to much previous research, democracy has a positive effect on GDP per capita. Also promising are data science techniques such as machine learning, text analysis, and sentiment analysis, which have pushed out the frontier of institutional analysis. New sorts of data can now be digitized and searched, such as medieval court decisions in England (see ► Chap. 39 by Grajzl and Murrell in this Handbook), and broader and more varied searches of datasets with billions of observations are possible, for example, data on taxpayers have been combined with test scores and teacher assignments from large urban school districts to study the income effects of better teachers (cited in Prüfer & Prüfer, 2018). New datasets have also been generated by mining information on the Internet and social media; for example, opinions based on Google searches and social media posts have been used to study factors affecting trends in racism (ibid.). Another promising development is the rise of studies of informal institutions, culture and beliefs (see citations in Alesina & Giuliano, 2015, and relevant chapters in Ménard & Shirley, 2018, 2022). This research promises to fill a gaping hole in our understanding, once troubling issues of classification and causality are resolved (on these issues see ► Chap. 34 by Mokyr in this Handbook).

It may be possible to expand our understanding with a pincer movement. Statistical analyses are already moving from aggregation to specificity; case studies can move from *sui generis* to comparative (see Skarbek, 2020). Case studies can be powerful tools when they are analytical narratives, cases that test logically consistent and falsifiable hypotheses with methodological rigor and also describe historical context, norms and beliefs, and institutional adaptations, all the rich nuances of the institutional setting (see Bates et al., 1998). Comparative analytical narratives—cases using a common methodology and common conceptual framework to assess a larger sample—would allow us to identify regularities with greater confidence. Although the task seems daunting, there have been examples (see Ostrom, 1990; Shirley, 2002; cases cited in Skarbek, 2020). Skarbek (2020) describes how comparative cases can be used to develop causal inferences using qualitative data. He also advocates process tracing: the systematic examination of the intermediate steps to determine how the process took place and generated the outcome of interest,

tracing causality through changes in independent variables through intervening variables to the dependent variable. He illustrates process tracing with studies of the social revolutions in France, Russia, and China and of the extortion by drug gangs in California prisons.

Deeper analysis of institutions **within** developing countries is another promising trend. The Spiller and Tommasi study of Argentina (2007) illustrates the explanatory power of analyses when the country-wide institutional framework can be held constant. Lack of reliable information can be a stumbling block to applying these tools in poorer countries, but lack of local researchers is often the more serious obstacle. In many developing countries, low pay, inadequate resources, and a sense of isolation drive the best scholars away from research or out of the country. Those who remain face an uphill battle getting funding to build databases, undertake serious research, and publish controversial findings.

A critical mass of local researchers is a prerequisite for understanding institutions fully, stimulating an informed debate, and fostering changes in belief systems, the first step to enduring institutional change. Since improvements in formal institutions hinge on changes in long-held beliefs, an important role for outsiders is to support this learning by helping build local knowledge and educational institutions while avoiding actions that fortify the defenders of the old order. Only when this minimum mass of human capital is in place will citizens of poorer countries begin to discover how to overcome fully the challenge of development.

Acknowledgement of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., & Johnson, S. (2005). Unbundling institutions. *Journal of Political Economy*, 113(5), 949–995.
- Acemoglu, D., & Robinson, J. A. (2000). Why did the west extend the franchise? Democracy and growth in historical perspective. *The Quarterly Journal of Economics*, 115(4), 1167–1199.
- Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Crown.
- Acemoglu, D., & Robinson, J. A. (2019). *The narrow corridor: States societies, and the fate of liberty*. Penguin Press.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5), 1369–1401.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2002). Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *The Quarterly Journal of Economics*, 117(4), 1231–1294.
- Acemoglu, D., Cantoni, D., Johnson, S., & Robinson, J. A. (2011). The consequences of radical reform: The French revolution. *American Economic Review*, 101, 3286–3307.
- Acemoglu, D., Egorov, G., & Sonin, K. (2020). *Institutional change and institutional persistence* (NBER Working paper #27482).
- Acemoglu, D., S. Naidu, P. Restrepo, J.A. Robinson (2019). Democracy does cause growth. *Journal of Political Economy*, 27(1), 47–100.
- Alchian, A. A. (1965). Some economics of property rights. *Il Politico*, 30(4), 816–829.

- Alesina, A., & Giuliano, P. (2015). Culture and institutions. *Journal of Economic Literature*, 53(4), 898–944.
- Algan, Y., & Cahuc, P. (2010). Inherited trust and growth. *American Economic Review*, 100, 2060–2092.
- Algan, Y., & Cahuc, P. (2014). Trust, well-being and growth: New evidence and policy implications. In P. Aghion & S. N. Durlauf (Eds.), *Handbook of economic growth* (pp. 49–120). Elsevier Science.
- Allen, D. W. (2000). Transaction costs. In B. Bouckaert & G. De Geest (Eds.), *Encyclopedia of law and economics, volume 1. The history and methodology of law and economics* (pp. 893–926). Edward Elgar.
- Alston, E., Alston, L. J., Mueller, B., & Nonnenmacher, T. (2018). *Institutional and organizational economics. Concepts and applications*. Cambridge University Press.
- Andrews, M., Prichett, L., & Woolcock, M. (2017). *Building state capability. Evidence, analysis, action*. Oxford University Press.
- Aoki, M. (2007). Endogenizing institutions and institutional changes. *Journal of Institutional Economics*, 3(1), 1–31.
- Arrunada, B., & Andonova, V. (2005). Market institutions and judicial rulemaking. In C. Ménard & M. M. Shirley (Eds.), *The handbook for new institutional economics* (pp. 229–250). Springer.
- Bates, R. H. (2001). *Prosperity and violence: The political economy of development*. W.W. Norton & Company.
- Bates, R. H., Greif, A., Levi, M., Rosenthal, J.-L., & Weingast, B. R. (1998). *Analytic narratives*. Princeton University Press.
- Berkowitz, D., Pistor, K., & Richard, J. F. (2003). Economic development, legality, and the transplant effect. *European Economic Review*, 47, 165–195.
- Bräutigam, D. A., & Knack, S. (2004). Foreign aid, institutions, and governance in sub-Saharan Africa. *Economic Development and Cultural Change*, 52(2), 255–285.
- Buchanan, J. M. (1977). The Samaritan's dilemma. In J. M. Buchanan (Ed.), *Freedom in constitutional contract*. Texas A&M University Press.
- Chemin, M. (2007). *Does judicial quality shape economic activity? Evidence from a judicial reform in India* (Centre Interuniversitaire sur le Risque, les Politiques Economiques et l'Emploi Working Paper 07-25).
- Coase, R. H. (1992). The economic structure of production. *American Economic Review*, 82(3), 713–719.
- Coase, R. H. (1998). The new institutional economics. *American Economic Review*, 88(2), 72–74.
- Dahlman, C. J. (1979). The problem of externality. *The Journal of Law and Economics*, 22, 1.
- David, P. (1985). CLIO and the economics of QUERTY. *American Economic Review*, 75(2), 332–337.
- De Vries, J., & van der Woude, A. (1997). *The first modern economy. Success, failure, and perseverance of the Dutch economy. 1500–1815*. Cambridge University Press.
- Diao, X., McMillan, M., & Rodrik, D. (2017). *The recent growth boom in developing economies: A structural-change perspective* (CEPR Discussion Paper No. DP11804).
- Dietrich, S., & Winters, M. S. (2021). Foreign aid and quality of government. In A. Bågenholm, M. Bauhr, M. Grimes, & B. Rothstein (Eds.), *The Oxford handbook of the quality of government* (pp. 449–471). Oxford University Press.
- Dijkstra, G. (2018). Aid and good governance: Examining aggregate unintended effects of aid. *Evaluation and Program Planning*, 68, 225–232.
- Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2002). *Appropriate institutions*. Paper presented at the World Bank Conference on Appropriate Institutions for Growth, Washington, September 13, 2002.
- Durlauf, S. N. (2018). *Institutions, development and growth: Where does the evidence stand?* (Economic Development and Institutions Working Paper #WPI8/041).
- Easterly, W. (2002). *The elusive quest for growth. Economists' adventures and misadventures in the tropics*. MIT Press.

- Engerman, S. L., & Sokoloff, K. L. (2002). Factor endowments, inequality, and paths of development among new world economies. *Economia*, 3, 41–109.
- Grajzl, P., & Murrell, P. (2022). Using topic modeling in legal history with an application to pre-industrial English case law on finance. *Law and History Review*, 40(2), 189–228.
- Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *American Economic Review*, 83(3), 525–548.
- Greif, A. (1994). Cultural beliefs and the Organization of Society: A historical and theoretical reflection on collectivist and individualist societies. *Journal of Political Economy*, 102(5), 912–950.
- Greif, A. (2006). *Institutions and the path to the modern economy. Lessons from medieval trade*. Cambridge University Press.
- Greif, A., & Laitin, D. D. (2004). A theory of endogenous institutional change. *American Political Science Review*, 98(4), 633–652.
- Greif, A., & Mokyr, J. (2017). Cognitive rules, institutions, and economic growth: Douglass North and beyond. *Journal of Institutional Economics*, 13(1), 25–52.
- Guiso, L., Sapienza, P., & Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20(2), 23–48.
- Hagemeyer, J., Michalek, J. J., & Svatko, P. (2021). Economic impact of the EU eastern enlargement on new member states revisited: The role of economic institutions. *Central European Economic Journal*, 8(55), 126–143.
- Henrich, J. (2016). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.
- Henrich, J. (2020). *The WEIRD people in the world. How the west became psychologically peculiar and particularly prosperous*. Farrar, Straus and Giroux.
- Huang, Y. (2023). *The rise and fall of the EAST: How exams, autocracy, stability, and technology brought China success, and why they might lead to its decline*. Yale University Press.
- Keefer, P., & Knack, S. (2005). Social capital, social norms and the new institutional economics. In C. Ménard & M. M. Shirley (Eds.), *Handbook for new institutional economics* (1st ed., pp. 701–726). Springer.
- Keefer, P., & Shirley, M. M. (2000). Formal versus informal institutions in economic development. In C. Ménard (Ed.), *Institutions, contracts, and organizations: Perspectives from new institutional economics* (pp. 88–107). Edward Elgar.
- Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics*, 112(4), 1251–1288.
- Koyama, M., & Rubin, J. (2022). *How the world became rich: The historical origins of economic growth*. Polity Press.
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. (1997). Legal determinants of external finance. *Journal of Finance*, 52, 1131–1150.
- La Porta, R., Lopez-de-Silanes, F., Shleifer, A., & Vishny, R. (1998). Law and finance. *Journal of Political Economy*, 106, 1113–1155.
- La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2008). The economic consequences of legal origins. *Journal of Economic Literature*, 46(2), 285–332.
- Martens, B., Mummert, U., Murrell, P., & Seabright, P. (2002). *The institutional economics of foreign aid*. Cambridge University Press.
- Ménard, C., & Shirley, M. M. (2018). *A research agenda for new institutional economics*. Edward Elgar.
- Ménard, C., & Shirley, M. M. (2022). *Advanced introduction to new institutional economics*. Edward Elgar.
- Milgrom, P., North, D., & Weingast, B. R. (1990). The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics and Politics*, 2, 1–24.
- Mokyr, J. (2016). *A culture of growth. The origins of the modern economy*. Princeton University Press.
- North, D. C. (1981). *Structure and change in economic history*. Norton.

- North, D. C. (1990). *Institutions, institutional change, and economic performance*. Cambridge University Press.
- North, D. C. (2005). *Understanding the process of economic change*. Princeton University Press.
- North, D. C., & Thomas, R. P. (1973). *The rise of the western world: A new economic history*. Cambridge University Press.
- North, D. C., & Weingast, B. R. (1989). Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century England. *The Journal of Economic History*, 49(4), 803–832.
- North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.
- Nunn, N. (2008). The long-term effects of Africa's slave trades. *Quarterly Journal of Economics*, 123(1), 139–176.
- Nunn, N., & Wantchekon, L. (2011). The slave trade and the origins of mistrust in Africa. *American Economic Review*, 101, 3221–3252.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
- Ostrom, E., Gibson, C., Shivakumar, S., & Andersson, K. (2002). *Aid, incentives, and sustainability: An institutional analysis of development cooperation*. Swedish International Development Cooperation Agency.
- Pistor, K., Raiser, M., & Gelfer, S. (2000). Law and finance in transition economies. *The Economics of Transition*, 8(2), 325–368.
- Prüfer, J., & Prüfer, P. (2018). Data science for institutional and organizational economics. In C. Ménard & M. M. Shirley (Eds.), *A research agenda for new institutional economics* (pp. 248–259). Edward Elgar.
- Prydz, E. B., & Wadhwa, D. (2019). Classifying countries by income. In *World development indicators, stories*. <https://datatopics.worldbank.org/world-development-indicators/stories/the-classification-of-countries-by-income.html>. September 9, 2019. Accessed 9 Aug 2023.
- Qian, Y. (2002). *How reform worked in China* (William Davidson Institute Working Paper 473).
- Rodrik, D. (2022). Prospects for global economic convergence under new technologies. In Z. Qureshi (Ed.), *An inclusive future? Technology, new dynamics, and policy challenges* (pp. 65–82). Brookings.
- Rodrik, D., Subramanian, A., & Trebbi, F. (2002). *Institutions rule: The primacy of institutions over geography and integration in economic development* (National Bureau for Economic Research Working Paper #9305).
- Sachs, J. (2003). *Institutions don't rule: Direct effects of geography on per capita income* (NBER Working Paper #9490).
- Schama, S. (1988). *The embarrassment of riches. An interpretation of Dutch culture in the Golden age*. Alfred A. Knopf.
- Shirley, M. M. (Ed.). (2002). *Thirsting for efficiency: The economics and politics of urban water system reform*. Elsevier Science.
- Shirley, M. M. (2008). *Institutions and development*. Edward Elgar.
- Siems, M. (2022). Legal origins. In J. Smits, J. Husa, C. Valcke, & M. Narciso (Eds.), *Elgar encyclopedia of comparative law*. Edward Elgar.
- Skarbek, D. (2020). Qualitative research methods for institutional analysis. *Journal of Institutional Economics*, 16(4), 409–422.
- Spiller, P. T., & Tommasi, M. (2007). *The institutional foundations of public policy in Argentina: A transactions cost approach*. Cambridge University Press.
- Svensson, J. (2000). Foreign aid and rent-seeking. *Journal of International Economics*, 51(2), 437–461.
- United Nations Development Programme. (2023). *Global multidimensional poverty index 2023. Unstacking global poverty: Data for high action impact*. United Nations Development Program.

- Vahabi, M., Batifoulier, P., & Da Silva, N. (2020). The political economy of revolution and institutional change: The elite and mass revolutions. *Revue d'Économie Politique*, 130(6), 855–889.
- Williamson, O. E. (1985). *The economic institutions of capitalism*. The Free Press.
- Williamson, O. E. (1996). *The mechanisms of governance*. Oxford University Press.
- World Bank. (2018). *World development report 2018: Learning to realize education's promise*. The World Bank.
- World Bank. (2023a). Poverty and inequality platform (version 2023 28-217-01-02-PROD) [dataset]. <https://pip.worldbank.org>. Accessed 31 July 2023.
- World Bank. (2023b). *World development indicators database*. <https://data.worldbank.org/indicators>. Accessed 2 Aug and 18 Oct 2023.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Institutional and Non-institutional Explanations of Economic Differences

30

Stanley L. Engerman and Kenneth L. Sokoloff

Contents

1	758
2	760
3	764
4	768
5	780
References	782

Abstract

In this essay, we outline some reasons why one should be cautious about grounding a theory of growth on institutions. We emphasize how very different institutional structures have often been found to be reasonable substitutes for each other, both in dissimilar, as well as similar, contexts. The historical record, therefore, does not seem to support the notion that any particular institution, narrowly defined, is indispensable for growth. Moreover, we discuss how the evidence that there are systematic patterns to the ways institutions evolve undercuts the idea that exogenous change in institutions is what powers growth. Institutions matter, but our thinking about how they matter should recognize that they are profoundly influenced by the political and economic environment, and that if any aspect of institutions is crucial for growth, it is that institutions change over time as circumstances change.

Stanley Engerman and Kenneth Sokoloff sadly passed away before this second edition of the Handbook of New Institutional Economics was created. Their chapter from the first edition remains a classic and emblematic of their invaluable contributions to the field.

S. L. Engerman
Department of Economics, University of Rochester, Rochester, NY, USA

K. L. Sokoloff
University of California, Los Angeles, CA, USA

Keywords

Institutional change · Development · Culture · Endogenous institutions · Franchise · Slavery · Colonialism

1

Economists have long been concerned with the explanation of differences across countries in levels of national income, population, and per capita incomes, as well as in their rates of growth. Because many of the processes of economic development operate over long periods of time, those studying the sources of these differences have quite naturally turned to the historical record for relevant evidence. Their concern with economic history thus comes not only from a desire to achieve a better understanding of the past, but also from a belief that such knowledge can serve as a guide for policymakers striving to improve the economic and social conditions of currently less developed nations. Many scholars have set about making contributions to knowledge through detailed investigations of the processes of growth in individual countries. Others have sought to discern what factors were crucial through comparative studies, focusing on issues such as why nations differed with regard to the timing of the onset of growth or how and why their records of achieved rates of growth varied over a long period of time.

Recently, considerable attention has been given to the question of why European nations and some of their overseas offshoots expanded more rapidly than did the economies of Asia, Africa, and Latin America after the eighteenth century, either generating new gaps in levels of income and rates of growth, or else greatly widening whatever differentials may have previously existed (see Table 30.1).¹ Previously the principal focus of historians examining the basis for differences in long-term economic performance had been with what led Great Britain to accomplish an Industrial Revolution sometime after the middle of the eighteenth century, and ahead of its European rivals (see Table 30.2).² Given the greater similarity of economic, political, and social structures among the European nations compared to those between Europe and the rest of the world, the factors highlighted in the discussions of the development of the Industrial Revolution are rather different from those generally featured in the broader geographic comparisons. In both cases, however, what the economists and economic historians are seeking to explain is why some nations in today's world remain poor, relatively and absolutely, and what conditions can be changed in order to achieve success in spurring growth and

¹ Among a vast literature, see North & Thomas, 1973; North, 1981; Rosenberg & Birdzell, 1986; Jones, 1987; Landes, 1998; Engerman & Sokoloff, 1997, 2002; Pomeranz, 2000; Acemoglu et al., 2001, 2002; Easterly & Levine, 2003. For even earlier discussions, see Weber, 1958, 1961; Sombart, 1969.

² See, for example, Hartwell, 1971.

Table 30.1 Levels of Per capita GDP and interregional differences, 1500–1998. (1990 dollars)

	1500	1820	1870	1913	1950	1973	1998
Western Europe	774	1232	1974	3473	4594	11,534	17,921
Western Offshoots	400	1201	2431	5257	9288	16,172	26,146
Japan	500	669	737	1387	1926	11,439	20,413
Asia (excl. Japan)	572	575	543	640	635	1231	2936
Latin America	416	665	698	1511	2554	4531	5795
Eastern Europe	483	667	917	1501	2601	5729	4354
Africa	400	418	444	585	852	1365	1368
World	565	667	867	1510	2114	4104	5709
Interregional Spreads	2:1	3:1	5:1	9:1	15:1	13:1	19:1

Source: As calculated in Maddison (2001)

Table 30.2 Per capita GDP in Western Europe, 1500–1998. (1990 dollars)

	1500	1600	1700	1820	1870	1913	1998
Austria	707	837	993	1218	1863	3465	18,905
Belgium	875	976	1144	1319	2697	4220	19,442
Denmark	738	875	1039	1274	2003	3912	22,123
Finland	453	538	638	781	1140	2111	18,324
France	727	841	986	1230	1876	3485	19,558
Germany	676	777	894	1058	1821	3648	17,799
Italy	1100	1100	1100	1117	1499	2564	17,759
Netherlands	754	1368	2110	1821	2753	4049	20,224
Norway	640	760	900	1104	1432	2501	23,660
Portugal	632	773	854	963	997	1244	12,929
Spain	698	900	900	1063	1376	2255	14,227
Sweden	695	824	977	1198	1664	3096	18,685
Switzerland	742	880	1044	1280	2202	4266	21,367
United Kingdom	714	974	1250	1707	3191	4921	18,714
Total Western Europe	774	894	1024	1232	1974	3473	17,921
World	565	593	615	667	867	1510	5709

Source: As calculated in Maddison (2001)

improving the welfare of the respective populations. It is this problem that the recent study of institutions has sought to help resolve and that probably represents its most significant contribution.

It is not necessary here to attempt to catalog the full set of explanations that have been given for differences in economic development, since many books and articles, published and forthcoming, have already done that. For present purposes, however, we highlight a transition over the last few decades from a concentration on the role of narrowly defined economic factors to a focus on the significance of various social structures and cultures in providing the conditions conducive to economic

development.³ Arguments based on conditions such as favorable natural resources (including accessible coal and iron, in the case of Britain), high rates of capital formation, and extensive markets or other circumstances that encourage a faster pace of technological change, which had long been central to our understanding of why some economies enjoyed better performance, have been replaced (or supplemented) by arguments about how differences across societies in political and cultural institutions arose, and how they influence the processes of growth.⁴ Although we cannot conceive of processes of economic growth that do not involve institutional change, in this essay we outline some reasons why one should be cautious about grounding a theory of growth on institutions. We emphasize how very different institutional structures have often been found to be reasonable substitutes for each other, both in dissimilar, as well as similar, contexts. The historical record, therefore, does not seem to support the notion that any particular institution, narrowly defined, is indispensable for growth. Moreover, we discuss how the evidence that there are systematic patterns to the ways institutions evolve undercuts the idea that exogenous change in institutions is what powers growth. Institutions matter, but our thinking of how they matter should recognize that they are profoundly influenced by the political and economic environment, and that if any aspect of institutions is crucial for growth, it is that institutions change over time as circumstances change.

2

A basic categorization of explanations for economic growth would include economic, cultural, political, and institutional factors. The importance of economic factors was much discussed in the ancient world, and among the numerous economic factors that have been considered since that time are: natural resources, such as the supplies of coal and iron; the opportunity to trade at low cost with other regions or nations, which provides markets that encourage specialization in producing goods in which the economy has a comparative advantage (and perhaps stimulating more rapid technical progress) as well as serve as a source of imports that a nation may be incapable of producing; climate, which can influence productivity through a variety of mechanisms; colonial empire, which might be associated with especially high private or social returns to investment; and the role of population change. Some contend, for example, that rapid population growth has sometimes proved beneficial,

³Early twentieth century scholars, such as Weber (1958, 1961), emphasized the role of culture, but the focus turned to real economic factors by the second half of the century. For a more extensive discussion of how thinking changed over time, see Arndt, 1978, 1987.

⁴On the role of coal, see Kindleberger, 1961, and the discussion by Parker in the same volume. Although attention to the significance of coal waned, it has recently revived with Wrigley (1988) and Pomeranz (2000). For discussion of the relative importance of changes in savings rates, or in investment opportunities, see Postan, 1935. For discussions of how the pace of technological change was responsive to economic factors, such as the extent of, or access to, markets, see Landes, 1969; Sokoloff, 1988.

fostering lower labor costs or the advantages in scale effects that come from higher total demand; others argue for the benefits of relatively slow growth in population, on the grounds that lower population density encourages higher per capita incomes and higher rates of capital formation. These, and other so-called economic explanations, say little explicitly about non-economic factors and institutions, although this does not mean that the latter are not implicit in the analysis.

Discussions of the role of non-economic factors (encompassing the cultural, political, and institutional) in accounting for differences across societies in economic development can also be traced back many centuries. Several of the classic theories for the rise of European capitalism and the onset of modern economic growth are based on conditions that clearly fall outside of the conventional economic sphere, such as the spread of particular religious beliefs, whether the Protestantism pointed to by Max Weber, the Judaism highlighted by Werner Sombart, or shifts in the orientation of dominant religions.⁵ Other arguments stress the important contributions of the advance of scientific and rational thought, or the impact of changing tastes for consumer goods and the effects on choices between work and leisure on the supply and intensity of labor during early industrialization.⁶ Changes in legal systems, in degrees of trust and the extent of social capital, and in the nature of political organization and the degree of democracy, have more recently been advocated as critical factors explaining differentials in economic performances.⁷ Although changes in these kinds of circumstances, such as in religion, are generally treated as exogenous to the economy, the nature of the interaction between economic and so-called “non-economic” factors may be complex. The contrast in views between Weber and R.H. Tawney on the relationship between religious changes and the rise of capitalism in Britain, and in northern Europe more generally, for example, corresponds to similar debates over the sources of change in many other purportedly non-economic conditions relevant to economic growth.⁸ Moreover, the implications of the very slow diffusion of cultural change (and of institutional change more generally) and economic growth around the world represent a puzzle for those who believe that introducing exogenous changes in these facets should have significant, favorable effects and constitute a viable instrument of economic policy.⁹

Quite a wide range of non-economic conditions relevant to growth have featured in debates over why Britain was the first industrial nation. Many can be subsumed under the blanket category of culture, where cultural factors are understood to

⁵ See the discussion of non-conformists in Britain in Weber (1958), Sombart (1969), McClelland (1961), and Fogel (2000).

⁶ See Steuart, 1767; Gilboy, 1932; Nef, 1958; De Vries, 1994; Jacob, 1997.

⁷ See, the discussions of different views in Berman (1983), Putnam (2002), and O'Brien (1988).

⁸ Whereas Weber is well known for his theory of how the content of Protestant thought may have encouraged believers to behave in ways we associate with capitalism, Tawney highlighted how economic change supported change in religious beliefs. See Weber, 1958; Tawney, 1926.

⁹ For a recent restatement of how the immutability of culture can explain continued backwardness, see Landes, 1998.

include: religion, particularly the impact of non-conformists in the development of technology and entrepreneurship; the scientific spirit and the expansion of knowledge, including a willingness to search out new methods and technologies; and the emergence of an educational system that permitted a wide diffusion of information and skills among the population. Culture has also been defined to include family and kinship patterns; tastes and preferences regarding work versus leisure; time preferences determining the levels of savings and consumption; and the development of a widespread desire to maximize financial profit or pursue material gain more generally. Proponents of the view that cultural change was responsible for economic change generally point to their coincidence in eighteenth century Britain, and presume that culture consists of behaviors and values that are determined independently of economic factors. This may, however, be an artificial distinction, because the economic effects of cultural factors, if not the cultural beliefs themselves, are often greatly influenced by the relative costs of different patterns of behavior and the amounts of income that people are willing to forgo to obtain chosen ends.¹⁰

The recent work on the significance of institutions for understanding why Britain industrialized first, and for understanding differences in economic performance more generally, gives relatively little attention to the role of culture per se. In emphasizing property rights and other aspects of the British legal framework, it breaks sharply from the previous stream of work on institutions by economists who emphasized culture in treating the evolution of economies, as part of a critique of classical economic theory.¹¹ Current thinking about institutions instead follows the pioneering approach of Douglass North in grounding the analysis of the causes and consequences of institutions and institutional change in theory.¹² This perspective defines institutions, though difficult to do with precision, as encompassing the specific organizations or rules that constrain and influence human behavior. A key aspect of these humanly devised rules is that they structure human actions by providing incentives that shape economic and political organization. Formal rules, together with the informal constraints that develop, influence the costs of production and of transactions within society. Among the institutions that are most important for economic performance are those involving the definition and enforcement of property rights, between the government and private parties and between the individuals within a society. The link between appropriate institutions and economic growth is that well-adapted institutions reduce the costs of production and distribution, allowing private agents greater scope to benefit from specialization, investment, and trade. Institutions, as human-imposed constraints, are not the only constraints that society or private actors confront, since there are others due to the state of

¹⁰For the linkage between morality and economics, see the discussion of Quakers and slavery in Smith (1979). Also see Fogel, 1989.

¹¹See the discussion of institutions, and the approaches of the British and Germans, in Cunningham (1890–1892).

¹²See North, 1981, 1988, 1990; North & Thomas, 1973; North & Weingast, 1989; Davis & North, 1971.

technological knowledge, demographic forces, nature (including climate and topography), as well as other features of the environment that may also have implications for the patterns of economic activity.¹³

Institutions, as described, play several roles in the economy. They influence the beliefs and behaviors of individuals and groups, and thus the preferences and priorities expressed through both private and public decisions. Another important role of institutions is to provide efficient property rights, trust, and effective incentives, thus facilitating the organization and conduct of appropriate and constructive transactions and interactions among individuals and firms. Indeed it is claimed, by North and others, that it was sound property rights and incentive schemes made possible by its distinctive institutions that were key to Britain industrializing first.¹⁴ No economic development is possible without secure property rights. However, the specification of formal rights is only one part of society's problem. The nature of the enforcement of institutional provisions, both in terms of accepted legitimacy and effectiveness, is critical to the success of whatever institutions exist. Similarly, legislative decisions and judicial rulings can also influence outcomes, whether or not they seem consistent with the circumstances under which the respective institutions were originally adopted. Enforcement is sometimes bilateral, between individuals, with no government role to ensure compliance, but in other cases enforcement requires governmental participation and action.

Although those who stress the importance of the institutional framework have somewhat different concerns than those who highlight the significance of culture, the two perspectives share an emphasis on the extent to which "non-economic" variables evolve independently of the processes of economic growth. Indeed, proponents of both views champion how the appreciation of this pattern, as well as the impact of those variables on the economy, constitute a salient intellectual advance over the earlier (circa 1950s through 1970s) literature on economic development, which focused primarily on economic variables, such as natural resources, physical capital, human capital (mainly education), exchange rates, and technical change. That generation of economists certainly accepted the importance of institutions, culture, or political stability, but presumed either that the appropriate institutions and beliefs existed already, or else that they would evolve in constructive directions relatively easily as the economic factors that could generate economic growth were in place. Since economic forces obviously do not operate in a vacuum, it may have been difficult for them to conceive of a non-institutional interpretation of economic growth.

¹³For a classic treatment of how factor endowments can help to shape culture and institutions, see Tocqueville 1835. For interesting discussions of the influences of factor endowments and political forces, see Brenner, 1985; Engerman & Sokoloff, 1997, 2002; North, 1981, 1990; North et al., 2000.

¹⁴A major concern of the property rights literature is with private agents being secure from expropriation by the state. See North, 1981; North & Weingast, 1989; Knack & Keefer, 1997; Acemoglu & Robinson, 2000; Haber et al., 2003.

Even as a purely logical construct, a wholly non-institutional (or anti-institutional) explanation of economic growth seems implausible (as would an explanation that takes no account of real economic factors), but as discussed below, debate on the relative importance of institutional and non-institutional forces has continued.¹⁵ The essential questions, thus, seem to us to be empirical. How much of the variation in economic performance across country and over time can be attributed to differences in institutions, with pure economic factors constant or endogenous with respect to institutions, and how much is due to differences in the economic variables, with institutions constant or endogenous? What are the processes that govern the ways specific institutions evolve, and under what circumstances can the introduction of exogenous institutional changes be considered viable economic policies?

3

Although in principle these questions can be framed as empirical issues, it is far from easy to clearly distinguish between, or gauge the relative power of, the institutional and non-institutional explanations of economic differences. No one would claim that there is a general answer, and indeed few, if any, individual cases seem not to allow some role for each type of explanation. There has been considerable interest in recent years, however, in a manner of posing the problem that might appear to make the empirical work more tractable: Are the key elements in determining institutions exogenous or endogenous? This distinction has been with us a long time, as in the debates over the superiority of British institutions, but it has figured prominently in the study of how the various economies established as colonies by Europeans (or others) developed over time. Even in the absence of a substantial indigenous population in the area of settlement, the presence of one group in the colony, arriving from the metropolis, and another remaining in the metropolis, meant that there were different circumstances for institutional development. In principle, therefore, a researcher could evaluate just how much of an impact the different circumstances had on the ways in which the institutions evolved. That many of the European countries established multiple colonies, in very different environments, further enhances the quality of the information arising from the natural experiment.

If the institutions in the colonies were, or remained, the same as those of the metropolitan nation (or perhaps the same as those of the indigenous societies that predated the arrival of the Europeans), they might be regarded as exogenous. In such a case, the institutions could be reasonably interpreted as evolving independently of the conditions in the respective colonial economies, and systematic patterns in subsequent differences in economic performance across the economies could, after

¹⁵ Even Weber's discussion of the role of Calvinism points out that the relation "was true only when some possibility of capitalistic development in the areas in question was present." See Weber, 1958, p. 190.

controlling for the purely economic factors, be attributed to institutions. If, however, the institutions in the colony diverged in ways that could be explained as adaptations by the respective population to the different environment, natural or human, then it would support the view that institutions were endogenous with respect to circumstances. Because institutions are human-fashioned structures that presumably reflect the efforts of populations trying to make the best of the opportunities and problems they face, most observers would be surprised if they were not at least partially endogenous. Indeed, most scholars feel that the institutions that emerged across the colonies established by the Europeans do seem to have varied systematically with aspects of the environment such as climate, land type, and natural resources. Some would go even further and suggest that the direction of institutional change is often endogenous to the growth process, as changes in technology and income generate changes in tastes, changes in the returns to organizing production and transactions in various ways, and changes in patterns of behavior more generally.

To acknowledge that there is some endogeneity to institutions does not imply that institutions are unimportant, or that they have only a limited impact on economic performance. Endogenous institutions, once in place, can prove as crucial as if they were exogenous, and they might persist for as long, or even longer. The key difference between those who contend that institutions are exogenous and those who argue they are endogenous is not with their impact and influence, but instead with where institutions come from and with the extent to which they are—or might be expected to be—revised over time.

It is widely recognized that it is sometimes useful to fix some types of institutions over time. Credible commitment to property rights is perhaps the classic example of the value of certainty about policy action.¹⁶ More generally, however, allowing some flexibility in institutions, such that they can be altered to make it easier for private or public agents to take fuller advantage of the new opportunities that arise as technology or the environment changes, would normally be expected to foster better economic performance and more rapid growth. Among the many such innovations in institutions that could be cited to illustrate the utility of institutions changing as conditions change are: those that created the modern patent system; those that extended suffrage to a broad range of the population; those that provided tax-supported primary and secondary schools that were free to those who chose to attend; the introduction of taxes on income; and those that had profound impacts on the structure of financial markets and institutions (such as the introduction of limited liability for shareholders in corporations, the establishment and evolution of public agencies to regulate the issuance and trading of securities, and the establishment of central or quasi-central banks such as the Federal Reserve or the Bank of England). Determining the optimum degree of flexibility, and designing mechanisms well suited to respond constructively to ever-changing circumstances with institutional change, are complex issues. While some see the role of constitutional provisions as a means of ensuring stability in the decision-making process and institutions more

¹⁶North & Weingast, 1989.

generally, most constitutions do have provisions for amendments, and allow some degree of legislative and governmental flexibility in setting the legal structure. Allowance for modifications to the laws need not harm the potential for growth, nor even yield instability, particularly given that the voting and other costs of implementing changes are typically high.¹⁷ Indeed, there are likely more cases of how nations and economies suffered from inflexible institutions than from excessive flexibility.¹⁸

Perhaps the most important elements of institutional structures are those that ensure an ability to adapt to different conditions and to adjust to new circumstances as seems necessary, rather than those that entail the retention and maintenance of any specific set of policies. The capability for adaptation may ultimately be more significant for economic growth than the continuation of any particular set of beliefs, rules, or behavior. Among the characteristics of a society that might be expected to enhance institutional flexibility are a population's level of education, their political liberties, the degree of decentralization in political or economic structures, and the extent of competition within and across polities. Some of these conditions may be relevant in considering the case of European expansion, where colonies were established in a wide range of environments, and settlers had to grapple with new sorts of climates, soil types, natural resources, and economic problems. As discussed below, although there were surely some specific cultural carryovers from Europe early in the settlement process, it is not clear that these factors were immutable or remained unchanged for long periods. The confrontation with new environments led to many adaptations, adjustments, and innovations in institutions, as colonial populations sought to make the most out of opportunities for economic improvement in circumstances that were unfamiliar, or at least different from those that had shaped the evolution of institutions in the Old World.

Another issue that is central to understanding how institutions matter for growth concerns the likelihood that no one particular narrowly-specified institution is required, as there are often alternative institutional forms or structures that are reasonable substitutes for each other and may achieve similar economic performance. Those who hold this view that non-optimal institutions may still be consistent with high rates of economic growth, though perhaps not the highest rate that was possible, often point to the stark contrasts across industrialized countries in the importance of banks relative to securities markets in financial intermediation, in the reliance on common or civil law, in how bankruptcy laws balance the rights of creditors and debtors, in systems and levels of taxation, and in the division of power between the executive, the legislature, and political party structures. These and many other examples, historical and contemporary, suggest the usefulness of institutions

¹⁷ In some cases of change there may be required compensation to be paid to those whose condition is weakened, requiring increased taxation of other members of the population. Nearly all serf and slavery systems that ended during the nineteenth century did so with compensation paid to property holders, not laborers, but there was no compensation of losers when the slave trade was ended.

¹⁸ See the discussions in Elbaum and Lazonick (1986).

generally in helping societies take advantage of the opportunities the environment offers them, but support the idea that no single institutional solution is crucial. In this way, the role of institutions might be considered analogous to the role of technology, in that the processes of change are important but no single method of accomplishing a goal is indispensable.¹⁹

A perhaps more serious issue is that among the feasible set of institutional solutions to a general problem, different approaches may have different implications for different segments of the population. Depending upon the manner in which institutions evolve, or are designed, in a society, they may develop to favor the interests of more powerful groups at the expense of others, or even of the population at large. For example, elites might prefer policies that raise their share of national income, even if they reduce long-run rates of growth. The nature of the political power structure in society is critical in determining which institutions are adopted. The suffrage, or the distribution of political influence more generally, may be rather broad and inclusive, with a relatively large share of the population able to vote.²⁰ Or, alternatively, the franchise may be limited, by requirements of literacy, wealth, nationality, age, and gender, with only a small minority of the population able to vote and to directly influence policy. When the suffrage is restricted, many members of society have only very limited political influence and no direct voice in establishing the institutional framework. Even a very small segment of the population, but one with highly disproportionate political power, would be able to establish institutions, legal codes, and property rights regimes that serve their own interests, and be able to exclude other members of society from benefits.²¹ Thus, there could well be a well-defined and enforced set of property rights, but one coincident with a large component of the population being outsiders to decision-making in society.²²

¹⁹ Thus, as suggested by Lance Davis and North, evaluating institutional change may be subject to the same type of benefit-cost analysis as are other economic factors. See Davis & North, 1971.

²⁰ See the discussions of the evolution of restrictions on suffrage in Engerman and Sokoloff (2001) and Keyssar (2000).

²¹ Of course, other conditions matter as well. Where there is extreme economic inequality, for example, an elite might be able to leverage its wealth into disproportionate political influence through informal channels. Another example of context mattering is where labor is scarce, such as on a frontier. The desire of an elite, even one with a monopoly on political voice, to attract migrants might lead to policies that groups that have no formal representation value. For a discussion of how this might have operated on the U.S. frontier, see Engerman & Sokoloff, 2001.

²² In most cases, we would expect that groups with no formal political influence would have quite circumscribed access to scarce resources. This limited access might be due to the laws explicitly favoring the dominant groups, even if the property rights allowed the “outsiders” were enforceable. A de facto limited access might, however, arise if the outside group lacked the financial or other resources necessary to take advantage of an opportunity they had a de jure legal right to. Thus, even though all citizens might be entitled to bring civil suits to enforce contracts, the poor may find themselves less able to act on this right. The case of married women is an interesting one to consider in this regard, as in the United States (and many other countries) the law allowed them only very limited rights as to owning property or entering into contracts as compared to those allowed men or single women, until the second half of the nineteenth century. See Khan, 1996.

Slave societies in the Americas often had well-defined institutions and property rights, and were capable of rapid economic growth, but part of their population had no rights and no means to obtain any. They provide a powerful example, albeit an extreme one, of how the determination of the size and nature of the elite groups, by political, economic, and/or military means, is critical to the establishment of institutions. Being excluded from voting does not necessarily mean a failure to benefit from economic change, nor that there will not be subsequent improvements in the rights to suffrage, but the limited nature of the decision-making group still raises important issues for our understanding of the distribution of rewards from economic activity. More generally, the observation that societies vary in how much influence different segments of the population can exert in shaping institutions implies that there may be systematic patterns in how flexible they are in adapting (or innovating) their institutions to enhance the ability of their populations to take advantage of new opportunities created by changes in the environment.

4

It is no doubt easier to isolate the effect of institutions if we believe that they are exogenously determined by the forces of past history or by forces outside the current economy. Among the factors that have sometimes been suggested as playing this role are: externally generated changes in mentalité due to changes in religious belief or secular attitudes; the outcome of a military conflict, either due to externally generated changes in the power structure or to internal revolutionary actions that altered the balance of political power; the non-military introduction of new foreign influence and contacts, reflecting, in part, improvements in transportation and communication; and, as shall be discussed in more detail below, the settling of new areas by people from a distant metropolis, whose institutions could be regarded as exogenous to the new area of settlement. In this context we consider non-institutional explanations to be not an absence of institutions (since that is not possible), but the presence of institutions regarded as endogenous to the socio-economic process, even when the circumstances giving rise to the institutions are themselves exogenous to the economy (e.g., climate and natural resources).

In evaluating whether institutions are endogenous, there are several approaches that could be taken. One concerns the impact of resources and natural and human endowments on institutions. A number of scholars have recently argued that there were systematic patterns in the types of institutions that evolved as settlers in European colonies adjusted to conditions that differed from those of the metropolis regarding disease environments and economic opportunities.²³ Subjecting this notion to empirical testing is complicated by the enormous range of institutions that could attract attention, some of which reflect metropolitan carryovers, while others developed quite differently in the colonies than they did in the metropolis.

²³ See Engerman & Sokoloff, 1997, 2002; Acemoglu et al., 2001, 2002; Easterly & Levine, 2003.

Metropolitan institutions did not necessarily disappear in the process of settlement, but many were modified depending on the conditions of the particular settlement. Thus British New World colonies might have employed British law, and French colonies continued French legal institutions, but British and French temperate zone colonies differed in many important ways from their respective Caribbean colonies. French and British Caribbean colonies had greater similarities to one another than they did with either their mainland counterparts or the metropolis. Although the initial controls over free white labor may have been much the same in all colonies, only some British colonies and only some French colonies came to rely on free, rather than slave, labor.²⁴ Climate and resources were the most powerful determinants of the geographic incidence of slave labor, irrespective of the metropolitan institutional structure. Slavery was legal in all the British colonies until the Revolutionary War, and differences in legal circumstances did not account for the major differences in its prevalence across them, such as between New England and the British West Indies.²⁵

Another approach to the question of whether institutions are endogenous is to consider whether economic growth itself influences people's attitudes, and the nature of the economy's institutions. Does economy itself contain the seeds of its own limitations, whether due to its failure or its successes? Karl Marx is certainly the most prominent of historical economists to have posited a sequence of self-generated endogenous changes in society, from feudalism to capitalism to socialism, with each of the first two stages being successful at first but then failing due to internal contradictions. Joseph Schumpeter claimed that the declining belief in the value of capitalism, which developed with economic growth, weakened capitalism's survival power, and he expected the "march into socialism" to occur based on the economy's success. Mancur Olson argued that as economies develop over time, vested interests operating in their own self-interest emerge and reduce the future growth of the economy through their success at rent seeking. Other notions of how self-generated changes in the economy stimulate institutional change highlight the impact of technological change on the scale of enterprise and the development of bureaucracies in business and government.

Scholars interested in how institutions evolve have recently devoted much attention to the contrasts between colonial and metropolitan influences on institutions in newly settled areas. A long-standing disagreement, tracing back centuries, regarding the 13 colonies that became the United States has been the causes of North-South differences in economic and demographic structures, including the explanation of the differences in the relative importance of slavery. Did those settling in different parts of the mainland arrive with rather different cultural patterns from Britain, differences that persisted after settlement, or did the various colonists from Britain arrive with basically similar cultural beliefs, but then adjust their institutions once

²⁴For an interesting analysis of adjustments to the laws governing slavery, see Cottrell, 2001.

²⁵For a discussion of the fascinating case of how a colony established with a prohibition on slavery came to have it lifted, see Wood, 1984.

established in the New World and confronted with a rather different set of conditions?'.²⁶ The evidence to date seems to favor the latter view. Not only has recent work demonstrated that even the Puritans were deeply influenced by the environment in selecting institutions for their two New World colonies, but studies of those Englishmen who came to populate the various settlements in the Americas emphasize how they were drawn from roughly the same social classes.²⁷ In the words of Edward Channing:

Historical writers have been altogether too prone to draw a hard and fast line of demarcation between the settlers of the Southern colonies and those who founded colonies north of the fortieth parallel. . . It is sometimes said that the Northern colonist came to the New World for conscience sake and the Southern planters sought wealth alone; but no such generalization can truthfully be made. Moreover, it is oftentimes the custom to point out some mysterious differences between the Virginian and the New Englander, which can be expressed by the words 'cavalier' and 'Puritan'. . . No such characterization is possible.²⁸

This perspective receives strong support from the record of slavery in the Americas. The basis for the success (to the owners) of slave labor in one area, and its failure in another, depended less on the initial attitudes of most settlers than upon the influence of climate and soil resources on the nature of those crops which could be grown and the technology and scale of efficient crop production. Wherever the soils and climates were suitable for growing sugar, the most valuable commodity in world trade during the seventeenth and eighteenth centuries and a crop that could be produced at the lowest cost on large slave plantations (under the gang labor system, which allowed owners to achieve very high labor intensity), slavery became the dominant institution of labor (and those of African descent constituted a dominant share of the population). Elsewhere, where soils and climates favored agricultural products (such as grains and hays) and where the gang labor system and slave labor offered no particular advantages, landowners had to rely more on free (often their own) labor, as the productivity of slaves in such settings would not warrant the high prices for slaves that prevailed on world markets. The populations of these settlements accordingly came to be much more homogeneous in wealth, human capital, ethnicity, and other dimensions. Thus, the factor endowments in the various colonies had a major impact on determining which labor institutions were dominant, the distribution of rewards between laborers and landowners, and the nature of political participation and decision-making. Because slavery was legal in all of the European colonies in the Americas, it is evident that not only did these natural forces lead to differences in institutions, but they also led to different outcomes from similar institutions. Although the Old World background was surely important, it is difficult to explain the extreme differences among the various areas within each colonial empire without reference to the effects of the New World circumstances.

²⁶ See, for example, Greene, 1988, 1993, 2002; Fischer, 1989.

²⁷ Kupperman, 1993.

²⁸ Channing, 1926, pp. 145–146.

The early history of the New World colonies established by the European nations permits one to examine some of the implications of focusing on exogenous factors in institutional development, as opposed to viewing institutions as largely endogenous. The locations of settlements were themselves subject to some choice, based on the demographic and economic characteristics of different locations. Moreover, the pattern of initial settlement was modified over time, as settlers learned more about prospects in different areas. In the settlement of the Caribbean by the British and the French, for example, the adjustments in terms of crops and labor institutions that took place over the first half-century of settlement were rather different from those that were to emerge in subsequent years. The problems that arise from selection by the colonizing powers notwithstanding, the natural experiment arising from the variety of settling metropolises—including Spain, Portugal, Britain, France, and Holland—and the extreme diversity of environments found among the colonies creates a wonderful laboratory in which to study the relationships between factor endowments, institutions, and economic growth. It should not be surprising that many scholars have been attracted to work in this laboratory. Their work, of course, does not amount to a comparison of institutional and non-institutional factors in economic growth, since everyone agrees on the importance of institutions. Rather, the work has sought to determine whether institutions can be understood as exogenous to the circumstances or economic system, or whether the environment or circumstances more broadly exert a powerful influence on how institutions emerge and evolve over time. Put simply, where do institutions come from?

A key economic question is the explanation for the post-1900 differences in levels of per capita income between the countries of mainland North America and those of Latin America (Table 30.3), differences that were much smaller during the colonial period. A closely related issue is why Latin America is the region of the world today with the most extreme inequality in income. Since the nations of South and Central America were settled mainly by the Spanish (the Portuguese settled Brazil), and the United States and Canada mainly by the British (pre-1763 Canada by

Table 30.3 Per capita GDP in selected new world economies, 1700–1997

Country	GDP per capita relative to the United States			
	1700	1800	1900	1997
Argentina	–	102	52	35
Barbados	150	–	–	51
Brazil	–	50	10	22
Chile	–	46	38	42
Cuba	167	112	–	–
Mexico	89	50	35	28
Peru	–	41	20	15
Canada	–	–	67	76
United States ^a	550	807	3859	20,230

Source: Sokoloff and Engerman (2000)

^aU.S. per capita GDP is measured in 1985 dollars

the French and pre-1664 New York by the Dutch, among the relatively less dominant settling nations), the traditional, and still popular, explanation holds that the different cultures, religions, and institutions of Britain and Spain could alone explain the divergent paths of economic development. Since there were sharp contrasts between the home countries of Britain and Spain in terms of economic and political structures, it is argued that the transfer of Old World institutions established the behavior of the economies and societies of the colonies in the New World, as differences in property rights determination and enforcement, legal frameworks more generally, economic goals, and religious beliefs were carried over, with little or no modification, into the new areas of settlement. The institutions that failed to generate sustained economic growth on the Iberian Peninsula likewise failed to do so in the New World, whereas the institutions that had evolved over centuries in Britain worked on both sides of the Atlantic. The logic is that either the political elites of the metropolis were carried over into the colonies, providing the political and legal framework for the successful carryover of institutions, or that the elites in the New World, though different from those of the metropolis, were able to use the same institutional structures to achieve similar ends by similar means.²⁹

An alternative explanation, one that has gained an increasing number of adherents of late, focuses on the economic and geographic circumstances in the area of settlement, and their influence on the determination of institutions in the new areas. As with the previous argument, there is a long literature on the role of climate and resources in influencing institutions and economic development. While a more detailed examination was provided by Montesquieu in the 1740s, a similar argument was made considerably earlier by Plato.³⁰ The links include the nature of the effects of climate on the willingness to work, the desire to emigrate or immigrate, the role played by slavery in society, and related economic concerns. Whether seen as the basic cause of the specific set of institutions, or as a reason to modify some pre-existing set of exogenous institutions, settlement societies can be argued to have been significantly influenced by factors other than some unchanged metropolitan institutions. Indeed most settlements made dramatic changes in their institutions after they were first established, searching for ways to enhance their profitability and survivability. The impact of climate and resources can also help to explain why the different areas settled by the same metropolitan power had rather different economic structures and performances (e.g., New England and the British West Indies), and why geographically contiguous and resource-similar areas settled by different metropolitan powers (as the British and French in the West Indies, as well as the Spanish, Danish, and Dutch there) came to resemble each other in many important ways. Indeed, recent scholarship has found strong evidence of the

²⁹It may be, however, that the distance between colony and homeland weakens the ability of the metropolis to control the settlers, weakening the nature of any transfer of political structure. See Greene, 1988, 2002. For the argument that the institutional heritage that British colonies drew on was more conducive to long-run growth than that the Spanish colonies worked from, see North, 1988.

³⁰see Montesquieu 1949; Plato., 1980, book five, section nine.

systematic effects of initial factor endowments on the types of institutions (including institutions involving suffrage and the conduct of elections, schooling, finance, the disposition of public lands, property rights, and intellectual property) that evolved in different colonies (and on long-term economic performance in these colonies), both in the Americas and elsewhere, highlighting how limited is the explanatory power of national heritage is.³¹

A specific example of how institutions can be altered to fit changing circumstances, and of how the distribution of power (both political and economic) as well as the environment influence outcomes, is provided by the adjustments in the societies of the Americas to the abolition of slavery.³² This most dramatic institutional change of the nineteenth century was, almost everywhere, imposed on a resistant slaveholding class in the aftermath of armed conflict or by a government elected by a population dominated by non-slaveholders (including European parliaments). All New World societies ended slavery between 1777 and 1888, and the nature of the abolitions was similar, with most providing some form of compensation, in cash, bonds, or labor time to the slaveowners, with very little or nothing going to the former slaves.³³ Nevertheless there were some striking differences in the range of post-emancipation responses. In the British West Indies, for example, slavery was abolished by 1834 and all colonial governments had the same basic goal of inducing labor to work on plantations and imposed legislation to try to accomplish this end. Different environments led to different outcomes, however, as

³¹ There is some question of what the comparison of institutions would indicate if some different dates were used for the evaluation. Spain arrived in America earlier than the British, went to areas with greater wealth and resources, and it took about another hundred years before the British arrived and were forced to go to areas that they regarded as clearly less promising for economic growth. For a comparison based on the year 1700 we would find the Spanish position seemingly more favorable than that of the British, reflecting the early economic advantages of the areas of Spanish settlement, and, as pointed out by John TePaske, the Spanish had three successful centuries in the Americas, and the British only two. See TePaske, 2002. Also see Engerman & Sokoloff, 1997, 2002; Acemoglu et al., 2002; Easterly & Levine, 2003.

³² The recent literature on the role of institutions in economic growth has raised important questions as to the extent to which political power is independent of economic power, and to the relative significance of political inequality and economic inequality for how institutions evolve. A full discussion of these issues is beyond the scope of this essay, but we would argue that although they are clearly related, and there is certainly an association across societies (or over time) between political power and economic power, the correlation is far from perfect and it is not all that uncommon for them to diverge. Their relative weight in the processes of institutional development likely varies with context. Moreover, some institutions may be more sensitive to political inequality, while others depend more on the extent of economic inequality. For example, the distribution of political influence would generally be expected to have a greater impact on public institutions, such as laws, and the distribution of wealth matter relatively more for the kinds of private institutions—such as financial institutions—that evolve.

³³ The United States, in 1865, was the one nation to free its slaves without any form of compensation provided in the form of cash, apprenticeship, or a “law of the free womb,” which required the free-born offspring of slaves to labor for the mother’s master into their late teens or twenties. Even Haiti agreed to pay compensation to the French after 1825, as a condition for the right to engage in trade with France.

Table 30.4 Land/Labor ratios, changes in sugar production in the British slave colonies prior to and after emancipation, and contract labor immigration

	Land/Labor Ratio ^a	% Change in Annual Sugar Prod. 1824–33 to 1839–46	Ratio of Sugar Production in 1887–96 to 1839–46	Contract Labor Immigration 1834–1918 (gross inflow)
Antigua	3.1	+8.7%	1.5	2600
Barbados	1.7	+5.5	3.5	–
St. Kitts ^b	2.9	+3.8	2.7	2900
Nevis	5.0	–43.1	(2)	(2)
Trinidad	47.7	+21.7 ^c	3.0	157,700
British Guiana	832.4	–43.0	3.4	301,000
Mauritius	8.0	+54.3	3.1	451,800
Dominica	16.3	–6.4	0.7	6000
St. Lucia	15.5	–21.8	1.7	5200
Montserrat	4.6	–43.7	2.5	–
St. Vincent	5.7	–47.5	0.7	5600
Tobago	8.8	–47.5	(4)	(4)
Jamaica	12.2	–51.2	0.6	53,900
Grenada	6.3	–55.9	0.0	6200

Source: Engerman, 1996

^aSquare Miles per thousand total population, just prior to abolition

^bNevis data merged with St. Kitts after 1882

^cTrinidad output did decline slightly after abolition, and it was not until 1845 that the 1834 level of output was regained. Tobago data merged with Trinidad after 1891. The 1877–86 level of sugar production in Tobago was about one-third less than it was in 1824–33

evident in the corresponding variation in the ratio of land to labor (see Table 30.4).³⁴ Areas of high population density such as Barbados maintained plantation systems and high sugar output, while those with low population density, with abundant frontier land, initially saw the end of the plantation system and a decline in sugar output. In cases where the islands had been relatively unproductive, sugar output continued to decline and the plantation system was never re-introduced. However, in those areas where land was highly productive and had been growing rapidly before emancipation (such as Trinidad and British Guiana), plantation systems returned in several decades, but were based on indentured labor drawn mainly from India, and not on ex-slaves. Thus the elite's ability to achieve their desired end, extracting the returns to the land they owned, was influenced by various other conditions, including resource endowments, and their efforts to achieve their goals, subject to the dissimilar constraints they faced, led to differences in institutional development.

³⁴See Engerman, 1982, 1996. For a classic treatment of the relationship between the land to labor ratio and institutions, see Domar, 1970.

Another example of how the evolution of institutions across New World societies reflected adjustments to different or changed circumstances is provided by the history of how broadly the franchise was extended over time and what fractions of respective populations actually voted in elections. Since most of the societies in the Americas had achieved independence from their colonial masters, and were at least nominal democracies, by the middle of the nineteenth century, suffrage institutions had a direct bearing on the extent to which elites based largely on wealth, human capital, and gender held disproportionate political power in their respective countries, and on their ability to shape government policies. The ability and inclination of the elites to maintain disproportionate political influence through the formal rules associated with the electoral process varied with a variety of circumstances. Among these circumstances was the extent of inequality in wealth, human capital, and political influence that existed at the time of independence, when conventions were generally held to draw up constitutions for the new nations. One simple or straightforward explanation of this pattern is that the greater the disparity in resources (which we have argued in other work was due to factor endowments during initial colonization), the greater the elite's ability to frame the rules in such a way as to preserve their relative political power.³⁵ Among the other factors that appear to have had significant effects on the way institutions evolved, however, was the relative scarcity of labor. Although elites were generally reluctant to share their access to political influence and economic opportunity with other segments of the population, they were more likely to do so in settings where they would benefit from attracting or retaining a scarce resource—labor.

The evidence on the evolution of suffrage institutions in the New World is quite consistent with this view. Summary information on how the right to vote was restricted across New World societies in the nineteenth and early twentieth centuries is reported in Table 30.5. The estimates reveal that the United States and Canada were the clear leaders in doing away with restrictions based on wealth and literacy and introducing the secret ballot, and achieving much higher fractions of the populations voted in these countries than anywhere else in the Americas. These societies were distinguished for their relative equality, population homogeneity, and scarcity of labor, and it is notable that others of British heritage, such as Barbados, generally retained stringent restrictions on the franchise well into the twentieth century. Moreover, it is striking that the leaders in extending suffrage in South and Central America, such as Uruguay, Argentina, and Costa Rica, are generally regarded as having been historically the most egalitarian of Latin American societies, and having initial factor endowments most closely resembling those of the United States and Canada.

The contrast between North and South America in the application of binding restrictions on the franchise was not so evident at the outset. Despite the sentiments popularly attributed to the Founding Fathers, voting in the United States was largely

³⁵ For discussions of how factor endowments are the principal source of the differences in inequality, see Engerman & Sokoloff, 1997, 2002.

Table 30.5 Laws Governing the Franchise and the Extent of Voting in Selected New World Countries, 1840–1940

Period and country	Year	Lack of secrecy in balloting	Wealth requirement	Literacy requirement	Percent of the population voting
1840–80					
Chile	1869	No	Yes	Yes	1.6
	1878	No	No	No ^a	–
Costa Rica	1890	Yes	Yes	Yes	–
Ecuador	1848	Yes	Yes	Yes	0.0
	1856	Yes	Yes	Yes	0.1
Mexico	1840	Yes	Yes	Yes	–
Peru	1875	Yes	Yes	Yes	–
Uruguay	1840	Yes	Yes	Yes	–
	1880	Yes	Yes	Yes	–
Venezuela	1840	Yes	Yes	Yes	–
	1880	Yes	Yes	Yes	–
Canada	1867	Yes	Yes	No	7.7
	1878	No	Yes	No	12.9
United States	1850	No	No	No	12.9
	1880	No	No	No	18.3
1881–1920					
Argentina	1896	Yes	Yes	Yes	1.8 ^b
	1916	No	No	No	9.0
Brazil	1894	Yes	Yes	Yes	2.2
	1914	Yes	Yes	Yes	2.4
Chile	1881	No	No	No	3.1
	1920	No	No	Yes	4.4
Colombia	1918 ^c	No	No	No	6.9
Costa Rica	1912	Yes	Yes	Yes	–
	1919	Yes	No	No	10.6
Ecuador	1888	No	Yes	Yes	2.8
	1894	No	No	Yes	3.3
Mexico	1920	No	No	No	8.6
Peru	1920	Yes	Yes	Yes	–
Uruguay	1900	Yes	Yes	Yes	–
	1920	No	No	No	13.8
Venezuela	1920	Yes	Yes	Yes	–
Canada	1911	No	No	No	18.1
	1917	No	No	No	20.5
United States	1900	No	No	Yes ^d	18.4
	1920	No	No	Yes	25.1

(continued)

Table 30.5 (continued)

Period and country	Year	Lack of secrecy in balloting	Wealth requirement	Literacy requirement	Percent of the population voting
1921–40					
Argentina	1928	No	No	No	12.8
	1937	No	No	No	15.0
Bolivia	1951	—	Yes	Yes	4.1
Brazil	1930	Yes	Yes	Yes	5.7
Colombia	1930	No	No	No	11.1
	1936	No	No	No	5.9
Chile	1920	No	No	Yes	4.4
	1931	No	No	Yes	6.5
	1938	No	No	Yes	9.4
Costa Rica	1940	No	No	No	17.6
Ecuador	1940	No	No	Yes	3.3
Mexico	1940	No	No	No	11.8
Peru	1940	No	No	Yes	—
Uruguay	1940	No	No	No	19.7
Venezuela	1940	No	Yes	Yes	—
Canada	1940	No	No	No	41.1
United States	1940	No	No	Yes	37.8

Source: Engerman, Haber, and Sokoloff (2000)

^aAfter having eliminated wealth and education requirements in 1878, Chile instituted a literacy requirement in 1885, which seems to have been responsible for a sharp decline in the proportion of the population that was registered to vote

^bThis figure is for the city of Buenos Aires, and it likely overstates the proportion who voted at the national level

^cThe information on restrictions refers to national laws. The 1863 Constitution empowered provincial state governments to regulate electoral affairs. Afterwards, elections became restricted (in terms of the franchise for adult males) and indirect in some states. It was not until 1948 that a national law established universal adult male suffrage throughout the country. This pattern was followed in other Latin American countries, as it was in the United States and Canada to a lesser extent

^dConnecticut and Massachusetts introduced literacy requirements during the 1850s. Sixteen other states, seven southern and eleven non-southern, introduced literacy requirements between 1889 and 1926

a privilege reserved for white men with significant amounts of property until early in the nineteenth century. By 1815, only four states had adopted universal white male suffrage, but as the movement to do away with political inequality gained strength, the rest of the country followed suit: virtually all new entrants to the Union extended voting rights to all white men (with explicit racial restrictions generally introduced in the same state constitutions that did away with economic requirements), and older states revised their laws in the wake of protracted political debates (see Table 30.6). The key states of New York and Massachusetts made the break with wealth restrictions in the 1820s, and the shift to full white adult male suffrage was largely complete by the late 1850s (with Rhode Island, Virginia, and North Carolina being

Table 30.6 Summary of economic-based qualifications for suffrage across the United States, 1787–1860

	Qualification in 1787 or Year of Entry	Year Economic Qualifications Ended, or Qualif. in 1860
<i>Original Thirteen</i>		
New Hampshire	Tax	1792
Massachusetts	Property	1821 (prop), tax req. in 1860
Rhode Island	Property	1842 (prop), tax req. in 1860
Connecticut	Property	1818 (prop), 1845 (tax)
New York	Property	1821 (prop), 1826 (tax)
New Jersey	Property	1807 (prop), 1844 (tax)
Pennsylvania	Tax	Tax req. in 1860
Delaware	Property	1792 (prop), tax req. in 1860
Maryland	Property	1802
Virginia	Property	1850
North Carolina	Property	1856 (prop), tax req. in 1860
South Carolina	Tax	1810 (tax)
Georgia	Property	1789 (prop), 1798 (tax)
<i>New States</i>		
Vermont	None (1791)	
Kentucky	None (1792)	
Tennessee	None (1796)	
Ohio	Tax (1803)	1851 (tax)
Louisiana	Tax (1812)	1845 (tax)
Indiana	None (1816)	
Mississippi	Tax (1817)	1832 (tax)
Illinois	None (1818)	
Maine	None (1819)	
Alabama	None (1819)	
Missouri	None (1820)	

Source: Engerman and Sokoloff (2001)

the laggards). The relatively more egalitarian populations of the western states, which were anxious to increase their populations, were the clear leaders in the movement. The rapid extension of access to the franchise in these areas not coincidentally paralleled liberal policies toward public schools and access to land, as well as other policies that were expected to be attractive to potential migrants. The frontier states in the West continued to be labor scarce, with low female to male ratios, and it is perhaps not surprising that late in the nineteenth century they were the leaders in extending suffrage to women, as well as in greatly strengthening the property rights available to married women.

Similar political movements with similar outcomes followed with a short lag in the various Canadian provinces, but the analogous developments did not occur in Latin America until the twentieth century. As a result, through 1940 the United States and Canada routinely had proportions voting that were 50–100% higher

than their most progressive neighbors to the South, three times higher than Mexico, and up to five to ten times higher than countries such as Brazil, Bolivia, Ecuador, and even Chile. It is remarkable that as late as 1900, none of the countries in Latin America had the secret ballot or more than a miniscule fraction of the population casting votes. The great majority of European nations, as well as the United States and Canada, achieved secrecy in balloting and universal adult male suffrage long before other countries in the western hemisphere, and the proportions of the populations voting in the former were always higher, often four to five times higher, than those in the latter. Although many factors may have contributed to the low levels of participation in South America and the Caribbean, wealth and literacy requirements were serious binding constraints well into the twentieth century.

What accounts for this pattern of diffusion of universal male suffrage across New World societies? One obvious explanation, noted above, is that differences in the degrees of inequality in wealth, human capital, and political influence were related to the likelihood of adopting such an institutional change. The cross-sectional patterns, as well as the histories indicating that the attainment of universal male suffrage and of the secret ballot was often the product of a long series of hard fought political battles, with the elites more likely to be opposed to liberalizing the franchise, are certainly consistent with this view.³⁶ Another important factor, however, was the desire to attract immigrants. It is striking that pioneers in extending suffrage, such as new (those after the original thirteen) states in the United States, Argentina, and Uruguay, did so during periods in which they hoped to attract migrants, such that the rights to suffrage formed part of a package of policies thought to be potentially attractive to those contemplating relocation. When elites—such as large holders of land or other assets—desire common men to locate in the polity, they thus may choose to extend access to privileges and opportunities without threat of civil disorder; indeed, a polity (or one set of elites) may find itself competing with another to attract the labor or whatever else is desired. Alternative explanations, such as the importance of national heritage, are not very useful in identifying why Argentina, Uruguay, and Costa Rica pulled so far ahead of their Latin American neighbors, or why other British colonies in the New World lagged behind Canada.³⁷

³⁶The achievements of a broadening of the formal requirements for suffrage, as well as of the more administrative procedures governing the conduct of elections, tend to overstate the reduction in the extent of political inequality. Economic elites always enjoy disproportionate informal political influence, and there is likely more scope for them to bypass formal procedures and channels in contexts where there is extreme inequality. Hence, it may not be so surprising that the outcomes did not improve more than they did after the extension of the suffrage in many of the Latin American societies.

³⁷For a fuller discussion of the issues involved in the evolution of suffrage institutions in the New World, see Engerman & Sokoloff, 2001. Also see Acemoglu & Robinson, 2000, for treatment of the somewhat different pattern in Europe.

How broadly a society chooses to extend the franchise is a fundamental political arrangement. Many scholars, including us, have noted that differences across polities in suffrage institutions, as well as changes over time, can often be related to decisions about government policies, such as what to do about public lands, what types of schooling and other public services and investments to support (and how to raise the revenue for them), and how to regulate financial institutions. As important as it is, however, the breadth of the franchise, or even the rate of participation in elections, is far from the only feature of political systems that matters. Many authoritarian governments stage elections, and voting is often as irrelevant in these contexts as it was in the Soviet Union. Hence, although we have highlighted the uneven evolution of suffrage institutions across the Americas, we do not find it at all surprising that the attainment of something approximating universal adult suffrage throughout the hemisphere has thus far failed to dramatically reduce the disparities in the structure of institutions or in economic performance across countries. Not only is much more involved in determining political outcomes, but there are many factors other than the distribution of political influence that shape the development of institutions.

This brief excursion into specific historical examples of the evolution of key institutions illustrates several key points. The first obvious implication is that important changes in institutions, such as the adaptations to the plantation system after the abolition of slavery in the British West Indies and the expansion of the fraction of the population eligible to vote, do occur, and occur in different directions in different countries and contexts. A second observation is that there appears to be a substantial systematic component to the variation in how these institutions evolved, with significant explanatory power coming from circumstances that were likely largely exogenous to previously existing institutions.

5

There are always institutions present, and we cannot conceive of a framework for making sense of the processes of economic development that does not include a role for them. Moreover, there is no doubt that for any given society, some institutions may limit the extent to which it realizes its potential economic output, while alternatives might do better. That said, it is unclear how firmly theories of economic growth can be grounded on institutions. Economists do not have a very good understanding of where institutions come from, or why some societies have institutions that seem conducive to growth, while others are burdened by institutions less favorable for economic performance. Until they do, it will be quite difficult to specify the precise role of institutions in processes of growth.

As we have sought to highlight in this essay, what little we presently know about the evolution of institutions suggests caution about making strong claims about their relationship to growth. First, it is clear that very different institutional structures often seem to be reasonable substitutes in being conducive to growth, both in dissimilar, as well as in similar, contexts. Narrow definitions of institutional

requirements for growth do not, accordingly, seem appropriate. Second, the case for attributing growth to institutions is weaker if institutions are endogenous rather than exogenous, and the evidence that there are systematic patterns to the ways institutions evolve makes the latter view problematic. The recent studies of the natural experiment in institutional development provided by the European colonization of the Americas (and of many other parts of the globe), for example, imply that the broad environment (reflecting factor endowments, social arrangements, or technology) had powerful effects on the sorts of institutions that evolved in respective colonies. Institutions obviously matter for growth, but the way we understand how they matter will be somewhat different if the agents and other forces shaping institutions are responsive to the conditions they face than if institutions develop independently of (or could be imposed in any) context.

The recognitions that the institutional structure appropriate for one environment may not be appropriate for another, and that the history of institutions in high-performing societies is one of change over time in response to changing circumstances, suggest a different perspective on the relation between institutions and growth. Although we all understand that there are favorable aspects to governments making credible commitments to various obligations or arrangements they enter into, such as enforcement of property rights, it is also clear that in theory one would want institutions to vary over time and place with the environment, technology, and values. One might, therefore, think of societies with institutions conducive to growth as being those that have exhibited greater institutional flexibility—where by institutional flexibility we mean the ease with which institutional adaptations that respond constructively to changes in circumstances are innovated and/or diffused. Societies with good institutions would, therefore, have institutions well-adapted for economic performance in their specific settings because they had implemented a series of institutional modifications or innovations (public and private) that cumulatively generated improvements in welfare. Societies with bad institutions are those with institutional inflexibility, whose institutions did not respond constructively to take advantage of the opportunities created by their environment and state of knowledge. Such a framework would encourage scholars interested in the relation between institutions and growth to devote more attention to the factors influencing the rate and direction of institutional change, and rather less attention to the quest for a set of institutional structures that would be universally effective at promoting growth.

Acknowledgments We would like to express deep appreciation for the research assistance of Elisa Mariscal, Patricia Juarez, and Luis Zegarra. We have also benefited from discussions with Daron Acemoglu, Lance Davis, David Dollar, David Eltis, Jeff Frieden, Avner Greif, Stephen Haber, Karla Hoff, Daniel Kaufmann, Zorina Khan, Naomi Lamoreaux, Frank Lewis, Peter Lindert, John Majewski, Douglass North, James Robinson, Jean-Laurent Rosenthal, Mary M. Shirley, Joel Slemrod, Peter Temin, John Wallis, and Jeffrey Williamson. We gratefully acknowledge the financial support we have received from the National Science Foundation, as well as from the Academic Senate at the University of California, Los Angeles.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., & Robinson, J. A. (2000). Why did western Europe extend the franchise? *Quarterly Journal of Economics*, 115, 1167–1199.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91, 1369–1401.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2002). Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *Quarterly Journal of Economics*, 117, 1231–1294.
- Arndt, H. W. (1978). *The rise and fall of economic growth: A study in contemporary thought*. Longman Cheshire.
- Arndt, H. W. (1987). *Economic development: The history of an idea*. University of Chicago Press.
- Berman, H. J. (1983). *Law and revolution: The formulation of the western legal tradition*. Harvard University Press.
- Brenner, R. (1985). Agrarian class structure and economic development in pre-industrial Europe. In T. H. Aston & C. H. E. Philpin (Eds.), *The Brenner debate: Agrarian class structure and economic development in pre-industrial Europe*. Cambridge University Press.
- Channing, E. (1926). *A history of the United States* (Vol. I). Macmillan.
- Cottrell, R. J. (2001). The long lingering shadow: Law, liberalism, and cultures of racial hierarchy and identity in the Americas. *Tulane Law Review*, 76, 11–79.
- Cunningham, W. (1890–1892 [1882]). *The growth of English industry and commerce* (2 vols, 2nd ed.). Cambridge University Press.
- Davis, L. E., & North, D. C. (1971). *Institutional change and American economic growth*. Cambridge University Press.
- De Vries, J. (1994). The industrial revolution and the industrious revolution. *Journal of Economic History*, 54, 249–270.
- Domar, E. D. (1970). The causes of slavery or serfdom: A hypothesis. *Journal of Economic History*, 30, 18–32.
- Easterly, W., & Levine, R. (2003). Tropics, germs, and crops: The role of endowments in economic development. *Journal of Monetary Economics*, 50, 3–39.
- Elbaum, B., & Lazouck, W. (1986). *The decline of the British economy*. Clarendon Press.
- Engerman, S. L. (1982). Economic adjustments to emancipation in the United States and the British West Indies. *Journal of Interdisciplinary History*, 12, 191–220.
- Engerman, S. L. (1996). The land and labour problem at the time of the legal emancipation of the west Indian slaves. In R. A. McDonald (Ed.), *West Indies accounts: Essays in the history of the British Caribbean and the Atlantic economy in honor of Richard Sheridan*. University of West Indies Press.
- Engerman, S. L., & Sokoloff, K. L. (1997). Factor endowments, institutions, and differential paths of growth among new world economies: A view from economic historians of the United States. In S. Haber (Ed.), *How Latin America fell behind*. Stanford University Press.
- Engerman, S. L., & Sokoloff, K. L. (2001). *The evolution of suffrage institutions in the Americas* (Working Paper 8512). National Bureau of Economic Research.
- Engerman, S. L., & Sokoloff, K. L. (2002). Factor Endowments, inequality and paths of development among new world economies. *Economia*, 3, 41–109.
- Engerman, S. L., Haber, S., & Sokoloff, K. L. (2000). Inequality, institutions, and differential paths of growth among new world economies. In C. Ménard (Ed.), *Institutions, contracts, and organizations*. Edward Elgar.
- Fischer, D. H. (1989). *Albion's seed: Four British folkways in America*. Oxford University Press.
- Fogel, R. W. (1989). *Without consent or contract*. W. W. Norton.
- Fogel, R. W. (2000). *The fourth great awakening & the future of egalitarianism*. University of Chicago Press.
- Gilboy, E. (1932). Demand as a factor in the industrial revolution. In *Facts and factors in economic history: Articles by former students of Edwin Francis Gay*. Harvard University Press.

- Greene, J. P. (1988). *Pursuits of happiness*. University of North Carolina Press.
- Greene, J. P. (1993). *The intellectual construction of America: Exceptionalism and identity from 1492 to 1800*. University of North Carolina Press.
- Greene, J. P. (2002). Transatlantic colonization and the redefinition of empire in the early modern era: The British-America experience. In C. Daniels & M. V. Kennedy (Eds.), *Negotiated empires: Centers and peripheries in the Americas, 1500–1820*. Routledge.
- Haber, S., Razo, A., & Maurer, N. (2003). *The politics of property rights*. Cambridge University Press.
- Hartwell, R. M. (1971). *The industrial revolution and economic growth*. Methuen.
- Jacob, M. C. (1997). *Scientific culture and the making of the industrial west*. Oxford University Press.
- Jones, E. L. (1987). *The European miracle: Environments, economics, and geopolitics in the history of Europe and Asia*. Cambridge University Press.
- Keyssar, A. (2000). *The right to vote: The contested history of democracy in the United States*. Basic Books.
- Khan, B. Z. (1996). Married women's property laws and female commercial activity. *Journal of Economic History*, 56, 356–388.
- Kindleberger, C. P. (1961). "International trade and investment and resource use in economic growth" and William N. Parker, "comment". In J. J. Spengler (Ed.), *National resources and economic growth*. Resources for the Future.
- Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff. *Quarterly Journal of Economics*, 112, 1251–1288.
- Kupperman, K. O. (1993). *Providence Island. 1630–1641: The other Puritan colony*. Cambridge University Press.
- Landes, D. S. (1969). *The unbound Prometheus*. Cambridge University Press.
- Landes, D. S. (1998). *The wealth and poverty of nations: Why some are so rich and others so poor*. W.W. Norton.
- Maddison, A. (2001). *The world economy: A millennial perspective*. OECD.
- McClelland, D. C. (1961). *The achieving society*. Van Nostrand.
- Montesquieu, C. d. S. b. d. (1949 [1748]). *The spirit of the laws*. Hafner Press.
- Nef, J. U. (1958). *Cultural foundations of industrial civilization*. Cambridge University Press.
- North, D. C. (1981). *Structure and change in economic history*. W.W. Norton.
- North, D. C. (1988). Institutions, economic growth and freedom: An historical introduction. In M. Walker (Ed.), *Freedom, democracy, and economic welfare*. Fraser Institute.
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- North, D. C., & Thomas, R. P. (1973). *The rise of the western world: A new economic history*. Cambridge University Press.
- North, D. C., & Weingast, B. (1989). Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century England. *Journal of Economic History*, 49, 803–832.
- North, D. C., Summerhill, W., & Weingast, B. (2000). Order, disorder, and economic change: Latin America versus North America. In B. B. de Mesquita & H. L. Root (Eds.), *Governing for prosperity*. Yale University Press.
- O'Brien, P. K. (1988). The political economy of British Taxation, 1660–1815. *Economic History Review*, 41, 1–32.
- Plato. (1980). *Laws*. Basic Books.
- Pomeranz, K. (2000). *The great divergence: Europe, China, and the making of the modern world economy*. Princeton University Press.
- Postan, M. M. (1935). Recent trends in the accumulation of capital. *Economic History Review*, 6, 1–12.
- Putnam, R. D. (Ed.). (2002). *Democracies in flux: The evolution of social capital in contemporary society*. Oxford University Press.

- Rosenberg, N., & Birdzell, L. E., Jr. (1986). *How the West grew rich: The economic transformation of the industrial world*. Basic Books.
- Smith, A. (1979 [1776]). *Inquiry into the nature and causes of the wealth of nations*. Clarendon Press.
- Sokoloff, K. L. (1988). Inventive activity in early industrial America: Evidence from patent records, 1790–1846. *Journal of Economic History*, 48, 813–850.
- Sokoloff, K. L., & Engerman, S. L. (2000). Institutions, factor endowments, and paths of development in the new world. *Journal of Economic Perspectives*, 14, 217–232.
- Sombart, W. (1969 [1913]). *The Jews and modern capitalism*. Burt Franklin.
- Steuart, J. S. (1767). *An inquiry into the principles of political economy*. A. Millar and T. Cadell.
- Tawney, R. H. (1926). *Religion and the rise of capitalism: A historical study*. Harcourt, Brace and World.
- TePaske, J. J. (2002). The vital peripheries of colonial South America. In C. Daniels & M. V. Kennedy (Eds.), *Negotiated empires: Centers and peripheries in the Americas, 1500–1820*. Routledge.
- Weber, M. (1958 [1905–6]). *The protestant ethic and the spirit of capitalism*. Scribner's.
- Weber, M. (1961 [1927]). *General economic history*. Collier Books.
- Wood, B. (1984). *Slavery in colonial Georgia, 1730–1775*. University of Georgia Press.
- Wrigley, E. A. (1988). *Continuity, chance, and change*. Cambridge University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Desiree Desierto and Mark Koyama

Contents

31.1	Introduction	786
31.2	Defining Institutions and Institutional Change	787
31.3	Institutional Functionalism	790
31.4	Conflict and Cooperation as Drivers of Institutional Change	794
31.4.1	Early States	796
31.4.2	The Demise of Serfdom in Western Europe	797
31.4.3	Institutional Change in Early Modern England	800
31.5	A Theory of Cooperation and Conflict	801
31.5.1	Feudalism	802
31.5.2	Magna Carta and Institutional Change in Medieval Europe	802
31.6	Critical Junctures and the Role of Individuals	804
31.7	Concluding Comments	805
	References	806

Abstract

How does institutional change take place? This question poses a number of challenges for scholars working in the New Institutional tradition. We first discuss how institutions have been conceptualized as either “rules of the game” or as “equilibria.” The rules of the game perspective suggests that institutions perform useful functions and change when circumstances make these functions obsolete except when transaction costs render this change too costly. The institutions as

We thank Mary Shirley and two anonymous referees for comments.

D. Desierto (✉)

Department of Economics, George Mason University, Fairfax, VA, USA

e-mail: ddesiert@gmu.edu

M. Koyama

Department of Economics, CEPR, and Mercatus Center, George Mason University, Fairfax, VA, USA

e-mail: mkoyama2@gmu.edu

equilibria perspective, in contrast suggests that both cooperation and conflict drive institutional change and that institutional change need not be efficient. We propose a framework for simultaneously depicting institutional change driven by both cooperation and conflict. Finally, we discuss the role individuals can play in shaping institutional change.

Keywords

Institutions · Institutional change · Conflict · Cooperation · Efficiency

31.1 Introduction

Institutional analysis was a core part of economics (or political economy) for Adam Smith (1776) and his contemporaries. Then in the mid-twentieth century, institutions largely fell out of the corpus of economic theory taught to graduate students or published in leading journals. Since the 1990s, this trajectory has been reversed. Multiple Nobel Prizes and John Bates Clark awards have been awarded for the study of institutions. The work of Douglass North and his various collaborators re-established the importance of institutions for understanding economic history and economic development. Institutions are mainstream again.

Nonetheless, the question, “How do institutions change?” remains an open one. Indeed, it is perhaps the most challenging question facing those who study institutions. In this chapter, we provide a survey of various approaches to the topic of institutional change.¹

We first outline the challenge posed by the topic to scholars of institutions. Our central claim is that the tools involved in studying the effects of institutions are not always appropriate for studying how institutions change. We go on to analyze a prominent set of hypotheses of how institutions emerge and why they disappear that we group under the label “institutional functionalism.” Institutional functionalism provides a useful set of tools for studying institutions. However, distinguishing between a strong and a weak version of the institutional functionalism hypothesis, we argue that it does not provide a fully satisfying account of why and how institutions change. In its place, we argue for a more game-theoretic approach to the study of institutions that both complements existing functionalist perspectives and provides greater insight into why institutions change over time.

We provide several examples drawn from the recent literature in economic history: the emergence of the state; Magna Carta and the gradual emergence of constitutional government; the decline of serfdom; and the Glorious Revolution. We go on to discuss the role of critical junctures and individuals in institutional change. We discuss how in our preferred framework individuals can act both as vehicles for

¹See Ménard and Shirley (2022), ► [Chap. 8](#) for a related discussion of how institutions change, though one with a different emphasis.

exogenous change and in helping select between multiple equilibria. Finally, in our conclusion, we discuss the scope for future work.

31.2 Defining Institutions and Institutional Change

The purpose of this chapter is to review competing theories of institutional change. Different theoretical frameworks serve different purposes, and, as we will go on to show, provide more or less insightful accounts of how institutions change over time.

The first challenge we face is in defining institutions. There are competing definitions of institutions. Different definitions are more or less useful depending on the application.

Many scholars and practitioners work with Douglass North's definition: institutions as rules of the game (North, 1990).² However, as we argue in this chapter, while this definition has been fruitfully employed by applied scholars examining the *effects* of institutions, it poses problems for the study of *institutional change*. To study institutional change, we prefer Greif's perspective (2006) which involves viewing institutions as equilibrium outcomes, jointly determined by the beliefs, strategies, and payoffs of the relevant players. This game-theoretic framing can account for why institutions are sometimes stable and sometimes subject to change.

Nonetheless, the institutions-as-rules (of the game) framework provides a useful starting point for institutional analysis. Describing institutions as rules means that they provide incentives that structure individual behavior. The analogy here is with sports. Different sports have different rules, and these different rules incentivize different behaviors. From the perspective of any individual, who is responding to these rules, they are effectively given or exogenous. Private property rights or the laws that support a system of private property rights are rules that structure individual behavior, such as the decision to invest or to improve a piece of land, in a particular way. A different set of rules, say communal proper rights, provide a different set of incentives.

North's work was predominantly historical. But beginning in the 1990s, many scholars became interested in the contemporaneous relationship between institutions and income per capita. When taking the institutions-as-rules framework to the data, it is natural to think about institutions as explanatory variables. The researcher is typically interested in the "effects of institutions." In the context of a standard regression framework, placing institutions on the right-hand side of a regression equation, suggests that institutions be interpreted as "exogenous variables." Recognition that institutions are not exogenous led to an influential stream of research, notably the classic Acemoglu, Johnson, and Robinson paper on colonial institutions (2001), and Dell's study of the Peruvian Mita (2010). These papers introduced causal methods to the empirical study of institutions. However, the aim of these papers was to obtain causal estimates of how institutions shape incomes. They are not set up to study the process of institutional change itself.

²See, for instance, Koyama and Rubin (2022), ► Chap. 3.

Naturally this institutions-as-rules framework is not designed to study institutional change: if the rules of the game are *given*, how do they change? Indeed, this criticism is not new. Greif (2006) argued that the “rules of the game” perspective was limited precisely because it struggled to explain how institutions change over time. The alternative approach, developed by Aoki (2001) and Greif (2006), suggests viewing institutions from an equilibrium perspective. Greif cautioned against reliance on a single definition noting that “[i]t is sometimes appropriate when examining the issue to consider institutions as exogenous structures; other times they are best considered as endogenous to the interacting individuals. In yet other cases, it is appropriate to study them as reflecting the actions and interests of some individuals but not others” (Greif, 2006, p. 13).³ The definition Greif provides is a general one: “An institution is a system of social factors that conjointly generate a regularity of behavior” (Greif, 2006, p. 30).⁴

In an influential paper, Greif and Laitin (2004) provide a specific explanation of endogenous institutional change. In contrast to the rules-of-the-game perspective, they develop an alternative view of institutions as equilibria (developed further in Greif (2006) and Greif and Kingston (2011)). This perspective emphasizes changing beliefs, motivations and behavior. Institutions cease to be in equilibrium when they induce behavior that is inconsistent with the equilibrium; that is, when they are no longer self-reinforcing. This can be due to exogenous shocks or changes in what Greif and Laitin (2004) call “quasi-parameters,” variables which evolve due to the operation of the institution itself.

Related to this criticism is the idea that institutions can comprise different institutional elements or components. Take, for example, Timur Kuran’s work on Islamic institutions (2010).

Kuran argues that a nexus of interlinking economic and legal institutions combined to delay the transition from personal to impersonal institutions in the Islamic Middle East. For example, he argues that the individualized nature of Islamic law made it difficult for an organizational form like the corporation to emerge. This legal individualism was complemented by the highly egalitarian Islamic inheritance system. By limiting the proportion of an estate that could be passed onto a single heir, Islamic inheritance laws made it likely that most businesses would be dismantled after a single generation. These institutional elements, and others such as the prohibition on charging interest, interlocked and complemented each other. Institutional change in the Islamic world after 1800, (when the potential gains from adopting Western technologies and organization forms became evident) was difficult to achieve, therefore, because it did not involve simply adapting or reforming a single institution but instead, changing a whole set of them.

The question of institutional change is also related to long-standing questions of “persistence” in economic history. This literature has considered both the persistence of institutional legacies and cultural traits (see Voth, 2021, for a review). To inquire: how do institutions change is also to examine those factors that prevent them from changing and cause them to persist.

³For Greif, viewing different definitions as mutually exclusive limits “the advancement of institutional analysis” (Greif, 2006, p. 13).

⁴For additional discussion see Greif and Kingston (2011).

Acemoglu and Robinson (2012) point out that institutions are “sticky.” Attempts to change institutions are often just nominal, but nonetheless institutions do change. To explain this, Acemoglu and Robinson (2012) introduce the concept of critical junctures. In the case of serfdom in Western Europe, for instance, the Black Death can be thought of as introducing one such critical juncture: “a major event or confluence of factors disrupting the existing economic or political balance in society” (Acemoglu & Robinson, 2012, p. 101). The problem, however, is that the authors do not make explicit what constitutes a critical juncture in history. Can the assessment of whether or not a given episode constitutes a critical juncture only be made *ex post*?

North and Thomas (1973) was also an influential contribution to the study of institutions within economic history. More recently, North, Wallis, and Weingast (2009) developed a highly insightful taxonomy of how institutional arrangements within societies cohere. The authors distinguish between various forms of what they call the “natural state” and open-access orders. Both the natural state and the open-access order should be understood as configurations of institutions and as equilibria. In a basic natural state there are no long-lived institutions outside of the state itself which may itself be indistinguishable from the personal identity of the ruler. Natural states therefore are fragile and unstable. Open-access orders, by contrast, permit a wide array of organizational forms.

North, Wallis, and Weingast (2009) do not, however, provide a theory of institutional change. Rather they outline “doorstep conditions” that characterize the threshold at which societies transition from being natural states to being open-access orders. This typology is informative but it does not provide explicit guidance about what factors determine whether a society passes over this threshold. The absence of a compelling account of institutional change was widely noted by commentators. Robert Bates in his otherwise extremely positive review commented that “Concepts, such as ‘double balance’ or ‘door step conditions,’ seemed disembodied and, therefore, unable to provide an intuitively satisfying or logically compelling explanation of why some societies have developed while others have failed to do so” (Bates, 2010, pp. 755–756).

Another conceptual challenge stems from North’s distinction between formal institutions and informal institutions (see Hodgson, 2023a). North often argued that formal institutions had to be complemented with informal constraints (see North, 1990). But this raised a number of questions, in particular: Are informal constraints or informal institutions really institutions? Or is it more productive to bracket them under the heading “culture”? In this chapter, we predominantly focus on formal institutions. It goes without saying, however, that cultural beliefs can act as constraints on what institutional change is possible and that cultural beliefs can co-evolve with institutional development (see Bisin, Rubin, Seror, & Verdier, 2024).

The pace and nature of institutional change are another topic. Institutional change can be gradual and evolutionary, or it can be abrupt. The decline of serfdom in Western Europe was a gradual process that took many decades to unfold. The Glorious Revolution was a distinct and abrupt episode of institutional change (even if we allow that it took several years for aspects of it to be solidified—the Bank of England had to wait until 1694 to be established). Similarly, the institutional changes brought by the French Revolution and subsequent Napoleonic Wars were

confined to a discrete period of time (c. 1789–1815) (Acemoglu et al., 2011). The end of slavery in the United States following the Emancipation Proclamation in 1863 and subsequent Union victory was another important and abrupt institutional change, though the legacies of this victory were contested for decades.

There are also examples that fit the idea of “exogenous” institutional change, that is, institutional change imposed from outside. The imposition of Western colonial institutions is often viewed in this light, for example by Acemoglu, Johnson, and Robinson (2001, 2002) (though it was in fact more complex than this).⁵ However, institutional change is often a highly endogenous process with multiple causes and feedback loops. This makes causal inference challenging.⁶ Of course, many instances of institutional change combine exogenous and endogenous elements. For example, the demise of the serfdom (discussed in Sect. 31.4) is often thought to have been an endogenous response to the exogenous shock of the Black Death (see Jedwab, Johnson, & Koyama, 2022). Dittmar and Meisenzahl (2020) note that before the Reformation period, plague shocks did not directly impact local institutions or politics. However, in the wake of the ideological competition opened up by the Reformation, local plague shocks did lead to institutional change because they discredited the response of local religious establishments.⁷

In what follows, we argue for a game-theoretical approach to institutional change. We will focus on formal institutions, rather than the evolution of informal norms. We will ask what factors led to particular institutions disappearing and new institutional forms emerging. As this is a dynamic and rapidly evolving field of study, we are selective in the literature we review, and focus on recent contributions that are especially relevant to our concerns in this chapter. Before proposing our framework, however, we first discuss an alternative approach found in earlier literature.

31.3 Institutional Functionalism

One way of thinking about institutional change is functionalism. The term functionalism comes from cultural anthropology. Functionalism was perhaps the dominant paradigm in cultural anthropology for several decades after World War II before falling from prominence. It was used as a lens to study both cultural beliefs and

⁵There was no single model of colonialization. While it was true that the initial character English colonies settled in North America were driven by the profit-seeking motives of the Virginia or Plymouth Companies, later imperial settlements, and conquests were often responses to local political conflicts.

⁶This is Przeworski’s criticism of the new institutional economics (2004). Specifically, Przeworski argues that institutions cannot be the primary cause of economic development if they are themselves caused by other factors. He concludes that “Institutions and development are mutually endogenous and the most we can hope for is to identify their reciprocal impacts.” We believe that the approach outlined in this chapter addresses his criticisms.

⁷Desierto and Koyama (2024b) find that (pre-Reformation-era) plague shocks functioned as labor market shocks. They induced elites to pass more sumptuary laws but did not otherwise affect political institutions.

social structures. It allowed scholars to explain the prevalence of particular institutional arrangements, e.g., prohibitions on incest, human sacrifice, slavery, or marriage norms in terms of those factors that would make these institutions functional at a societal level (Radcliffe-Brown, 1952). While it was not always specified, functionalism was understood in terms of improving overall societal welfare. The institutions did not have to be understood in those terms by the individual agents who enacted and enforced them—for example, widespread incest taboos are not necessarily justified by a concern for the viability of offspring; but these “functional” reasons can account for the success and stability of the norm.⁸

Within anthropology, functionalism was criticized for its emphasis on equilibrium. By portraying cultures and social institutions as static, functionalism does not account for how cultural beliefs and social structures form or evolve. Although now largely disregarded by anthropologists, functionalism has had a major impact on how economists think about institutions.

Functionalism approaches offer powerful explanations of why particular institutional arrangements are stable. But can they also provide insights into the process of institutional change?

One angle is evolutionary theory. Alchian (1950) discussed the idea that more efficient institutional arrangements would be selected over time through a quasi-evolutionary process. This idea is also prominent in the later work of F.A. Hayek (e.g., Hayek, 1976, 1982, 1988): history can function as a sieve, eliminating inefficient institutional practices. According to Hayek, it is no coincidence, for instance, that institutions like private property and marriage are so common across human societies, whereas communal property arrangements or forms of “communism” appear rare, transitory, and dysfunctional among large-scale societies whenever they emerge. At the same time, Hayek’s appeal to evolutionary arguments was seen as overly teleological, even by friendly critics (see Vanberg, 2014).

Coase (1960) also influenced how economists thought about institutional formation and institutional change. He focused attention on the critical role of transaction costs in determining the outcome of economic cooperation. When transaction costs are high, the institutions that determine the initial allocation of property rights, shape the outcome of any economic arrangement. Gary Becker, Richard Posner, and Harold Demsetz were responsible for applying these insights to institutions such as the law, property rights, marriage and family law, and even belief systems.⁹

⁸For example, in explaining why in some societies, a sister’s son was allowed privileged relations with a mother’s brother, Radcliffe-Brown set aside the explicit, mythic, or pseudo-historical explanations provided by his interlocutors and instead advanced an “alternative method of dealing with the institution” which was to look “for a theoretical understanding of it as a part of a kinship system of a certain type, within which it has a discoverable function” (Radcliffe-Brown, 1952, p. 13).

⁹Posner (1980) laid the foundations for the economic analysis of many cultural beliefs that we would label superstition. This line of analysis has been fruitfully developed by Leeson (2017). Posner (1980) provides an “equilibrium” analysis and does not consider the problem of institutional change.

An early and insightful application of these ideas was Demsetz (1969) who argued that property rights only emerge when land becomes sufficiently valuable that it pays to enforce these property rights. Hornbeck (2010) provided empirical evidence for Coase and Demsetz. He examines the impact of barbed wire—a technology that lowered the costs of enforcing property rights. Prior to the invention of barbed wire, it was particularly costly to fence and protect land and thus to enforce property rights in areas lacking adequate material for fencing. In these areas, he finds that the introduction of barbed wire had a differential positive impact on land use and productivity.

The formal political economy literature has also provided analyses of endogenous institutional change. Kroszner and Stratmann (1998) argue for, and test, a theory for how interest competition led to the formation of long-standing committees within Congress in the twentieth century US. Special interest groups and legislators face a principal-agent problem because they cannot write binding contracts (as this would be bribery). Kroszner and Stratmann (1998) argue that the development of the committee structure within Congress solves this problem by making interactions between interest-groups and legislators repeated games.

North and Thomas (1973) espoused a form of neoclassical institutional functionalism. They were particularly interested in the question: “why property rights which cannot profitably be established at one point in time will later be economically justified.” Since it was obvious that “the benefits from developing new institutions and property rights must have risen relative to costs,” they sought to answer this question in “an analysis of those parameters which influence the relationships between benefits and costs” (North & Thomas, 1973, p. 8). Institutions that provided the right incentives, particularly what North and Thomas (1973, p. 157) called “efficient property rights,” were critical for economic growth. But these institutions themselves evolved in response to underlying parameters such as technology, transaction costs, and population.

economic growth will occur if property rights make it worthwhile to undertake socially productive activity. The creating, specifying, and enacting of such property rights are costly, in a degree affected by the state of technology and organization. As the potential grows for private gains to exceed transaction costs, efforts will be made to establish such property rights. (North & Thomas, 1973, p. 8).

Implicitly, this argument suggests a one-for-one mapping between underlying economic parameters and institutions. It assumes a world where institutional equilibria are unique.

Institutional functionalism fits naturally with much standard neoclassical economic reasoning. For this reason, many institutional studies take an implicitly functionalist approach. For example, the medieval and early modern craft and merchant guilds were interpreted by many scholars as emerging to solve market failures in the economy. These institutions were local monopolies that restricted entry and usually had some kind of legal or quasi-legal exclusive control of a given industry or trade in a city. Adam Smith had condemned guilds as monopolies, but

institutional functionalism suggests that, to have been so pervasive and long-lasting, they must have solved some important problem or sets of problems.¹⁰

Allen (2011) studies a host of pre-modern institutions through the lens of constrained efficiency. His thesis is that many seemingly odd or corrupt practices such as the sale of offices, dueling, and the patronage system made sense in a world where measuring performance was very costly. Because payment could not be made on the basis of measured output (or input), incentives had to be aligned through other means. For Allen “the major problem of the pre-modern world was the enormous role nature played in the ordinary business of life. Keeping time, measuring distance, and relying on a power were once problems of a first order. Once these problems were solved, new institutions were developed that characterize the modern world” (Allen, 2011, p. 21).

To be clear, there was considerable evidence in favor of such functionalist interpretations. Many of the practices of craft guilds could be rationalized as guaranteeing quality in the face of severe asymmetric information problems (Richardson, 2008). They also helped to incentivize investment in industry-specific human capital (Epstein, 1998). Merchant guilds could be rationalized as providing defense against banditry, as in Hickson and Thompson (1991), and as resolving credible commitment problems with the state, as in Greif, Milgrom, and Weingast (1994).

The claim that institutions that are no longer fit for purpose, perhaps because underlying economic conditions have changed, will be reformed, appears unobjectionable, at least as a necessary condition for institutional change. We call this the weak version of institutional functionalism. Greater conceptual challenges arise when scholars infer efficiency from existence. We label this the strong version of institutional functionalism. It is the working hypothesis that an observed institutional change is an efficient solution to a change in the underlying parameters. The reason why this may not be the case is the existence of multiple equilibria. Ogilvie (2007) similarly criticizes functionalist approaches to institutional analysis (which she brackets under the label “efficiency approaches” in her article “Whatever is, is right?”). She distinguishes between a strong version of the efficiency approach whereby only institutions that are efficient (subject to transaction costs) endure over time, and a weaker version, which allows for path dependency, so that institutions which were initially efficient solutions to social problems persist even after they have lost their initial rationale (Ogilvie, 2007, p. 656).

To take the example of the medieval and early modern craft guilds, simply proposing a functionalist explanation for medieval and early modern guilds does not guarantee that that is the actual reason for their prominence and longevity. Ogilvie (2011, 2019) argues that the practices of both merchant and craft guilds can be alternatively explained with a conflict-based framework.

The prevalence of multiple equilibria means that there is no one-to-one mapping between parameters, or economic fundamentals, and institutional outcomes. Many settings resemble coordination games. Therefore, inefficiency is quite possible. Societies can be “stuck” with inefficient institutions.

¹⁰For a broader discussion, see Koyama (2020).

This is clearly explicated by Young (1998) who offers an evolutionary theory of institutions.¹¹

Young begins with a bottom-up or spontaneous-order view of institutional formation and change: for “a great variety of social and economic institutions—language, codes of dress, forms of money and credit, patterns of courtship and marriage, accounting standards, rules of the road ...no-one willed them into being: they are what they are due to the accumulation of precedent; they emerged from experimentation and historical accident” (Young, 1998, p. 4).

In some settings, for example in bargaining games, there is a tendency for adaptive learning to select “efficient institutions” that is, institutions whose payoffs lie along the Pareto frontier. But more frequently, it is the risk-dominant equilibrium rather than the Pareto-dominant equilibrium that is selected.

A related problem to that of multiple equilibria is that too many social phenomena can be rationalized by rational choice theory if we allow individuals to care about things beyond, say, consumption (see Murphy, 2019). If rational choice theory can justify a wide array of institutional arrangements on efficiency grounds, then it also does not provide much guidance about why certain societies select certain institutional arrangements above others.

Albrecht, Hendrickson, and Salter (2022) argue that many apparent inefficient institutions only appear inefficient if one neglects to examine the costs of changing them. They distinguish between Kaldor-Hicks, or planner inefficiencies, and political inefficiencies, i.e., policies or institutions that are inefficient if we do not include the costs of reform. Evolutionary forces constrain these political inefficiencies: in the long-run even political inefficiencies will be eliminated by a process akin to natural selection (i.e., a society that is rife with political inefficiencies may be conquered or absorbed by a neighbor). This will push institutional change in the direction of greater planner efficiency.¹²

Taking stock, institutional functionalism clearly has its place. As a working hypothesis, it can generate testable predictions that can be taken to the data. As a theory of institutional change, however, it is at best incomplete and often tautologous. Understanding how institutions change requires moving beyond functionalist explanations. In the following sections, we then sketch what we believe is a more complete framework for understanding how and why institutions change over time.

31.4 Conflict and Cooperation as Drivers of Institutional Change

Having considered functionalism as a set of explanations for institutional change, we now consider alternative approaches. Critics of institutional functionalism or efficiency-based explanations often posit conflict as an alternative approach.

¹¹ The underlying logic was already understood by Schelling (1960).

¹² A related discussion about the normative implications of the efficiency perspective is Leeson (2020).

Because rules determine how resources are distributed, they are subject to conflict: “Which institution (or set of institutions) results from this conflict will be affected not just by its efficiency but by its distributional implications for the most powerful individuals and groups” (Ogilvie, 2007, p. 662). Ogilvie (2007, p. 663) argues that

... a ‘conflict’ approach which incorporate institutions into its analysis without assuming such activities to be efficient can explain many facts about pre-modern institutions that ‘efficiency’ views cannot. A conflict approach would agree, for instance, that there is a good economic reason for the long existence of serfdom. But this was not that it efficiently solved market imperfections in public goods, agricultural innovations, or investment. Rather, serfdom created an economy of privileges that hindered efficient resource allocation in labour, capital, and output markets.

An important insight that underlines the conflict theory of institutions is the absence of a political analogue to the Coase theorem (Acemoglu, 2003). The Coase theorem is typically interpreted as saying that in the absence of transaction costs, efficiency will obtain regardless of the initial allocation of property rights (Coase, 1960). Acemoglu (2003) argues that this reasoning cannot usually be generalized to the political setting. The main reason for this is that contracts that are enforceable across all states often cannot be written. Agents in the political arena cannot make credible commitments.

Acemoglu and Robinson (2000, 2006) applied a game-theoretic framework to study democratization. Democracy emerges in the context of political conflict between the rich elites and the masses. Elites concede democratic rights in response to the threat of revolution from below because otherwise their promises of economic redistribution would not be credible.

While the role of conflict has received attention in the literature, it is not the only driver of institutional change. While individuals and groups can fight over the distribution of resources and surplus, they can also cooperate to preserve and grow that surplus.

Indeed, scholars have theorized about the origins of the state and political order more generally through two distinct analytical frameworks—contractual theories, and approaches based on predation or conflict.

Contractual theories focus on the cooperative nature of institutional change in that they explain how a particular political organization could be justified by *voluntary* consent. However, as they are typically intended as normative theories, they are not accurate descriptive accounts of how states arose. Hobbes’s version of the state of nature and his contractual justification for absolutism is merely the most famous of these (Hobbes, 1996). Locke used a similar framework to articulate a constitutional regime, as did other seventeenth and eighteenth century social contract theorists such as Pufendorf and Vattel (Tuck, 1979).

Conflictual accounts provide purported positive theories for the emergence of the state. Karl Marx developed such a conflictual account, as did Friedrich Nietzsche and Franz Oppenheimer. Oppenheimer saw the genesis of the state in the act of violence such as “subjection of peasant folk by a tribe of herdsmen or sea nomads” (Oppenheim, 1922, p. 29). Indeed, as a positive description of state formation it

seems accurate: many and perhaps most actual cases of state genesis involve violent conquest. This is inscribed in the formation myths of the Romans, many of the polis of ancient Greece and such modern nations as England and France (formed by the Germanic invaders in the fifth and sixth centuries).

Of course, we do not have historical accounts of the rise of the earliest states, so to understand cases of pristine state formation, we remain reliant on rational reconstruction as informed by both theory and whatever archaeological evidence can be brought to bear on the question.

The idea that both cooperation and conflict are critical to understanding state formation was already noted by North (1979, p. 251):

The contract approach may explain why the state potentially can provide a framework for economizing on the use of resources and therefore can promote wealth. But as both the third party to every contract and the ultimate source of coercion, the state becomes the field on which the battle for control of its decision-making power is fought. All sides wish to be able to redistribute wealth and income in the interest of their own group. This was Madison's main point in Federalist Paper No. 10, written 200 years ago. The contract theory explains the gains of initial contracting (in comparison to anarchy) but does not explain the subsequent maximizing behavior of constituents with diverse interests. The predatory theory ignores the initial gains of contracting and focuses on the extraction of rents from constituents by those who gain control of the state.

In other words, cooperation generates surplus, which can then be expropriated through conflict.

We selectively survey some recent scholarship that demonstrates the forces of conflict and cooperation during important historical episodes of political and economic development.

31.4.1 Early States

A key idea for coercion-based theories of the state is geographical circumscription—the idea that geographical features such as deserts or mountain ranges surrounding a piece of fertile land limited the outside options of would-be subjects (Carneiro, 1970). Allen (1997) used this argument to examine state formation in Egypt. Egypt's early state formation is suggested by its geography. Agriculture first developed along the fertile Nile valley, and this area of lush farmland was surrounded by inhospitable desert. Geographical circumscription was thus possible earlier than elsewhere as its geography helped rulers corral tax payers. Early statehood in Egypt brought stability, but may have made the average Egyptian worse off, as the surplus created by agriculture could be extracted by the political elite.

Mayshar, Moav, and Neeman (2017) and Scott (2017) both emphasize the importance of the ability of a state to ascertain and appropriate agricultural output. Standard accounts of the emergence of the state, link this process to agricultural productivity. As agriculture developed and became more productive, a surplus emerged that could be used to support political elites. Mayshar et al. (2017)

challenge this account. They argue that the ability of political elites to extract a surplus from the farming population depended on their ability to expropriate. In other words, it relied on their military superiority and their ability to accurately observe how much output had been produced.

What about an efficiency-based explanatory for early state formation? According to this perspective, the emergence of states can be rationalized as a response to a collective action problem. Allen, Bertazzini, and Heldring (2023) study the formation of early states in Mesopotamia (modern day Iraq). As agriculture in this region (and period) was dependent on natural irrigation from rivers (rather than being rain-fed), they argue that a shift in the location of a river created an exogenous source of variation in demand for public goods provision and hence a “state.” They find that a shift of a river away from a location is associated with a three percentage point increase in the probability of that location being part of a city state. This, the authors argue, provides support for cooperative rather than conquest-based theories for the origins of the state.

31.4.2 The Demise of Serfdom in Western Europe

In 1300, 50% of people in England were servile and 50% of the land was held on villein tenure, but by 1500 serfdom had largely disappeared. Land was held through a variety of different forms of copyhold (Bailey, 2014, p. 28). Restrictions and feudal duties associated with the possession of land had also mostly gone.

Serfdom was a long-standing legal and economic institution. It was characterized by a coercive relationship between lord and serf. Villeins were legally obliged to hold land and pay labor service and customary dues. Lords had corresponding obligations to their serfdoms but these could be breached and, at least in theory, lords had the discretionary power to impose additional incidents and fees (Bailey, 2014, p. 37). How serfs were actually treated varied locally (“Sale of serfs, and arbitrary punishments and restrictions upon them, were unusual”). In theory, lords were able to extract the entire surplus beyond that required for subsistence in the form of labor dues or fines. In practice, however, many labor dues were commuted.

The disappearance of serfdom in England was part of a wider European phenomenon that has attracted considerable scholarly attention. Domar (1970, p. 21) outlined a simple economic model of serfdom that generates a trilemma: one cannot have abundant land, an aristocracy that lives off labor income, and free labor. Which combination prevails depends on political factors or institutions. Domar believed that his story fit Russia fairly well but he recognized that it did not explain Western Europe. Domar (1970, p. 28) noted that

...when we come to the depopulation caused by the Black Death after 1348, my hypothesis is of little value in explaining the subsequent cause of events. Why did serfdom fail to come back after such a sharp increase in the land/labor ratio? ...I have none to offer and have to fall back on political factors

Historians writing in the Marxian tradition did offer an explanation based on political factors.

Rodney Hilton emphasized class conflict and the growing self-assertiveness of peasants following the Black Death (Hilton, 1969, 1985). While the initial response of lords was repression, in the long-run peasant resistance brought about an improvement in labor market conditions. The Peasants Revolt of 1381 was decisive in this development according to this perspective.

The timing of this account seems incompatible with what Bailey (2014) observes from the study of 28 English manors from East Anglia, Oxfordshire, and Buckinghamshire. Bailey finds that serfdom began a sharp decline immediately following the Black Death. Labor services were already being commuted from 1300 onwards but the rate at which this occurred accelerated after 1350. Soon after 1400 labor service mostly disappeared (barring seasonal work requirements which lasted until the 1450s). He concludes that villein tenures disappeared between 1380 and 1420 while personal serfdom disappeared between 1380 and 1450.

The question then is what caused this decline? North and Thomas (1973), North (1990) suggested an important factor: changing relative prices. But they noted that the effects of changing relative prices might be complex. Acemoglu and Wolitzky (2011) propose a principal/agent model that addresses this topic. In their model, coercion increases effort by the agent. Greater demand for labor (because the value of marginal land has increased or perhaps because of the invention of a new technology) increases coercion. This model, therefore, recovers Domar's logic that greater labor scarcity can lead to more labor coercion (a labor demand effect) but it embeds it in a richer framework. Specifically, Acemoglu and Wolitzky (2011) allow for an outside option effect. If greater economic scarcity also increases outside options for workers (by raising per capita income and hence demand for manufactured goods), then this causes labor coercion to fall. According to this framework, therefore, if the outside option effect dominates the labor demand effect, then a negative demographic shock can cause a decline in labor coercion.

These predictions are broadly consistent with what historians observe happening after 1348. Nonetheless, they abstract away from the interplay of cooperation and conflict among elites.

The demographic shock of the Black Death provided a crucial background condition. Figure 31.1 depicts the collapse in population that took place after 1348 and its subsequent failure to recover until after 1500. Real wages initially did not rise (because prices rose faster than nominal wages). But as Fig. 31.2 also shows, from the 1380s onwards, wages began to grow, peaking in the early or mid-fifteenth century. Many economic historians have viewed this period as a golden age for workers. The evidence for this is mixed as we have to consider the decline in international trade and a decline in hours of work available. Bailey (2021, p. 254) cautions "against highly optimistic claims for the living standards of workers c. 1400" nonetheless he notes that it "seems reasonable to suppose that the average annual income within the lower orders of society rose by c.40 percent."

Nonetheless, the effects on the allocation of land were immediate. Black Death mortality increased the turnover of free peasant land: "the epidemics increased the

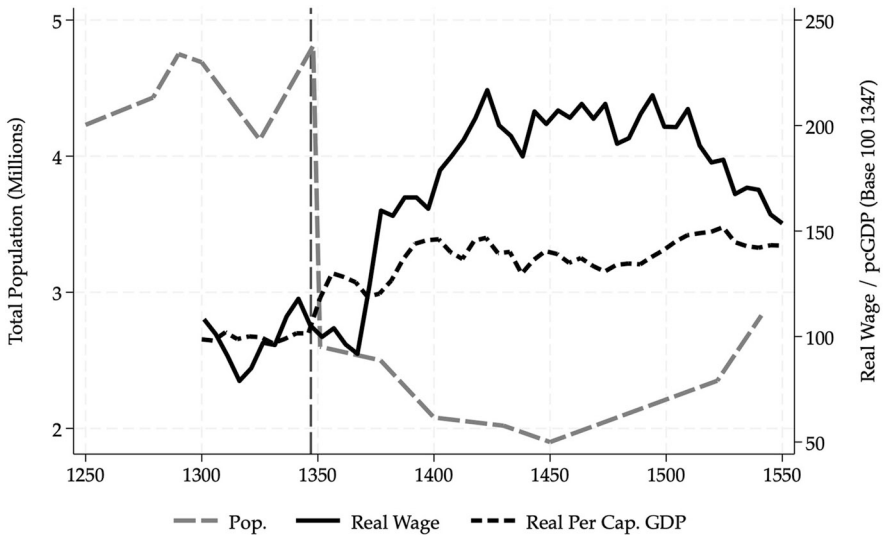


Fig. 31.1 The demographic shock of the Black Death. Population estimates are from Broadberry et al. (2015)

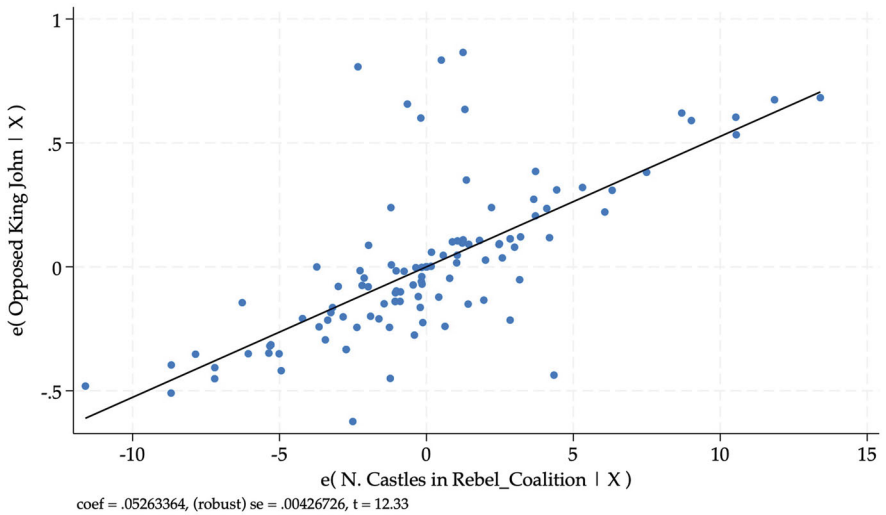


Fig. 31.2 The positive relationship between the number of castles owned by rebels in a baron’s family network and that baron’s probability of rebelling against King John. (Data source: Desierto, Hall, and Koyama 2024)

incidence of the failure of heirs, and the length of time land remained in the lord’s hands without a tenant, creating even more opportunities and incentives for land-lords to change the terms of tenure” (Bailey, 2021, p. 305)

Rather than being a response to conflict between peasantry and lords, Bailey argues that the evidence is consistent with labor scarcity inducing workers to move.

Peasants abandoned farms where conditions were unsatisfactory and moved where conditions were better. This rather than manumission—the contractual lifting of servile obligations—or peasant resistance drove the decline in serfdom.

The decline of serfdom was also linked to the rise of the state (Peters, 2022).¹³ Palmer (1993) argued that the late fourteenth century saw a dramatic increase in the scope and in particular the legal capacity of the English state. He argued that “Governance after the Black Death was qualitatively different, exhibiting a government intent on using the law to control society, to preserve as far as possible the status quo.” Crucial to Palmer’s argument were the attempts by the English state to regulate labor markets. “The enforcement of labor regulation was still primarily local, but local power was now more effectively exercised by lords representing central authority over all available workers instead of by lords exercising their own inherent authority over their unfree tenants” Palmer (1993, p. 5). In a similar fashion, Bailey (2021, p. 300) argues that individual lords were not well equipped to respond to the new challenges posed by the radically altered land/labor ratio.

This example reiterates the importance of political and institutional factors. Had landlords been a sufficiently cohesive class, they could have coordinated on limiting this movement. Indeed, they attempted to do so for some decades in the late fourteenth century. But ultimately the incentives to defect were too great and attempts to restrain peasant demands collapsed.

31.4.3 Institutional Change in Early Modern England

Since North and Weingast (1989), scholars of institutions have focused on England as a canonical example of institutional change (e.g., Pincus, 2009; Pincus & Robinson, 2014; Cox, 2011, 2015; Acemoglu & Robinson, 2012; Hodgson, 2023b). We revisit this episode from the perspective of recent contributions to the literature and from the broader sweep of English history.¹⁴

Historians caution that the Glorious Revolution should be seen in the context of a gradual process of institutional change in which Parliament became more integral to the governance of England over time. Its origins go back to the thirteenth century (Butt, 1989). Angelucci, Meraglia, and Voigtländer (2022) examine the consequences of grants given to certain towns to govern themselves and collect their own taxes in the twelfth and thirteenth centuries. They find that towns that received such grants went on to obtain parliamentary representation. Moreover, it was these towns that in the seventeenth–nineteenth centuries played a role in strengthening Parliament and pushing for more inclusive institutional change (for example, by raising regiments to fight for Parliament in the Civil War or voting for the Great Reform Act of 1832).

¹³ Belich (2022) also emphasizes the importance of the plague for the development of Europe’s warfare states.

¹⁴ A vast literature emerged in response to North and Weingast (1989). The majority of this literature, however, focused on whether the consequences of the Glorious Revolution were as dramatic and discontinuous as North and Weingast (1989) originally argued.

The Reformation was another decisive shock to England's political institutions. Rubin (2017), Greif and Rubin (2024) argue that since the Tudors had come to power during a civil war, they could not rely on heritage and tradition, while the Reformation removed the Catholic Church as a source of legitimacy. In response the Tudors relied increasingly on Parliament. This is one concrete instance of how cultural change, in this case religious change, could bring about institutional change.

According to Greif and Rubin (2024), different principles of legitimation and different legitimating agents after 1520 set England on a particular institutional trajectory. Besides relying more on Parliament to enact laws, the Tudors increased the number of Members of Parliament and increased the number of constituencies. This set in motion the conflicts that would occur in seventeenth century England and result in civil war and ultimately in the Glorious Revolution.

Other factors also played an important role. The growth of Atlantic trade empowered merchants and their allies in Parliament, as argued by Acemoglu, Johnson, and Robinson (2005). Jha (2015) examines the role of financial markets in facilitating institutional change. He argues that access to shares in colonial trading ventures allowed a subset of elites to coordinate in opposing King Charles I during the English Civil War. This is part of a broader research agenda in which Jha proposes that financial instruments can help foster broad political coalitions that favor peace and beneficial reforms (see Jha, 2018; Jha & Shayo, 2019). Political parties also emerged in this era (see Jones, 1961; Pincus & Robinson, 2014) and as argued by Desierto and Koyama (2024a), they played an important role in allowing elites to coordinate on specific policies.

The Glorious Revolution was undoubtedly an important landmark in the institutional development of England. It placed lasting limits on the discretionary authority of the Crown and established parliamentary sovereignty. But it was itself the product of many decades of conflict and political deals (cooperation) between members of the political elite and the Crown.

31.5 A Theory of Cooperation and Conflict

Our historical examples demonstrate that some institutional changes result from conflict, and some from cooperation. But under what conditions do individuals voluntarily come together, or else fight with one another, to bring about institutional change? We think that existing scholarship struggles to answer this, because it has yet to provide a clear theoretical framework that *simultaneously* depicts cooperation and conflict. The problem is primarily methodological. Conflict is naturally modeled within a non-cooperative framework, while cooperation fits within a cooperative, game-theoretic framework. To depict both conflict and cooperation thus entails finding a way to merge these two distinct frameworks.

A promising avenue is found in Ray (2007), in which the process by which players come to an agreement is modeled as a non-cooperative game. The cooperative game approach tends to focus on coalitions as fundamental units that can block or approve proposals. Ray (2007) instead focuses on the formation of such coalitions as a result of non-cooperative bargaining among proposers and responders.

Desierto and Koyama (2024) and Desierto, Hall, and Koyama (2024) use this framework to depict the political economy of feudalism and the emergence of the Magna Carta in England. We summarize below.

31.5.1 Feudalism

The rise of feudalism can be seen as involving a form of secondary state formation as new types of polities were fashioned in the wake of the barbarian invasions of the early Middle Ages. The feudal system provides an illuminating illustration of the importance and simultaneous interplay of both cooperation and conflict in establishing order. A key feature of the feudal world is that there were no political institutions that confer power to a ruler and guarantee the stability of their rule. Instead, the ruler had to continuously hold together a coalition of elites, each of whom has their own resources and can, at any point, withdraw from the coalition, peacefully or violently (through rebellion).

Desierto and Koyama (2024) thus depict a feudal order as a coalition of elites held together by non-binding agreements, which may be completely voluntary or otherwise backed up by the threat of violence. Individual elites each have their own economic and military resources which they can pledge to a ruler in exchange for some share of the entire realm. This however, need not always be a peaceful affair. The ruler can forcibly take those resources, and bring its owner into the ruler's coalition through conquest. Any elite can also exit the coalition by withholding or removing their resources from the ruler's jurisdiction through rebellion. It is the forging and re-forging of the ruler's coalition of elites that describes how a feudal realm expands and consolidates into larger territories, or fragments into smaller units. One important result is that consolidation is more likely when the resources of elites are difficult to appropriate by any ruler, e.g., when an elite has a castle that can defend their surrounding lands and therefore keep them away from the ruler should the elite rebel. This makes rebellion more costly to the ruler, which makes the ruler more likely to prevent rebellion by sharing more wealth with elites. The elites then remain loyal, and the realm is consolidated.

31.5.2 Magna Carta and Institutional Change in Medieval Europe

Constraining the grabbing hand of the state is a key theme in institutional economics. With some exceptions, less attention has been paid to the early growth of these institutions in medieval Europe.¹⁵ The issuing of Magna Carta in 1215 was one of the important early instances of institutional change in English history, a critical juncture in the language of Acemoglu and Robinson (2012). While it by no means fully succeeded in its own time, as King John reneged on the initial agreement, and

¹⁵For example, Salter and Young (2023). There is a rich empirical literature on medieval city-states, see for instance, Stasavage (2014). Puga and Treffer (2014) considers the case of institutional change in medieval Venice as it moved from being a relatively open oligarchy to a closed one.

further baronial wars were required to make subsequent kings respect the limitations on their authority, Magna Carta did lay down the foundation for the subsequent development of Parliament as a forum for the king to ask for taxes and consult with the realm (Koyama, 2016). Desierto, Hall, and Koyama (2024) study the baronial coalition whose rebellion brought about the Magna Carta.

Their analysis is both theoretical and empirical. Using the coalition-formation framework in Desierto and Koyama (2024), the authors consider how an exogenous shock to the power or legitimacy of the king can generate the formation of a sub-coalition of rebel barons, who then propose a new arrangement that limits the extractions by the king. Specifically, the rebels propose a new institution like the Magna Carta, which restricts the king's discretion over resources and ensures that barons receive equal rights. The king can agree to the proposal, or reject it, in which case he gathers his loyalists to fight the rebels. The victor thereafter implements their preferred arrangement—either the status quo or the rebels' proposal. An agreement like Magna Carta is thus obtained in equilibrium if it is an optimal contract between parties (e.g., the king and the barons) who are each capable of violence. The authors show that this equilibrium is more likely to occur when the king is highly extractive; when the barons have resources that are non-appropriable by the king and therefore easily withdrawn from the king's coalition; and when the initial distribution of resources among the barons is fairly egalitarian.

Underlying these results is a key theorem stating that any one baron is more likely to join the rebel coalition when the total (non-appropriable) resources of that rebel coalition are large. Desierto, Hall, and Koyama (2024) demonstrate whether this theorem holds empirically, as a test of the plausibility of the theory. To do this, they build a unique dataset comprising the universe of all English barons in the early thirteenth century and each of their family networks. They match each baron to their barony and thus have a measure of each baron's resources. The relevant resources are those that are non-appropriable, that is, those resources a baron can take back from the king should the baron rebel. One measure of the non-appropriability of resources is its defensibility. The authors thus count the number of castles held by every member of the rebel coalition to serve as proxy for the non-appropriable resources of the rebel coalition. The theorem predicts that this should increase the likelihood that any one baron joins the rebel coalition.

One obvious challenge in identification is to find variation in the resources of the rebel coalition across baronies, as there is only one rebel coalition. The authors then prove a stronger theorem stating that the resources of a *subset* of the rebel coalition—specifically those who are in the baron's social network, would increase the likelihood that the baron joins the rebel coalition.

To proxy for the baron's social network, the authors construct the family network of each baron. They then identify all the rebels in the network, and count all the castles held by these rebels.

The authors find that the number of castles held by the rebels in the baron's family network is positively associated with that baron also joining the coalition against King John (Fig. 31.2). To address possible endogeneity concerns, the authors instrument the number of castles held by rebels in the baron's family network with the number of castles held by the ancestors of those rebels within the baron's (current) family network.

31.6 Critical Junctures and the Role of Individuals

Functionalist explanations of institutional change, whereby institutions change when circumstances render their functions obsolete or inefficient (subject to the costs of changing institutions), leave little room for individuals. Institutions emerge to eliminate social waste because as Olson put it “big bills aren’t often dropped on the sidewalk” (Olson, 1996, p. 3). In contrast, the role of individuals in pushing institutional change is explicable in the game-theoretic framework we outlined above.

Individual actors can be conceptualized as potential sources or vehicles of exogenous change. Such individuals are rare, though as we discuss below, there are credible examples of such individuals, including Martin Luther or Napoleon. Alternatively, specific individuals can be thought of as playing a role in solving collective action problems such as coordination games. We provide examples below.

There are, however, many examples of institutional change that do not rely on the characteristics of specific individuals. In the case of the Magna Carta crisis, no single rebel baron appears to have been pivotal. The nominal leader of the rebellion, Robert FitzWalter was not himself critical as there were several other potential rebel leaders. Similarly, while some individuals were prominent in the English Peasant’s Rebellion such as Watt Tyler, serfdom would have declined in England regardless of any individual. On the other hand, other instances of institutional change seem to have been driven in part at least by specific individuals.

Martin Luther is often seen by historians as a decisive figure in sparking the Reformation.¹⁶ Becker, Hsiao, Pfaff, and Rubin (2020) provide evidence for the critical role of Martin Luther in diffusing Protestantism. In particular, it was Luther’s network of correspondents that facilitated the rapid spread of his ideas after 1517. Luther’s ties to local elites help to explain how the Reformation became instituted across Germany within a short time period. Luther can be seen as performing a dual role. He delivered an external shock to the legitimacy of the medieval Church. He also served to coordinate and articulate previously disparate dissatisfaction with Church practice.

We touched upon the importance of the Reformation for institutional change in England above. Many historians suggest that Henry VIII was decisive in the Reformation’s success in England. Had he not wanted or needed a divorce from Catherine of Aragon, it seems unlikely that the small Protestant movement in England would have succeeded.¹⁷

In some cases, groups of individuals may have played an important role in driving institutional change. Another celebrated instance of institutional change was the French Revolution. The majority of work by economists on the French Revolution has considered its consequences (see, for example Acemoglu et al., 2011) and less attention has been paid to its causes, though of course a vast historical literature has investigated this topic.

¹⁶A case could be made that Frederick the Wise, the Elector of Saxony who backed and sheltered Luther was in fact an individual who made a decisive difference to history.

¹⁷Henry VIII of course did later expropriate large resources from the monasteries (Desierto, Koyama and Shera 2024). This, however, was not part of his initial incentive to begin the Reformation.

Jha and Wilkinson (2023) study one of many intriguing links between the American and French Revolutions. Specifically, they examine the idea of revolutionary contagion among French soldiers who served in the American Revolutionary War. The argument is that exposure to liberal and revolutionary ideas in North America radicalized these soldiers and helped them to organize and that they then played a critical revolution during the Revolution in France.

To identify these effects Jha and Wilkinson (2023) exploit quasi-random variation in which regiments were sent to America and which served elsewhere in the war. Specifically, from Comte Rochambeau's original force, two regiments the Neustrie and Anhalt regiments were not sent to America. Comparing local areas with Rochambeau's veterans (i.e., those who actually served with him) with those from these regiments allows the authors to estimate the impact of serving in America. They find that places with more Rochambeau veterans were more in favor of democratic institutional change on the eve of the Revolution, more likely to have political clubs, and experienced more riots and unrest after 1789.

A final example of how individuals can matter for institutional change comes from the US Civil War (1860–1865). Dippel and Heblich (2021) look at the impact of the “Forty Eighters” on antislavery and pro-Union sentiment in the United States. The Forty-Eighters were individuals involved in the liberal 1848 revolutions that shook central Europe and who subsequently migrated and settled in the United States. Many of them became involved in antislavery mobilizing after the Kansas-Nebraska Act of 1854 and many were involved in encouraging enlistment in the Union Army.

Dippel and Heblich (2021) put together a dataset of almost 500 Forty Eighters. Using a matching approach, they find that the presence of the Forty Eighters increased enlistment into the Union army by around 2/3. They find evidence that the Forty-Eighters were also associated with civilian leadership and that their presence helped to deter desertions. While Dippel and Heblich (2021) frame their argument in terms of how individuals can help form social movements, we prefer to interpret it as speaking to the role individuals can have in bringing about institutional change.

31.7 Concluding Comments

This chapter has provided a survey of how recent scholarship has approached the question of institutional change. Institutional change is a challenging topic for social scientists to study. Many of our empirical tools are intended to study the effects of institutions; they attempt to isolate an exogenous component of institutions in order to estimate causal effects. By design, they cannot speak to the endogenous forces that cause institutions to adapt and change over time.

We have argued that the most fruitful way to approach the question of institutional change must allow for both conflict and cooperation and that game-theory offers a particularly useful framework. Interpreting institutional change through an efficiency lens has provided numerous insights but it is limited in cases where there are multiple possible equilibria, not all of them efficient.

Understanding institutional change also requires detailed institutional and historical knowledge. Each empirical setting is distinctive. Many of the papers we survey

introduce novel, highly disaggregated, datasets to provide much more granular analysis of how institutions change than were possible previously.

The topic of institutional change is of critical importance for understanding economic history and economic development. Recent work, particularly empirical work, has allowed us to make significant progress in explaining certain important episodes of institutional change, such as the decline of serfdom, Magna Carta, the Reformation, or the French Revolution. Nonetheless, much work remains to be done and many aspects of institutional change remain to be uncovered.

Acknowledgment of Funding Open-access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D. (2003). Why not a political Coase theorem? Social conflict, commitment and politics. *Journal of Comparative Economics*, 31(4), 620–652.
- Acemoglu, D., & Robinson, J. A. (2000). Why did the west extend the franchise? Democracy, inequality, and growth in historical perspective. *The Quarterly Journal of Economics*, 115(4), 1167–1199.
- Acemoglu, D., & Robinson, J. A. (2006). *The economic origins of dictatorship and democracy*. Cambridge University Press.
- Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail*. Crown Business.
- Acemoglu, D., & Wolitzky, A. (2011). The economics of labor coercion. *Econometrica*, 79(2), 555–601.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *The American Economic Review*, 91(5), 1369–1401.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2002). Reversal of fortune: Geography and institutions in the making of the modern world income distribution. *The Quarterly Journal of Economics*, 117(4), 1231–1294.
- Acemoglu, D., Johnson, S., & Robinson, J. (2005). The rise of Europe: Atlantic trade, institutional change, and economic growth. *American Economic Review*, 95(3), 546–579.
- Acemoglu, D., Cantoni, D., Johnson, S., & Robinson, J. A. (2011). The consequences of radical reform: The French revolution. *American Economic Review*, 101(7), 3286–3307.
- Albrecht, B. C., Hendrickson, J. R., & Salter, A. W. (2022). Evolution, uncertainty, and the asymptotic efficiency of policy. *Public Choice*, 192(1), 169–188.
- Alchian, A. A. (1950). Uncertainty, evolution, and economic theory. *Journal of Political Economy*, 58, 211–211.
- Allen, R. C. (1997). Agriculture and the origins of the state in Ancient Egypt. *Explorations in Economic History*, 34(2), 135–154.
- Allen, D. W. (2011). *The institutional revolution*. Chicago University Press.
- Allen, R. C., Bertazzini, M. C., & Heldring, L. (2023). The economic origins of government. *American Economic Review*, 113(10), 2507–2545.
- Angelucci, C., Meraglia, S., & Voigtländer, N. (2022). How merchant towns shaped parliaments: From the Norman conquest of England to the Great Reform Act. *American Economic Review*, 112(20), 3441–3487.
- Aoki, M. (2001). *Toward a comparative institutional analysis*. MIT Press.
- Bailey, M. (2014). *The decline of serfdom in late medieval England: From bondage to freedom*. Boydell and Brewer.
- Bailey, M. (2021). *After the Black Death: Economy, society, and the law in fourteenth-century England*. Oxford University Press.

- Bates, R. (2010). A review of Douglass C. North, John Joseph Wallis, and Barry R. Weingast's "Violence and social orders: A conceptual framework for interpreting recorded human history". *Journal of Economic Literature*, 48(3), 752–756.
- Becker, S. O., Hsiao, Y., Pfaff, S., & Rubin, J. (2020). Multiplex network ties and the spatial diffusion of radical innovations: Martin Luther's leadership in the early Reformation. *American Sociological Review*, 85(5), 857–894.
- Belich, J. (2022). *The world the plague made: The Black Death and the rise of Europe*. Princeton University Press.
- Bisin, A., Rubin, J., Seror, A., & Verdier, T. (2024). Culture, institutions and the long divergence. *Journal of Economic Growth*, 29, 1–40.
- Broadberry, S., Campbell, B. M. S., Klein, A., Overton, M., & van Leeuwen, B. (2015). *British economic growth, 1270–1870*. Cambridge University Press.
- Butt, R. (1989). *A history of parliament*. Constable and Company Limited.
- Carneiro, R. L. (1970). A theory of the origin of the state. *Science*, 169(3947), 733–738.
- Coase, R. H. (1960). The problem of social cost. *The Journal of Law and Economics*, 3(1), 1.
- Cox, G. W. (2011). War, moral hazard, and ministerial responsibility: England after the Glorious Revolution. *The Journal of Economic History*, 71(1), 133–161.
- Cox, G. W. (2015). Marketing sovereign promises: The English model. *The Journal of Economic History*, 75(1), 190–218.
- Dell, M. (2010). The persistent effects of Peru's mining Mita. *Econometrica, Econometric Society*, 78(6), 1863–1903.
- Demsetz, H. (1969). Information and efficiency: Another viewpoint. *Journal of Law and Economics*, 12(1), 1–22.
- Desierto, D., & Koyama, M. (2024). *Feudal political economy*. Economic Theory. <https://doi.org/10.1007/s00199-024-01583-8>
- Desierto, D., & Koyama, M. (2024a). *The birth of parties*. Unpublished manuscript.
- Desierto, D., & Koyama, M. (2024b). The political economy of status competition: Sumptuary laws in preindustrial Europe. *Journal of Economic History*, 84(2), 479–516.
- Desierto, D. A., Hall, J., & Koyama, M. (2024). *Magna Carta* (GMU Working Paper in Economics No. 23-34). GMU.
- Desierto, D. A., Koyama, M., & Shera, M. (2024, December). *Rents and reformation*. Manuscript.
- Dippel, C., & Hebllich, S. (2021). Leadership in social movements: Evidence from the "forty-eighters" in the civil war. *American Economic Review*, 111(2), 472–505.
- Dittmar, J. E., & Meisenzahl, R. R. (2020). Public goods institutions, human capital, and growth: Evidence from German history. *Review of Economic Studies*, 87(2), 959–996.
- Domar, E. D. (1970). The causes of slavery or serfdom: A hypothesis. *The Journal of Economic History*, 30(1), 18–32.
- Epstein, S. R. (1998). Craft guilds, apprenticeship, and technological change in preindustrial Europe. *The Journal of Economic History*, 58(3), 684–713.
- Greif, A. (2006). *Institutions and the path to the modern economy*. Cambridge University Press.
- Greif, A., & Kingston, C. (2011). Institutions: Rules or equilibria? In N. Schofield & G. Caballero (Eds.), *Political economy of institutions, democracy and voting* (pp. 13–43). Springer.
- Greif, A., & Laitin, D. D. (2004). A theory of endogenous institutional change. *American Political Science Review*, 98(4), 633–652.
- Greif, A., & Rubin, J. (2024). Endogenous political legitimacy: The Tudor roots of England's constitutional governance. *Journal of Economic History*, 84(3), 655–689.
- Greif, A., Milgrom, P., & Weingast, B. R. (1994). Coordination, commitment, and enforcement: The case of the merchant guild. *Journal of Political Economy*, 102(4), 745–776.
- Hayek, F. (1976). *Law, legislation, and liberty: Vol. II, The mirage of social justice*. University of Chicago Press.
- Hayek, F. (1982). *Law, legislation and liberty* (Vol. III). Routledge.
- Hayek, F. (1988). *The fatal conceit*. Routledge.
- Hickson, C. R., & Thompson, E. A. (1991). A new theory of guilds and European economic development. *Explorations in Economic History*, 28(2), 127–168.
- Hilton, R. (1969). *The decline of serfdom in medieval England*. Palgrave Macmillan.

- Hilton, R. H. (1985). A crisis of feudalism. In T. Ashton & C. Philpin (Eds.), *The Brenner debate, agrarian class structure and economic development in pre-industrial Europe* (pp. 119–137). Cambridge University Press.
- Hobbes, T. (1996). *Leviathan* (R. Tuck, Ed.). Cambridge University Press.
- Hodgson, G. (2023a, October). *Formal and informal institutions: Some problems of meaning, impact and interaction*. Draft Manuscript.
- Hodgson, G. (2023b). *The wealth of a nation*. Princeton University Press.
- Hornbeck, R. (2010). Barbed wire: Property rights and agricultural development. *The Quarterly Journal of Economics*, 125(2), 767–810.
- Jedwab, R., Johnson, N. D., & Koyama, M. (2022). The economic impact of the Black Death. *Journal of Economic Literature*, 60(1), 132–178.
- Jha, S. (2015). Financial asset holdings and political attitudes: Evidence from revolutionary England. *Quarterly Journal of Economics*, 130(3), 1485–1545.
- Jha, S. (2018). Trading for peace. *Economic Policy*, 33(95), 485–526.
- Jha, S., & Shayo, M. (2019). Valuing peace: The effects of financial market exposure on votes and political attitudes. *Econometrica*, 87(5), 1561–1588.
- Jha, S., & Wilkinson, S. (2023, September). *Revolutionary contagion*. Manuscript.
- Jones, J. (1961). *The first Whigs: The politics of the exclusion crisis, 1678–1683*. Oxford University Press.
- Koyama, M. (2016). The long transition from a natural state to a liberal economic order. *International Review of Law and Economics*, 47, 29–39.
- Koyama, M. (2020). A review essay on the European guilds. *The Review of Austrian Economics*, 33(1), 277–287.
- Koyama, M., & Rubin, J. (2022). *How the world became rich*. Polity Press.
- Kroszner, R. S., & Stratmann, T. (1998). Interest-group competition and the organization of congress: Theory and evidence from financial services' political action committees. *The American Economic Review*, 88(5), 1163–1187.
- Kuran, T. (2010). *The long divergence*. Princeton University Press.
- Leeson, P. T. (2017). *WTF*. Stanford University Press.
- Leeson, P. T. (2020). Economics is not statistics (and vice versa). *Journal of Institutional Economics*, 16(4), 423–425.
- Mayshar, J., Moav, O., & Neeman, Z. (2017). Geography, transparency and institutions. *American Political Science Review*, 111(3), 622–636.
- Ménard, C., & Shirley, M. M. (2022). *Advanced introduction to new institutional economics*. Edward Elgar.
- Murphy, R. H. (2019). The rationality of literal tide pod consumption. *Journal of Bioeconomics*, 21(2), 111–122.
- North, D. C. (1979). A framework for analyzing the state in economic history. *Explorations in Economic History*, 16(3), 249–259.
- North, D. C. (1990). *Institutions, institutional change, and economic performance*. Cambridge University Press.
- North, D. C., & Thomas, R. P. (1973). *The rise of the Western world*. Cambridge University Press.
- North, D. C., & Weingast, B. (1989). Constitutions and commitment: The evolution of institutions governing public choice in seventeenth century England. *Journal of Economic History*, 49, 803–832.
- North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.
- Ogilvie, S. (2007). 'Whatever is, is right'? Economic institutions in pre-industrial Europe (Tawney Lecture 2006). *Economic History Review*, 60(4), 649–684.
- Ogilvie, S. (2011). *Institutions and European trade: Merchant guilds, 1000–1800*. Cambridge University Press.
- Ogilvie, S. (2019). *The European guilds: An economic analysis*. Princeton University Press.
- Olson, M. (1996). Big bills left on the sidewalk: Why some nations are rich and others poor. *Journal of Economic Perspectives*, 10, 3–24.
- Oppenheim, F. (1922). *The state*. B.W. Huebsch.

- Palmer, R. C. (1993). *English law in the age of the Black Death, 1348–1381: A transformation of governance and law*. North Carolina Press.
- Peters, M. E. (2022). Government finance and imposition of serfdom after the Black Death. *European Review of Economic History*, 27(2), 149–173.
- Pincus, S. (2009). *1688 The first modern revolution*. Yale University Press.
- Pincus, S. C., & Robinson, J. A. (2014). What really happened during the Glorious Revolution? In S. Galiani & I. Sened (Eds.), *Institutions, property rights, and economic growth: The legacy of Douglass North* (pp. 192–222). Cambridge University Press.
- Posner, R. A. (1980). A theory of primitive society, with special reference to law. *Journal of Law and Economics*, 23, 1–53.
- Przeworski, A. (2004). The last instance: Are institutions the primary cause of economic development? *European Journal of Sociology*, 45(2), 165–188.
- Puga, D., & Trefler, D. (2014). International trade and institutional change: Medieval Venice's response to globalization*. *The Quarterly Journal of Economics*, 129(2), 753–821.
- Radcliffe-Brown, A. (1952). *Structure and function in primitive society*. The Free Press.
- Ray, D. (2007). *A game-theoretic perspective on coalition formation*. Oxford University Press.
- Richardson, G. (2008). *Brand names before the industrial revolution* (Working Paper 13930). National Bureau of Economic Research.
- Rubin, J. (2017). *Rulers, religion, and riches: Why the west got rich and the Middle East did not*. Cambridge University Press.
- Salter, A. W., & Young, A. T. (2023). *The medieval constitution of liberty: Political foundations of liberalism in the west*. University of Michigan Press.
- Schelling, T. (1960). *The strategy of conflict*. Harvard University Press.
- Scott, J. C. (2017). *Against the grain: A deep history of the earliest states*. Yale University Press.
- Smith, A. (1776). *An inquiry into the nature and causes of the wealth of nations*. Clarendon Press.
- Stasavage, D. (2014). Was Weber right? The role of urban autonomy in Europe's rise. *American Political Science Review*, 108, 337–354.
- Tuck, R. (1979). *Natural rights theories: Their origin and development*. Cambridge University Press.
- Vanberg, V. J. (2014). Darwinian paradigm, cultural evolution and human purposes: On F.A. Hayek's evolutionary view of the market. *Journal of Evolutionary Economics*, 24(1), 35–57.
- Voth, H. J. (2021). Persistence: Myth and mystery. In A. Bisin & G. Federico (Eds.), *Handbook of historical economics*. Academic.
- Young, H. P. (1998). *Individual strategy and social structure*. Princeton University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Commitment, Coercion, and Markets: The Nature and Dynamics of Institutions Supporting Exchange

32

Avner Greif

Contents

32.1	Introduction	812
32.2	Contract-Enforcement Institutions	814
32.2.1	Organic, Private-Order Institutions	817
32.2.2	Public-Order CEIs: The Legal and Regulatory Systems	821
32.2.3	Designed, Private-Order Institutions	824
32.3	Coercion-Constraining Institutions	831
32.3.1	Balancing Coercive Powers	833
32.3.2	A Ruler with an Independent Coercive Power	835
32.3.3	A Ruler with a Conditional Coercive Power	836
32.3.4	Balancing Coercive Power with Economic Power	837
32.3.5	The Origin and Dynamics of CCIs and Political Institutions	840
32.4	The Dynamics of Market-Supporting Institution and the Implied Dynamics of Markets and Political Institutions	844
32.4.1	The Demand for Public-Order and Designed CEIs	845
32.4.2	The Supply of Public-Order and Designed, Private-Order CEIs	846
32.4.3	Market Expansion and Political Development	849
32.5	Reflections on the Historical Development of Markets and Political Institutions	849
32.6	Concluding Comments	859
	References	862

Abstract

The development of market-based exchange relies on the support of two institutional pillars that are, in turn, shaped by the development of markets: (1) “Contract-enforcement institutions” determine the range of transactions in which individuals can commit to keep their contractual obligations. (2) “Coercion-constraining” institutions constrain those with coercive power from abusing others’ property rights and influence whether individuals will bring their goods to the market in the first place. Markets and political institutions co-evolve,

A. Greif (✉)
Stanford University, Stanford, CA, USA
e-mail: avner@stanford.edu

reflecting the dynamic interplay between coercion-constraining and contract-enforcement institutions. This chapter reviews the literature on contract-enforcement institutions and defines and elaborates on coercion-constraining institutions. It provides a tentative theory of the dynamics of market-supporting and political institutions and draws on history to illustrate the merits of this conjecture.

Keywords

Contract enforcement · Coercion-constraining institutions · Market supportive institutions · Political development

32.1 Introduction

Markets rest upon institutions. The development of market-based exchange relies on the support of two institutional pillars that are, in turn, shaped by the development of markets. Research in the field of new institutional economics has largely focused upon one such institutional pillar—"contract-enforcement institutions"—that determine the range of transactions in which individuals can commit to keep their contractual obligations. Yet, markets also require institutions that constrain those with coercive power from abusing others' property rights. These "coercion-constraining" institutions influence whether individuals will bring their goods to the market in the first place.

This chapter's discussion of market-supporting institutions is geared toward the issues we know the least about. First, the dynamics of market-supporting institutions and the implied dynamics of markets; second, the inter-relationships between the dynamics of market-supporting and political institutions where the latter comprise the rules for collective decision-making, political rights, and the legitimate use of coercive power. It argues, in particular, that neither the assertion that liberal political institutions lead to markets nor that markets lead to liberal governance are supported by theory or history. Markets and political institutions co-evolve through a dynamic interplay between contract-enforcement and coercion-constraining institutions.

Many successful market economies have prevailed in the past; there were adequate market-supporting institutions. Early successes, such as those in the Islamic world or China, were not indicators of later development. It was the commercial expansion that began in Europe during the late medieval period that led to the development of markets that support the complex, dynamic modern economy with its wide-scale reliance on impersonal exchange. Why did not early success lead to subsequent market expansion? More generally, what does determine the dynamics of market expansion? Addressing these questions is a key to understanding the "Rise of the West," the operation of market economies, and the factors that still hinder market development.

The argument advanced here is that markets can rest on different combinations of contract-enforcement and coercion-constraining institutions. Different combinations, in turn, can support distinct sets of exchange relationships, implying that

their relative efficiency depends on the details of the related economy. In particular, under conditions elaborated upon below, markets can prosper even in the absence of limited government and the rule of law. Equally important, different initial combinations of contract-enforcement and coercion-constraining institutions lead, in a non-deterministic way, to distinct dynamics of markets and political institutions.

The core idea developed here regarding this dynamic is the following. Contract-enforcement institutions *organically* (spontaneously) emerge in the initial stages of market development as unintended and unforeseeable results from the pursuit of individual interests. Yet, even in the same economic situation, different institutions can emerge as their details are influenced by various factors, including those that are cultural and social (e.g., Greif, 1994a; McMillan, 2002). The details of these initial contract-enforcement institutions influence which additional institutions the economic agents will find it profitable to use if made available and establish if possible. *Ceteris paribus*, initial contract-enforcement institutions influence what additional, *designed* (intentionally created) institutions the economic agents will “demand.”

The ability to effectively supply designed—private- or public-order—contract-enforcement institutions, depends on the prevailing coercion-constraining institutions, which are those that influence decisions regarding the acquisition and use of coercive power. This is the case because many designed institutions reveal information about wealth to those with coercive power. Wealth-revealing, designed, contract-enforcement institutions will be utilized only if coercion-constraining institutions are such that this information does not undermine the security of property rights.

Consider, for example, public-order, contract-enforcement institutions. When appealing to the court, using a land registry, applying for a business license, or submitting to a regulatory agency information regarding one’s wealth is generated. This information can be used by those with coercive power to identify and capture this wealth. Wealth-revealing public-order institutions can be *effectively* supplied only if the generation of this type of information does not lead to wealth confiscation. If this is not the case, even if public-order, contract-enforcement institutions are established, they will be underutilized. Distinct coercion-constraining institutions imply distinct abilities for *effectively* supplying various, designed, contract-enforcement institutions that further extend the market.

Interestingly, the coercion-constraining institutions conducive to the growth of the market also likely to lead to the endogenous emergence of political institutions associated with liberal societies in which market participants (the “commercial sector”) have political representation and influence. Political institutions—rules for collective decision-making, political rights, and the legitimate use of coercive power—that are actually followed are self-enforcing in the sense that following them is each political actor’s best response (e.g., Barzel, 2002; Greif, 1994b, 1998; Hardin, 1989; Weingast, 1997). An important determinant of a political actor’s decision to follow a political rule depends on his relative coercive power (e.g., Acemoglu & Robinson, 2000; Bates, 2001; Bates et al., 2002; Downing, 1992; Greif, 1994b, 1998). Because coercion-constraining institutions influence decisions regarding the acquisition and use of coercive power, they impact the set of self-enforcing political institutions.

In particular, the coercion-constraining institutions conducive to the effective supply of designed, wealth-revealing, contract-enforcement institutions are those in which the commercial sector has coercive (and economic) power that countervails the coercive power of others, such as rulers. Under certain conditions, the fear of risking a costly retaliation by the commercial sector would induce those with coercive power to prefer consulting it rather than taking unilateral actions affecting property rights. The commercial sector's power induces other political actors to provide it with a political voice through such means as political representation. Political voice precedes the establishment of, in particular, designed public-order institutions because it is required for the commercial sector to communicate the need for particular public-order institutions.

The views that market development requires appropriate political institutions (e.g., North, 1990; Weingast, 1997) or that political development follows the expansion of markets (e.g., Lipset, 1959) are too simplistic. Markets and political institutions co-evolve, reflecting the dynamic interplay between coercion-constraining and contract-enforcement institutions. It is no coincidence that the modern market economy and the liberal state jointly emerged.

The above analysis builds on and integrates elements of Old and New Institutionalism. Old Institutionalism (e.g., Menger, 1963 [1883]) emphasizes the distinctions and inter-relationships between organic (spontaneous) and designed (pragmatic) institutions; new Institutionalism emphasizes studying the micro-foundation of contract-enforcement institutions (e.g., Williamson, 1985; Greif, 1989, 1993) and the inter-relationships between the polity and the economy (e.g., North & Thomas, 1973). Finally, the discussion here highlights the importance of understanding coercion-constraining institutions, which influence the development of both markets and polities. It thus builds on, and contributes to, the emerging literature on the institutional foundations of social order (e.g., Bates, 2001; Bates et al., 2002; Greif, 1994a, b, 1998, *Forthcoming*).

This chapter is organized as follows: Sect. 32.2 reviews the literature on contract-enforcement institutions, whereas Sect. 32.3 defines and elaborates on coercion-constraining institutions. Once the nature of these two types of institutions is clarified, Sect. 32.4 provides a tentative theory of the dynamics of market-supporting and political institutions. Section 32.5 draws on history, particularly that of China and Europe, to illustrate the merits of this conjecture.

The theoretical conjecture and historical analysis presented in Sects. 32.4 and 32.5 are very rudimentary. Many aspects of the theory have yet to be worked out and additional important issues—such as agency problems within the polity—have yet to be integrated. Yet, this analysis makes an explicit conjecture regarding market expansion that links the institutional foundations of the market and the polity and is derived from history.

32.2 Contract-Enforcement Institutions

The extent of the market—the degree of voluntary exchange—is determined by its supporting contract-enforcement institutions (CEIs). CEIs determine the transactions to which one can credibly commit to fulfill contractual obligations and therefore the

exchange relationships into which economic agents will enter. CEIs determine who can exchange with whom and in what goods. A market extends when the available set of CEIs increases, allowing a greater number of people to enter into more exchange relationships in more situations. Studying the relative efficiency of markets and their dynamics requires examining their CEIs and their development.

CEIs are required to support markets because exchange is almost always sequential. Sometime elapses between the *quid* and the *quo*, providing one with the ability to renege.¹ This sequentially implies the fundamental problem of exchange: A necessary condition for exchange is for all sides to credibly commit to adhere to their contractual obligations, to the extent that each expects to be better off than refusing to exchange. One will not enter into an otherwise profitable exchange relationship unless the other party can commit *ex-ante* (when a decision is made whether or not to exchange) to fulfill his contractual obligation *ex-post* (when he can renege).

Institutionalists have so far concentrated on contract-enforcement institutions that link conduct in current exchange with future payoffs in a way that makes it known *ex-ante* that the best one can do *ex-post* is not to renege.² To illustrate the mechanism, suppose it is commonly known that failure to pay a debt implies future inability to borrow, and that a borrower values the gains from future credit more than those from reneging on his debt contract. Because the best a borrower can do *ex-post* is to pay his debt, he can *ex-ante* commit to doing so. The value of future economic exchange is placed as a bond to be lost in case of breach. One can similarly commit when it is known *ex-ante* that failure to pay a debt implies sufficiently large legal sanctions.

Although various CEIs differ in their details, their effectiveness depends on mitigating the same problems of making the threat of sanctions (or rewards) credible. Those who are to apply the sanctions should have the appropriate information regarding past conduct and the incentives to neither shirk their duty nor abuse their power. The offender should be precluded from fleeing to avoid sanctions and the sanctions should be sufficiently high to deter breach. The parties should also share expectations regarding what behavior constitutes a breach and that should be commonly known that the above conditions are met. The details of CEIs make a difference, however, as they determine who will be able to credibly commit in what

¹E.g., Greif, 1997, 2000; Aoki, 2001; Dixit [Forthcoming](#). Williamson (1985) stressed that sunk, relationships specific investment are important in generating sequentially. But, as discussed in Greif, 2000 exchange is sequential, for example, in financial transactions, in labor relationships, in agency relationships, and in the exchange of experience, goods whose attributes are revealed only with use. Indeed, even in spot exchange it is usually technologically impossible for the *quid* and the *quo* to be laterally exchanged simultaneously. The only exception is exchange in which goods with objectively known attributes are exchanged simultaneously.

²This is common in works focusing on the law (e.g., Williamson, 1985) or reputation mechanisms (e.g., Greif, 1989, 1993) to the exclusion of institutions based on internalized norms and intrinsic motivation.

exchange relationships. An objective and impartial court is effective only in exchanges in which it can *ex-post* verify conduct.

In considering the relative merits of various CEIs, it is useful to group CEIs according to whether they are “organic” or “designed” and whether they are “private-order” or “public-order.”³ According to Carl Menger (1883), organic (spontaneous) institutions emerge as an unintended and unforeseeable results of the pursuit of individual interests. Designed (pragmatic) institutions reflect intentional and conscious design and possibly the coordinated responses of many individuals. The former roughly corresponds to what North (e.g., 1990) defines as informal institutions and the latter, formal institutions. Private-order institutions, the importance of which has been emphasized by Williamson (1985), rely mainly on economic and social sanctions imposed by economic agents, while public-order institutions rely mainly on sanctions imposed by the state. In either of these cases, we can either study institutions as self-enforcing or not.⁴ In studying self-enforcing institutions, we attempt to study as endogenous the behavior of all relevant agents, including, such as judges, priests, or police officers.

In studying CEIs, the economics of information, contract theory and mechanism design have been extensively employed (Furubotn & Richter, 1997 provide a useful survey). Game theory, however, has been found particularly useful because it exposes the conditions necessary for threats and promises to be credible. It is common to model the contracting environment as some version of either a One-Sided Prisoners’ Dilemma game or Prisoner’s Dilemma.⁵ But the flexibility

³More generally, institutions differ according to the associated sanctions: whether they are economic, social, or coercive (in the form of legal sanctions or physical assaults); or by who imposes the sanctions: the interacting economic agents or a third party (either another economic agent such as a trade association or a non-economic agent, such as a legal agency). Actual institutions, as discussed below, often transcend these simple dichotomies. Some combine economic and coercive sanctions imposed by various agents. Organic institutions often evolve to acquire designed components while customs are often codified as laws.

⁴Self-enforcing institutions have been studied under various headings. Relational contracting (in which one’s conduct in a particular situation reflects considerations regarding its implications on the entirety of the relationship over time and situations; e.g., McMillan & Woodruff, 2000). Social norms (which are rules of behavior that are regularly adhered to although they are not legally enforced; e.g., Kandori, 1992). Self-governance (in the sense that the responsibility for contract enforcement is placed in the hands of the interacting individuals; e.g., Dixit, 2003a) and reputation-based institutions (in which one’s current conduct is motivated by fear of losing one’s reputation and hence future gains; e.g., Greif, 1989, 1997, Forthcoming.)

⁵One-Sided Prisoners’ Dilemma game (OSPD) (Greif, 1989, 2000) is known as the Game of Trust (Kreps, 1990). One player can initiate the relationship and if he has done so, the second player gets to decide whether to cheat or not. Cheating is more profitable to the second player but expecting cheating the first player is better off by not initiating the relationship. Greif (2000) argues that the commitment problem associated with exchange is well captured by this game. In the Prisoners’ Dilemma game (PD) both players move sequentially, each can choose to either cooperate (be honest) or defect (cheat). Defecting is a dominant strategy (it is the best action for each player whether the other cooperates or cheats). If both cooperate, they are better off than if they both defect (E.g., Kandori, 1992; Ellison, 1994.) Hodgson, 1998 reviews the use of evolutionary game theory to study CEIs.

of game theory permits the introduction of various assumptions, regarding, for example, the matching process among the potential parties to exchange or the information available to them, and whether the division of the surplus from exchange should be taken as given or not.⁶

In interpreting these theoretical analyses, however, it is important to recognize that they only expose the conditions required for particular behavior to be an equilibrium outcome and its implications. Empirical analysis, however, is needed to evaluate whether and how these conditions were fulfilled in a particular episode and how this situation became commonly known. This implies considering how the game (or, more generally, the environment) that was assumed in the theoretical analysis was generated.

32.2.1 Organic, Private-Order Institutions

In organic, private-order CEIs the credible threat by the economic agent(s) to impose sanctions deters breach (henceforth, I will refer to “organic, private-order CEIs” as “organic CEIs”). Organic CEIs are likely to emerge when economic agents face the prospect of beneficial, ongoing exchange while breaching a contract is (actually or statistically) observable. The credibility of one’s threat to terminate a relationship following a breach is not much of an issue. People are willing to spend resources to punish cheaters (e.g., Fehr & Gächter, 2000) and people often expect that past cheaters will cheat again. In either case, the effectiveness of reputation-based deterrence—the range of situations in which it can support exchange—increases as the value of future relationships become higher. It therefore increases, *ceteris paribus*, with the per-period gains from exchange, and the economic agents’ patience, and decreases in their alternative income outside the relationship. Organic CEIs are therefore more likely to emerge when the parties are locked into their relationships: the market is thin, and it is costly to find a new partner with whom to exchange. McMillan and Woodruff (2000) tested this proposition in contemporary Vietnam and indeed found that where the closest alternative trader is far away, more credit was granted.

The cost of terminating relationships can endogenously emerge as a part of the institutions. Suppose that exchange is supported only by the above bilateral reputation in which one is punished only by the person whom he cheated. Suppose that the one who cheated has to wait until another relationship is naturally dissolved before

⁶Greif (1989, 1993), for example, considers the implications of transferable utility (where the payoff distributions are endogenous to the interacting parties) and non-random matching. Milgrom et al. (1990) and Dixit (2003a) examined the implications of asymmetric games in which the players are not symmetric in their actions sets). Kranton (1996a), Ghosh and Ray (1996), and Kali (1999), Dixit (2003a) considered adverse selection situations in which players are of different types in the sense that some are less likely to cheat. Clay (1997a, b) examined the incomplete monitoring situation in which one can mistakenly believe that the other cheated.

he establishes a new relationship. If the implied cost of waiting is sufficiently high, a bilateral reputation mechanism can support exchange.⁷

The above discussion assumes a moral hazard situation. All agents are identical, and each faces the choice of whether to breach his contract or not. In reality, however, the economic agents have different unobservable characteristics that determine how likely they are to breach a contract. Some may be more patient than others or have better unobserved outside opportunities. In such adverse selection situations, where some agents are of a “bad” type and hence more likely to cheat, bilateral reputation can operate even if it is technologically costless to immediately find an alternative partner to an exchange. Because some agents are bad types, each agent is motivated to “test” new partners in exchange to discover their types. One “builds” a relationship by initially trading small and gradually increasing the stakes or one demands that his partner bear a sunk cost (e.g., giving gifts or wasting time) at the beginning of their relationship. Such initial sunk cost increases the cost of breach (Watson, 2002; Kranton, 1996a, b; Ghosh & Ray, 1996). In contemporary Africa, suppliers of input and credit were indeed found to initiate exchange in small amounts and gradually increase them (Fafchamps, 2004).

Bilateral relationships provide only limited and often costly contract enforcement. They require market imperfections to insure that the value of the exchange is higher than can be achieved by exchanging elsewhere. Each bilateral exchange may require the costs of building relationships or other sunk costs required to establish trust. Organic institutions based on bilateral reputation mechanisms require that one is able to directly monitor the performance of the other. Hence, these institutions are effective in exchanges such as that of goods where quality is easy to verify, or in credit relations. They are ineffective, however, in an exchange where outcomes are uncertain and not directly observable (e.g., in agency relations).

Organic, multilateral reputation institutions can support exchange in a wider range of situations than is possible under a bilateral reputation mechanism. More behavior can be monitored, better information circulated, and higher sanctions would be imposed on those who cheat. These institutions are based on information flows regarding past conduct among many economic agents and on economic sanctions imposed by those who were not cheated. Each individual is willing to exchange information with others, expecting to get information in return. Because economic agents have more to lose by revealing valuable information to competitors, the incentive to share information is higher if this is not the case.

The literature identifies three reasons why an individual might find it optimal to participate in a multilateral punishment against someone who did not cheat him. First, in moral hazard situations, the expectation that others will sanction an individual can be enough to motivate punishment, because punishment by others reduces the ability to perform and reduce loss due to breach (Greif, 1989, 1993;

⁷MacLeod & Malcomson, 1989 present a comprehensive analysis. In equilibrium, the unemployed agent cannot establish a new relationship by offering to, say, sell his product for less than the market price because at such lower prices it is not credible that one will not cheat.

Fafchamps, 2004). Second, in adverse selection situations, past cheating reveals bad types who will cheat again (e.g., Kranton, 1996a; Ghosh & Ray, 1996).⁸ Third, individuals participate in multilateral punishment according to an internalized sense of fairness among members of a social structure (Fehr & Fischacher, 2004).

Because organic CEIs based on multilateral punishment are not designed, they often have symbiotic relationships with social structures, norms, and cultural beliefs (e.g., Greif, 1994a). These social features provide the initial conditions for the rise of the economic institution which, in turn, reproduced these social features. Information flows, personal familiarity within such social structures as networks, communities, business groups, and religious groups, make multilateral (community or collective) punishment possible. Shared norms and beliefs and their representation in customs and merchants' laws coordinates on the actions that constitute cheating is provided by.

Individualistic searching for reliable business partners can also lead to the formation of a social structure in the form of a business network or an "old boys" network (Kali, 1999; Fafchamps, 2004). In either case, the gains implied by the economic institution to each of the members of the social structure motivate retaining affiliation with the social structure and hence, reproduce it. An economy can therefore end up segregated in the sense that exchange, beyond spot exchange, is conducted among members of groups based on non-economic attributes such as religious sects, lineages, etc. In this case, the overlay of an economic institution on a social structure also implies that social sanctions and social, cultural, racial, or religious exclusion supplement, or even replace, the disciplinary impact of economic sanctions.

To illustrate institutions based on a multilateral reputation mechanism, consider the one that governed agency relationships among the eleventh century (Jewish) Maghribi traders who were engaged in long-distance trade all over the Muslim Mediterranean (Greif, 1989, 1993). Among them, it was efficient—ignoring contractual problems—to operate through overseas agents rather than having each merchant travel abroad with his goods. But because agents could cheat while handling a merchant's capital abroad, to be employed they had to be able to *ex-ante* commit to be honest *ex-post*, after the goods were sent to them. Among the Maghribis, as arguably among others in similar situations, this commitment was achieved based on a multilateral reputation mechanism within a business network by an institution that can be referred to as a coalition.

The Maghribis employed each other as agents, shared the expectations that Maghribi merchants would hire only Maghribis agents, and that all of them would cease employing an agent who had cheated. Information flows within their commercial and social network enabled detection and circulation of information about

⁸Kandori (1992) and Ellison (1994) pointed out the possibility that in random matching PD games, a multilateral punishment can be supported by having the one who is being punished play cooperate when the other punishes him by playing cheat. Punishing is thus profitable to the one who punishes. One cooperates in his own punishment, in turn, because doing so implies that the punishment will cease after a finite number of periods.

cheating. Coordination on multilateral punishment was facilitated by a set of cultural rules of behavior—a Merchants' Law—that specified how an agent should act to be considered honest in circumstances not mentioned in a merchant's instructions.⁹ False accusations of cheating were curtailed by the extensive use of witnesses to testify to one's honesty. Multilateral punishment was self-enforcing because each merchant, expecting others to punish, found it in his best interest to punish as well.

Multilateral punishment enhanced efficiency and profitability relative to bilateral punishment (in which only the trader who was cheated retaliated), since it enabled the employment of agents even when the relationship between a specific merchant and agent pair was not expected to recur. The resulting additional gains from cooperation, the value of the information flows, and the expectations concerning future hiring ensured the "closedness" of the coalition. Merchants were motivated to employ only member agents while agents were motivated not to seek employment elsewhere because being a coalition member was profitable. Although wages were lower than under a bilateral punishment, employment was more certain. Finally, agents were motivated not to cheat in their old age, fearing that their children's reputations would suffer.

Social and economic sanctions often inter-related in organic institutions and this interrelationship can reflect strategic manipulation by individuals pursuing their own interests. This complexity is well illustrated in the study of contract-enforcement institutions in Mexican California conducted by Clay (1997a, b).¹⁰ This analysis touches upon two neglected issues in institutional economics: the interplay between social and economic sanctions and multi-tier institutions (although see Dixit, 2003b; Greif, 2004).

Social relationships were central to a contract-enforcement institution in the Mexican communities but there was no institution that enabled their members to commit to buy goods for credit from long-distance American traders. Among these traders, contract enforcement was achieved based on a coalition-like institution and economic sanctions. The two institutions were linked, however, to create a composite—two tiers—institution that enabled Mexicans to commit to pay their debts. This institution emerged when individual American traders created a link between the Mexican intra-community institution and the American traders' coalition. These individuals integrated into the communities by marrying local girls, converting to Catholicism, and settling in the villages. Hence, they gained access to the intra-community contract-enforcement institution. These American traders also retained their membership in the American traders' coalition. An American who settled in a community used the intra-community, social-based institution to collect debt from locals and he could also commit to not to breach a contract with the American trader

⁹ Rules that make the meanings of various actions common knowledge are central to any institution based on multilateral or third-party punishment. See Greif, 1993 for a discussion of economic institutions and Hardin, 1989 and Weingast, 1997 for political institutions.

¹⁰ For empirical analyses of the interplay between social sanctions and economic institutions, see Ellickson, 1991; Landa, 1994; Bernstein, 1992; and Rauch, 2001.

who provided the loan by the threat of being excluded from the American traders' coalition.

Efficiency-enhancing organic CEIs are not necessarily optimal given the existing monitoring, information, and enforcement technology as lack of coordination hinders adjustments. Each individual behaves optimally given the constraints and opportunities offered by others' behavior and expected behavior, but whenever the environment changes in a way that makes different behavior optimal, there is no mechanism to adjust expectations regarding others' behavior. This expected behavior therefore continues to influence choices thereby hindering adjustments. The efficiency of organic CEIs is therefore higher in a static economy where rules governing exchange and punishments do not have to adjust frequently.

Similarly, the efficiency of organic CEIs is higher when it is less important to adjust the number of interacting individuals. The optimal size of an organic CEI with multilateral punishment is theoretically linked to the speed of information transmission. The optimal size balances the benefit of greater participation with the cost implied by the delay in punishment due to the additional time required to transmit information. But the actual number of individuals governed by the institution reflects either the original social group around which the institution emerged, or individualistic search for worthy partners. In either case, the size may not be optimal. In the presence of adverse selection, organic CEIs based on multilateral punishment can be welfare-reducing as they lower the incentives among non-members to search for worthy partners (Greif, 1993; Kali, 1999; Annen, 2003).

Moreover, organic CEIs based on multilateral punishment reduce incentives to create institutions that would foster impersonal exchange and support the needs of a dynamic economy (Greif, 1994a; Kranton 1996b). More generally, the relative efficiency of organic CEIs declines as populations and markets grow. These institutions are relation-based and hence entail low fixed costs (they do not require special organization or other systemwide special investment) but high and rising marginal costs. The expansion of a business entails exchange where one has successively weaker relational links with the other party to the exchange. Expansion thus requires investment in building relationships and carries a greater risk of collapse (Li, 1999).

32.2.2 Public-Order CEIs: The Legal and Regulatory Systems

Designed CEIs can support exchange that organic CEIs cannot and often reflect a social or private response to gains from providing additional contract enforcement. They reflect intentional and conscious design and possibly the coordinated responses of many individuals and therefore are characterized by explicit rules that regulate and coordinate membership, behavior, processes for changing rules, and intentionally created organizations, such as guilds, courts, credit rating companies, escrow companies, and business associations.

A social response to the needs for additional contract-enforcement manifests itself in public-order CEIs, the legal and regulatory systems. Public-order CEIs coordinate behavior through laws and regulations, collect and process information using various

formal procedures and bureaucratic organizations and deter contractual breaches by threatening legal and regulatory sanctions. They can be particularly effective in influencing behavior because they rely on the authority of the state. They can, for example, employ coercive power to punish violators, impose various auditing and supervisory requirements, and one to appear in court and reveal information.

The legal system constitutes a third party that alters the costs of a contractual breach, enabling parties to commit to a contractual performance in cases in which contracts will not be self-enforcing based on the value of future relationships.¹¹ In addition, the law can foster the operation of organic CEIs. Boot et al. (1993) argued that legal contractual incompleteness can be used to signal reputation while Johnson et al. (2002) empirically found that in post-communist countries the law is used to reduce the cost and time required to build new relationships that are later sustained by reputation.

Compared to the costs associated with organic CEIs, the public-order institutions that support modern markets require high fixed costs. Large legislative, judicial, administrative, and coordination costs are required to establish the system and render it effective and credible in the context of a highly mobile, non-agrarian economy. Once established, the marginal cost of enforcement—and hence exchange relationships—is low and constant (e.g., Li, 1999). Indeed, until recently economists studied market economies assuming that public-order CEIs provide costless, perfect, objective, and impartial contract enforcement.

In reality, however, the legal system rarely provides costless, perfect, objective, and impartial contract enforcement (e.g., Williamson, 1985; Ellickson, 1991; Greif & Kandel, 1995). Public-order institutions that best approximate this situation operate in a few advanced contemporary countries and only in recent times. We know surprisingly little, however, regarding the institutional development that led to these modern successes.¹² But even in these cases, the operation of the law as a contract enforcer is restricted by various factors. For a court to enforce contracts, for example, it should be able to *ex-post* verify actions taken by the litigants and their impact on observed outcomes.¹³ Verification of past actions is often costly, particularly in labor relationships and in complex business transactions and production processes.

¹¹ Public-order institutions have also been successful in creating markets by compiling and making public the necessary information. Formal land titling and land registrars, for example, underpin mortgage markets in developed economies (Arruñada, 2003). Approval by the FDA is central to the trade in drugs.

¹² See Greif, 2001, 2004 regarding endogenously providing incentives for a partial law to provide impartial justice. Klerman, 2007 summarizes various theses and emphasizes competition among courts of law; Klerman & Mahoney, 2004 elaborate on the importance of freeing the legal system from political intervention.

¹³ In general, modern legal systems do not collect information in commercial cases. This has not always been the case. In late medieval Venice, the authorities *ex-ante* collected the information required to verify the conduct of agents in long-distance trade ventures (Gonzalez de Lara, 2002). I am not familiar with an analysis of the optimal scope of *ex-ante* or *ex-post* information collection by the legal system.

Apart from the need for verifiable information, other factors lessen the effectiveness of the legal system. States have a limited geographical scope while laws and regulations may be designed to achieve various policy objectives other than securing rights. Budget constraints and administrative capacity imply that legal proceedings may be time consuming. Direct legal costs, such as legal fees, and indirect costs, such as the opportunity cost of time, are often high. Economic agents' strategic responses to the incentives implied by laws and regulations can limit their effectiveness (e.g., Townsend, 1979). Polinsky and Shavell (2000) surveyed the literature considering how legal rules can be designed to maximize social welfare subject to these constraints. In the models they surveyed, it was implicitly assumed that the prevailing legal tradition is the European tradition of man-made, explicit law. In societies with other, religious, and customary legal traditions, different models may be needed.

Further limiting the operation of public-order institutions is the need to mitigate the associated agency problem. Decision-makers within public institutions, such as judges, police officers, and regulators, have to be provided with the incentive to use their power to protect rather than abuse property rights or otherwise reduce property-rights security. Judges can be bribed or take advantage of the fact that it is difficult to measure the quality of their services. The prevalence of corruption in much of the world testifies to the magnitude of this problem (Rose-Ackerman, 1999; Shleifer & Vishny, 1993).

The difficulty of providing appropriate incentives to judges and regulators often reflects their concern about their personal safety following an unfavorable judgment. Powerful members of a society can use coercive power and other means to obstruct justice and circumvent regulations (Glaeser & Shleifer, 2003). Making judges and regulators free of political control and relying relatively more on regulations rather than laws are ways of mitigating this problem. Yet, relaxing the ability to discipline judges can lead to more corruption. Similarly, regulatory agencies can be established to propagate the policy of legislators beyond these legislators' term rather than promoting welfare (e.g., Weingast, 1996).

These limitations of public-order institutions imply, in particular, that organic CEIs are more efficient when information is known but not verifiable, when the speed of resolving contractual disputes is important, and when the issues are too complex for the court to grasp at a low cost or when the issues require particular knowledge that the court lacks. Indeed, Macaulay's (1963) seminal work reveals the large extent to which organic CEIs govern contractual relationships in the contemporary USA (See also Bernstein, 1992). Where and when public-order institutions are ineffective due to corruption or budgetary constraints, the relative profitability of private-order institutions increases. Indeed, the informal sector is disproportionately larger in developing economies and countries in transition (Portes, 1994, p. 438; De Soto, 1989, p. 12, 131; Fauchamps et al., 1993; Greif & Kandel, 1995).

When effective public-order institutions exist, economic agents can respond to their limitations for fostering contract enforcement by appropriately structuring their contractual relationships, property-rights distribution, and organizational forms. Williamson (1985) has emphasized the importance of private-order institutions operating in the shadow of the law. Hostage-taking, vertical integration, and corporate governance are examples of such structuring (See Williamson, 1985;

Hart & Moore, 1999; Maskin & Tirole, 1999; Tirole, 2001. Dixit (Forthcoming) also explores the implications of bargaining in the shadow of the law).

Townsend (1979) initiates the analysis of situations in which one economic agent can, at a cost, falsify or verify information while the court can enforce contracts given the endogenous information structure. Theoretical and empirical analyses examined the structures of optimal contracts in such situations of costly falsification and verification (e.g., Lacher & Weinberg, 1989; Williamson, 2002). Recognizing the contractual and efficiency implications of costly state falsification and verification opens the way for considering the implied motivation to alter the institutions that influence these costs. Gonzalez de Lara (2002) presented such an analysis, when she considered how late medieval Venice structured its legal system to reduce these costs, thereby enabling progressively more efficient contracts and more savings to be invested in profitable long-distance trade.

32.2.3 Designed, Private-Order Institutions

A private response to the needs for additional contract enforcement—beyond that provided by existing organic and public-order CEIs—manifests itself in designed, private-order CEIs (Henceforth, I will refer to these institutions as “designed CEIs” although public-order institutions are also designed). These are intentionally established by economic agents in response to profit opportunities entailed by improving contract enforcement beyond what is possible otherwise.

Designed CEIs, like organic ones, are based on the expected responses of economic agents. They are established either by the interacting parties (e.g., business associations) or by a third party (e.g., the stock exchange and credit rating companies). Institutionalists examined mainly designed CEIs that foster contract enforcement by the credible threat of imposing economic sanctions by the economic agents. Similar to public-order institutions, designed CEIs are intentionally planned and have explicit rules and intentionally created organizations. Because designed CEIs rely on sanctions by the economic agents, they are analytically similar to organic ones. But because designed CEIs are intentional, they often have features similar to those of public-order institutions, such as formal procedures to resolve disputes and impose fines.

The organizations and rules central to designed CEIs increase the disciplinary impact of economic sanctions by changing the information structure, providing coordination, and more generally by altering the strategic interaction among the economic agents. The organizations central to designed CEIs—or the people who control them—are often directly motivated by their economic interests and would not abuse their power-seeking profit.¹⁴ This mechanism, as elaborated below, often has to be supplemented by legal means.

Information intermediaries are pervasive in the modern economy, taking such diverse forms as auditing firms, credit reporting firms, the Better Business Bureau,

¹⁴For a formal analysis, see Milgrom et al., 1990; Dixit, 2003a; Greif, 2004.

credit rating firms, the Consumer Report, and business associations.¹⁵ They reduce the cost of acquiring information regarding an economic agent's past conduct, ability to perform, and professional credentials. They improve information quality, increase the speed of its circulation, and even certify one's identity in cyberspace. Information intermediaries improve monitoring, aggregate and track information regarding past conduct, and enable one to signal his reputation by paying fines, providing arbitration, and checking the quality of goods and services upon delivery or because of a complaint. Organizations, such as business associations, coordinate responses to contractual breaches, thereby fostering the certainty that a multilateral punishment will be imposed but reducing the risk of being improperly punished.

Other organizations foster contract enforcement by altering the structure of the interactions among the economic agents, particularly by replacing infrequent interactions among any two economic agents with frequent interaction with the organization. This changes the set of self-enforcing beliefs regarding conduct. A prominent example is a credit card company. When one pays with a credit card, the (possibly) infrequent transactions between the seller and the buyer are replaced by the frequent transactions between the buyer and the credit card company. Failing to pay a debt to a credit card company, unlike to a one-time seller, entails losing the gains from future purchases using the card.¹⁶ Exchanges, such as the London Exchange or the NYSE, operate on the same principle. The infrequent interaction between two particular sellers and buyers of securities is replaced by frequent ones between each seller and buyer and the exchange itself. Escrow companies, large retailers, and hotel chains foster contract enforcement in a similar fashion.

Hotels and other chains illustrate how organizations also foster contract enforcement by aggregating reputation. The organization is structured in a way that reputational consideration makes it credible that it will monitor and discipline its constituting members. A hotel chain is motivated to maintain the same level of service quality in its individual hotels, expecting that once disappointed, a customer will shun its other hotels (Ingram, 1996). A chain enables more commitment than is possible by individually owned hotels. Stock exchanges are similarly motivated by reputational considerations to discipline their member traders.

The credible threat by an organization to discipline its members, increases the value of membership, as each member can attract more trade. This higher value, in

¹⁵In 1985, for example, a credit reporting firm, currently called 'Seafax', began selling information regarding the past conduct of buyers in the fresh fish industry via the internet. It responded to the profit opportunity presented by the fact that sellers of perishable goods need to quickly market their product. They have little time to verify the creditworthiness of a new customer. Indeed, during the 1980s, the US fresh fish industry was characterized by repeated interactions (lock-in relationships) in which those selling fish were in a weak bargaining position. Seafax capitalized on this situation by becoming an information intermediary. It now provides information regarding companies in all segments of the perishable food industry in North America and its business news is updated every half hour. For analysis of such intermediaries, see Bernstein, 1996; Klein, 1997.

¹⁶A credit card company is also more likely to seek legal sanctions, given that its reputation is on the line and that credit card's debt aggregates the monetary value of many purchases.

turn, provides the organization with an effective disciplinary device in the form of exclusion. Indeed, reputation considerations motivate the NYSE to examine and certify the creditworthiness of its traders, judge disputes among members, suspend members who are at fault, and screen listed firms for their quality (Banner, 1998). Arguably, similar considerations motivated craft guilds and credit cooperatives (Guinnane, 1994) in the pre-modern economy and large producers, and business associations in modern times.

Because formal organizations are central to designed CEIs, these institutions' cost structure is similar to that of public-order CEIs. There is a high fixed cost in initially setting up an institution, as it requires the acquisition of organizational capacity (such as storing information), specifying and making common knowledge rules and processes (regarding membership, filing a complaint, and sanctions), generating awareness of the new organization, and creating the belief that it can support contract enforcement in a way that is beneficial to its customers. Once established, however, designed CEIs exhibit low marginal cost of expanding the number of individuals covered by the system or the number of transactions each is engaged in.

This assertion regarding the nature of the marginal cost, however, is too simplistic. It considers only the technological determinants of costs. But strategic considerations of these costs muddy the water. The organization (or the individuals who control it) can gain from abusing the information and power at their disposal. A credit rating company can gain by extorting money for good reports or charging customers for unjustifiably improving their rating. A business association can let its members go unpunished for selling defective goods and keeping the gain or sharing it among themselves. Ignoring possible legal sanctions, such behavior is likely to be punished eventually by customers who take their business elsewhere. The threat of losing future business to discipline the organization, however, means that its profits need to be sufficiently high. The costs of the institution also include the markup—beyond the technological marginal costs—required for providing the organization with the appropriate incentives. The cost of designed private and public CEI's can also be higher than technologically warranted because they are often a natural monopoly and hence, ignoring other factors, would not change marginal cost for their services.

The high fixed costs associated with designed institutions imply that organic ones would be more efficient when there is little to gain from expanding the number of exchanging individuals or when the loss due to bilateral repeated relationships is relatively small. There is rationale behind the observation that the largest (in terms of coverage) designed CEIs seem to be located in the consumer and retail sectors. The credit card companies and stock exchanges are but two examples. In any case, designed CEIs can substitute for organic ones and, due to their lower marginal costs and designed features, may be more efficient in providing contract enforcement in impersonal, complex exchange and in a dynamic environment.

We do not have a systematic body of knowledge regarding either the relative efficiency of public-order and designed, private-order CEIs or the factors influencing their relative efficiency. Public-order CEIs entail the potential benefits of impartial third party enforcers but also entail the various costs elaborated above. The organizations or individuals central to designed CEIs would enforce contracts only to the

extent that they can directly profit from doing so. Competition among designed CEIs would therefore not necessarily be beneficial as it is the expectation of future profits that motivates the CEIs not to abuse their enforcement ability. Yet, without competition, the CEIs would charge the price for its service that maximizes its profit rather than gains from exchange. Relative to public-order institutions, designed ones are likely to be better able to learn from the market's feedback, diversify their products, rely on tacit knowledge and statistical measures of performance, provide faster service and be more cost effective than a legal system.

Irrespective of their relative costs or other implications, designed CEIs can often substitute public-order institutions. The Bourse of Amsterdam was the most important and best organized in Europe during the seventeenth century. Yet, many of the financial instruments traded in it, such as short sales, forward contracts, options, and hypothecation of shares as collateral, were either in legal limbo or actually illegal. Reputation sustained trade until the time when these instruments became legal.¹⁷ The threat of economic sanctions can achieve the same deterrence as the equivalent legal sanctions.¹⁸ Even today, where legal systems are relatively ineffective, as for example, in contemporary Mexico, economic sanctions replace legal ones (e.g., Woodruff, 1998). Even where the legal system is well developed, as in the USA, similar sanctions are important in commercial transactions (e.g., Bernstein, 1992, 1996).

Arguably, designed CEIs substituted for the public-order CEIs in past economies to a larger extent than in modern ones as the latter were relatively undeveloped and the state had limited administrative capacity. (Although historically the distinction between private and public was less sharp than in modern time). Furthermore, there is evidence to support the claim that designed CEIs were central to the historical process through which various market economies grew and led to the conditions favorable to the establishment of public-order CEIs. Markets did not wait for public-order institutions provided by a centralized, territorial state. Rather, they developed based on designed CEIs.¹⁹

Consider the case of impersonal exchange characterized by separation between the *quid* and the *quo* over time and space. Despite the lack of impartial legal enforcement provided by the state, such exchange prevailed in late medieval Europe based on an institution that can be referred to as the Community Responsibility System (CRS). This was a designed system (although perhaps with organic roots dating back to the sixth century) with explicit rules and an organization that built on the fact that merchants were members of particular communes that had intra-commune contract-enforcement institutions. These intra-community institutions provided the foundation for an institution that provided contract enforcement in

¹⁷Stringham, 2003. Quinn, 1997 and Neal & Quinn, 2001 report similar findings regarding goldsmithbankers in London around the same time.

¹⁸But legal sanctions can be imposed when economic sanctions cannot due to, for example, budget constraints, outside economic opportunity, etc.

¹⁹See Greif, 2000 regarding Europe, and Schaede, 1989; Ramseyer, 1991; Ryser, 1997; Okazaki, 2002; and Kambayash, 2002 regarding Japan.

inter-community impersonal exchange despite the absence of a state with effective public-order institutions (Greif, 2002, 2004).

Under the CRS, communities established organizations in trading centers that enabled merchants to learn the communal and personal identities of their (otherwise unknown) partners in an exchange. If a member of community A, for example, had cheated a member of community B each and every member of community A was held responsible by community B for the damage. Hence, community A had the choice of either ceasing to trade with community B or compensating for the damage and seeking retribution from the individual who cheated. This joint-liability—which was neither contractual nor voluntary for an individual merchant—implied that each community was endogenously motivated to utilize its community enforcement institutions to discipline a merchant member who cheated in inter-community exchange. Anticipating compensation, merchants were motivated to learn the communal and personal identities of their partners to an exchange and could credibly commit to complain in a case of default despite the cost involved. Hence, communal courts provided impartial justice in inter-communal disputes although a community's courts were partial. Indeed, they provided impartial justice because they represented the interests of their merchants and cared about their reputations.

Institutions, such as the CRS, that fall in between the way we model private and public-order institutions, probably constituted an important step in the development of market economies. Indeed, if exchange historically began based on organic CEIs founded on personal relationships within relatively small groups, the emergence of the impartial legal system to facilitate impersonal exchange requires an explanation. Why was a high, fixed-cost, legal system for impersonal exchange established if the volume of impersonal exchange was low? We know that contracting efficiency alone does not lead to a transition from one system of contract enforcement to another (Greif, 1994a; Kranton, 1996b). Private-order institutions, such as the CRS, generated the initial volume of impersonal exchange required to justify the high sunk cost of the public-order institutions. Indeed, these institutions were established following a decline in the economic efficiency and political viability the CRS (Although, as discussed below, the success or failure of this transition depended on the existence of complementing institutions).

Designed CEIs seem to be equally prominent in modern market economies. They are such an integral part of these economies that it is easy to lose sight of their importance. They manifest themselves in organizations such as banks, credit cooperatives, credit card companies, consumer groups, escrow companies, trading companies, wholesalers, chain stores, hotel chains, banks, trade associations, unions, trading companies, trade and industry associations, stock exchanges, clearinghouses, credit rating agencies, credit bureaus, and better business bureaus. Their operation is reflected in brand names, copyrights, audits, guarantees, accreditations, etc.²⁰

²⁰Separating the identity of a business owner and that of the business itself fosters contract enforcement. The option of selling the business and its reputation increases the cost of a breach (Kreps, 1990 and Tadelis, 2002.)

Although the associated institutions have many other functions, such as reducing search costs, matching savers with investors, and smoothing consumption, contract enforcement seems to be an integral and important part of their operation.

Indeed, although we lack a systematic analysis of the relative importance of different types of institutions in various economies, designed CEIs seem to be the hallmark of advanced market economies. Fafchamps (2004) examined the institutional foundations of markets in contemporary Africa. He found that all have organic institutions, and some have effective public-order institutions. Where they uniformly fall behind is in their designed CEIs. Similarly, the engine of growth in modern economies has been the rise of their service and consumers' goods sectors. One can conjecture that this rise is due to their designed CEIs. Such CEIs have a relative advantage in these sectors due to the difficulties of measuring quality, the relatively small sums, and the many instances of infrequent interactions.

We similarly lack a systematic analysis of the relationships among and designed and public-order CEIs. Designed institutions are often established in response to their details and to circumvent the need and reduce the cost of using public-order institutions. Title insurance companies and escrow companies are examples. Public-order institutions are often required to prevent designed CEIs from using their economic power to curtail rather than expand exchange and efficiency. A drastic example is that of the Hanseatic League. This was an inter-city alliance that initially secured the property rights of German merchants abroad, and thereby promoted Baltic and Atlantic trade during the late medieval period. The same economic power that enabled the Hanseatic League to check the coercive power of rulers abroad enabled it also to acquire exclusive trading rights. It eventually turned into a welfare-reducing monopoly that restricted the entry of more efficient traders (Greif, 1992).

Although many designed institutions are completely private-order, others critically depend on complementary public-order institutions. Indeed, many designed institutions are what can be referred to as "quasi-private." Contract enforcement between the interacting economic agents in the main exchange relationships does not depend, or does so only marginally, on the law. The threat of economic sanctions provides the appropriate motivation. Yet, the law is critical to the operation of these institutions by creating the various conditions necessary for them to function and mitigating auxiliary contractual relationships.

The ability of credit rating agencies to secure property rights in exchange, for example, critically depends on public-order institutions that increase the cost of falsifying one's identity. Hotel chains, large producers, wholesalers, and banks commit by placing their reputations and hence future business as bonds. Yet, the value of this reputation critically depends on public-order institutions that protect their brand names. Accreditation and seals-of-approval provided by business associations or the Better Business Bureau motivate their holders to adhere to their contractual obligations. This motivation, however, benefits from the increase in the cost of forgery that public-order institutions create. Public-order contributes to maintain the value of organizations' reputations by increasing the cost for copy cats to enter, assume the organization's identity and cheat its customers.

Similarly, the designed CEIs in the modern economy imply separation between ownership and control. The corporate governance literature explores the contractual implication of the associated agency problems, concentrating, in particular, on inducing agents to exert the level of effort most beneficial to the firm in the presence of asymmetric information and contract incompleteness. Yet, an additional first-order problem is preventing these agents from directly taking actions that benefit themselves but harm the company. The corporate governance literature usually ignores such possibilities. Managers are assumed to be able to supply inefficient levels of effort but are unable to take the money and fly to Bermuda. More generally, those in control can, for example, transfer assets to their private accounts, get kickbacks from suppliers, provide their customers with defective products or false information, raise their salaries, and give themselves bonuses, and extort money using their ability to inflict sanctions. The firm or organization and its owners may care about the implied reputational or other losses, but those in control may not care about either their own reputations or the organization's (e.g., Johnson et al., 2000).

Very little systematic analytical and empirical attention has been devoted to the importance of quasi-private institutions. Even less attention has been devoted to hybrid institutions that combine elements of private and public-order and/or combine economic, social, and coercion sanctions. The CRS discussed above represents such a hybrid, as the concern of a community's reputation motivated it to use its coercive power against a community member who defaulted in inter-community relationships. Medieval guilds often motivated their members by economic, social, legal, and even religious sanctions (e.g., Richardson, 2002; Olds & Liu, 2000). The famous Grameen Bank lends to an individual through a formal contract but conditions its lending to others on repayment, thereby soliciting social pressure to motivate repayment (e.g., Ghatak & Guinnane, 1999). Several works examined the role of private coercion—organized crime—in providing contract enforcement (e.g., Gambetta, 1993; Greif & Kandel, 1995; Dixit, 2003a).

Many other aspects of designed CEIs are still waiting for a rigorous analysis. We do not have a theory for the internal organization of business associations and similarly designed CEIs, or the conditions under which reputation considerations prevent abuse of information and organizational power. The organizations central to many designed CEIs are often natural monopolies implying that despite their low (technical) marginal cost, their prices may be high, diminishing their usefulness. Even in this case, competition may be beneficial for reducing prices and ensuring quality of service. Yet too much competition can erode the reputational incentives that provide motivation to these organizations not to abuse their information and power. Conversely, firms strategically respond to the information regarding their reputations that is produced, for example, by credit rating companies.

The dynamic process of the emergence of designed CEIs or how exactly they acquire the information critical to their operation has not been rigorously studied either. Empirical studies (e.g., Hoffman et al., 2000) suggest the importance of initially having sufficiently valuable information which enables the organization to both gain from its distribution and use it to acquire additional information. The implications of increases in market scale (e.g., through globalization) on designed CEIs, has also not been studied. Conversely, the implications of international trade

on designed CEIs have barely been considered (Although see Greif, 1992; Yarbrough & Yarbrough, 2003).

Various contract-enforcement institutions—private and public, organic, and designed—have their distinct advantages and disadvantages and each complements and substitutes for the other. More reliance on organic CEIs, however, is optimal in relatively small and static economies and when there is less to gain from impersonal exchange. More reliance on public-order and designed CEIs is optimal in relatively large, dynamics economies with much to gain from impersonal exchange. Arguably, public-order institutions are more important in industrial economies in which complexity is relatively low and transactions relatively large. In this case, the court can relatively easily verify past actions, and the threat of a law suit is credible given the sums involved. Designed CEIs are relatively more important in complex, consumer- and service-oriented economies in which actions are difficult to verify *ex-post* and the sums are relatively small.

32.3 Coercion-Constraining Institutions

As noted by John Locke, among many others, the feasible extent of markets depends on protection from coercive power. For one thing, displaying one's goods in the market facilitate their confiscation by those with coercive power. In small social units, such as tribes, communities, and clans, the use of coercive power to capture another's property is likely to be effectively curtailed by the density of personal, social, and economic ties and the relatively even distribution of coercive power. In larger social units, this is not the case leading many, such as Hobbes, to argue that a state is required to protect rights and foster markets. This implies a dilemma: a state strong enough to protect rights is also strong enough to abuse them.

It has been suggested that this dilemma can be resolved by political institutions limiting rulers' power: limiting their prerogatives and placing political decision-making rights in the hands of asset holders (e.g., North & Weingast, 1989; Weingast, 1997). Yet, as was noted in the chapter by Mary M. Shirley, this answer is unsatisfactory. Market economies often prosper despite the lack of such political development. China had an extensive market economy during its Imperial past in the absence of such political rules. More generally, from 1950 to 1990, the rate of growth in national income of democracies and dictatorships was almost the same (Przeworski et al., 2000). Conversely, the political rules of modern developed economies were often adopted in developing economies without disciplining rulers or fostering markets. The political rules of the modern state are neither necessary nor sufficient for markets. Understanding the prevalence of markets and the co-evolution of markets and political rules requires examining deeper factors.

Accordingly, the following concentrates on coercion-constraining institutions (henceforth, CCIs). CCIs influence decisions at the social level regarding the acquisition and use of coercive power. Effective CCIs make violence economically productive as it is used to protect property rights from abuses, such as expropriation by the state, the ravages of a civil war, and large-scale military raids. They secure property rights by discouraging those who can acquire coercive power to abuse

rights from doing so, and by motivating those who have coercive power—rulers, the elite, states (and I henceforth use these terms interchangeably)—to protect rights. These CCIs rely on balancing one's coercive power with either the coercive power of others or their ability to inflict economic sanctions on one who abuses rights. CCIs deter abuse of rights by creating the shared beliefs that attempting to do so will lead to a costly retaliation.

There is no one-to-one correspondence between states and CCIs. There can be effective CCIs in the absence of a state while in predatory states there are no effective CCIs. Yet, the common denominator of CCIs can be illustrated by considering the argument that a state strong enough to protect property rights is also strong enough to abuse them. Two assumptions are implicit in this statement. The first is that only the state has coercive power. But the existence of a state does not preclude social units within it—including those composed of economic agents—from having actual or potential coercive power. Countervailing coercive power can constrain the state. Indeed, states' coercive power has historically often been no more than the aggregation of the coercive power of its composing social units. A balance of coercive powers within a polity—among such social units as towns, clans, tribes, and classes—can limit the abuse of rights.

The second assumption implicit in the above statement is that having coercive power implies an unconstrained ability to gain from abusing rights. This, however, is not the case. The costs and benefits to a ruler from abusing rights depend on the state's administration's capacity to gain information regarding assets that can be captured, capturing them, and transforming the proceedings into goods and services beneficial to the ruler. A ruler's costs and benefits from abusing rights depends on administrative capacity and who controls the administration. Furthermore, administration is required for effective ruling. The court and the army need to be provided for, and resources have to be mobilized to advance the ruler's policies. Those controlling the administration are therefore in a position to take actions that are costly to the ruler. In particular, if the state's administration is controlled by the asset holders, abusing their rights can undermine, rather than foster a ruler's welfare. The capacity of, and control over the state's administration can be structured in a way that the ruler can either credibly commit not to abuse rights or is deterred from doing so.

Furthermore, various technological and institutional factors limit a ruler's grabbing hand. These factors influence a ruler's benefits from abusing a particular asset and the extent to which economic agents can respond to expropriation by economically sanctioning the ruler. For example, a ruler gains little from capturing an asset whose value is lost without the original owner's complementary human capital, and abusing alien merchants' assets can drive the merchants away, depriving a ruler of trade benefits. Hence, even a ruler with a monopoly over coercive power can be deterred from abusing rights by countervailing economic powers reflecting either administrative structure or institutions taking advantage of limits to a ruler's grabbing hand.

CCIs that are based on either countervailing coercive or economic powers reflect the same principle. The expected responses of those with coercive and economic powers influence one's decision regarding abusing their assets. Protection is afforded, however, only to those who can retaliate. This section presents the principles on which various CCIs rest (although a CCI often rests on several principles), their origin, dynamics, and relationships with political development.

32.3.1 Balancing Coercive Powers

In CCIs based on balancing coercive powers, the expected violent response of the asset holders deters one from abusing these assets. Bates et al. (2002) provided a general framework to study the nature and costs of various such CCIs in polities with and without a ruler.²¹ The analysis considers the strategic decisions regarding acquiring and using coercive powers focusing on equilibria in which no abuse occurs. By imbedding this problem in a resources allocation problem, it is possible to measure the cost of securing rights under various CCIs. It is the difference between the equilibrium and the first-best allocation of resources.

Consider an infinitely repeated interaction among agents (e.g., clans, tribes, and towns). In each period, each agent simultaneously chooses how to allocate a finite amount of effort among production, investment in coercive power, or leisure (which cannot be expropriated). After observing the choice of the other, each can sequentially decide whether to raid the other—at some cost—or not. The amount gained from raiding increases in one's relative coercive strength, the superiority of defense over offense, the amount produced by the other, and the share of the products lost due to a raid.

No Ruler: Mutual Deterrence Among Social Units When there is no ruler, there are equilibria in which each agent sufficiently invests in coercive power to deter the other from raiding him. Theoretically, such a “mutual deterrence” equilibrium is more likely to exist the more the agents value the future; the higher the cost of raiding; the more the military technology favors defense; the higher is the share of the products lost due to a raid; and when the agents also gain from economic cooperation (e.g., trading or joint production that they can lose following a raid). Furthermore, the higher the value of these parameters, the lower is the cost of securing property rights as the equilibrium allocation of effort is closer to first best. Yet, such a mutual deterrence equilibrium generically entails a socially wasteful allocation of effort to acquire coercive power. Furthermore, it reduces the incentives to make productive investments because wealth requires resources to protect it. Finally, mutual deterrence equilibria are unstable in the sense that even transitory changes in relative might or wealth lead to military conflicts.

Tribal societies provide the classic example of polities without a ruler, but such polities even exist in urban settings. The commune of Genoa was a polity in which security of property rights was based on mutual deterrence. By the end of the eleventh century, the decline in the Muslim and Byzantine naval powers provided maritime cities, such as Genoa, with an opportunity to gain from expanding its overseas trade.²² The Genoese responded by establishing a commune in which

²¹ See also Greif, 1994b, 1998; Muthoo, 2000; and Bates, 2001. For works considering the relationships between military technology and security of rights, see, e.g., Skaperdas, 1992; Skaperdas & Syropoulos, 1996; Konrad & Skaperdas, 1996; Grossman, 1997; Grossman & Kim, 1995; Moselle & Polak, 2001.

²² See Epstein, 1996 for Genoa's history. For analysis, see Greif, 1994b, 1996, 1998, Forthcoming.

mutual deterrence among Genoa's heavily armed clans secured property rights. Clans jointly mobilized their resources to equip the navy, and the military force required to expand Genoa's commerce. The extent of inter-clan cooperation was limited, however. Consistent with the claim regarding the disincentive for wealth expansion under mutual deterrence, the clan who gained most from the previous expansion ceased cooperating in further expansion. A richer Genoa would have required increasing its military investment to deter others from attacking it.

Wide-scale commercial expansion did not occur until 1155 when the constraint implied by mutual deterrence was relaxed by an external military threat from the German Emperor. Inter-clan military conflict would have made the city more vulnerable to attacks by the Emperor. As each clan had to devote fewer resources to deter the other, each was willing to cooperate by advancing commerce. Yet, as Genoa grew in wealth and the external threat unexpectedly subsided (due to a civil war in Germany) in 1164, a prolonged period of civil war ensued.

Make me a King Extending the model to include a ruler with coercive power enables examining the conditions under which he would better secure rights. Assume that a ruler can use his coercive ability, after paying a fixed cost, to capture assets. He will capture the assets with some positive probability (which can depend on the agent's coercive power). Apart from the actions discussed above, each agent can now also either pay a tax or not.

Consider the equilibria in which property rights are secured. Each economic agent is deterred from raiding by the military ability of the others, but in addition, he is also deterred by the threat of the ruler retaliating against one who raided. The ruler is motivated to retaliate by the expectation that as long as he does so and refrains from abusing rights, the agents will pay him taxes. If he fails to retaliate or abuse rights, the agents will revert to playing another equilibrium strategy in which they invest more in military ability and consume more leisure. In this case, the ruler loses tax revenues. In such equilibria, the coercive power of each player—the ruler and the economic agents—is constrained by the coercive power of the two other players.

Distortions are caused by two factors. If the relative coercive power of the ruler is not sufficiently high, in and of itself, to deter the economic agents from raiding each other, each of them will invest in acquiring coercive ability. Effort allocation is therefore distorted. As the ruler's coercive power increases, however, and his expected retaliation is sufficient to deter raiding, this distortion approaches zero. A second distorting factor is taxation. Although taxation is socially beneficial as it motivates the ruler not to abuse rights, it causes the economic agents to consume more leisure than in the first-best allocation. The per-period equilibrium taxation (on-the-equilibrium-path), and hence distortion, is lower the lower a ruler's per-period payoff after he abused rights (that is, off the-equilibrium path). Intuitively, if his continuation value is lower, a lower per-period tax (on-the-equilibrium-path) is sufficient to deter the ruler from abusing rights (The distortion declines in the ruler's discount factor. Olson (1993) presents a non-strategic analysis of this important consideration).

The extent to which an equilibrium with a ruler can approximate a world with perfectly secured property rights and a first-best allocation of effort thus depends on

two factors. The higher a ruler's coercive power, the less effort each agent has to allocate to deter raiding; the lower the ruler's continuation value after he abuses rights, the lower the per-period tax required to deter abuse, and hence the lower is the access consumption of leisure due to taxation. At the limit, the allocation of resources approaches the first-best as the ruler's coercive power increases but his continuation value decreases. The ruler's coercive power constrains the economic agents' coercive power while their ability to acquire coercive power and consume leisure constrains the ruler's coercive power.

Central to the economic efficiency of the CCIs with a ruler is his coercive power on- and off-the-equilibrium path. The CCIs are more efficient the higher is the ruler's coercive power on-the-equilibrium-path (that is, as long as he did not abuse rights or renege on punishing those who did) and the lower is his coercive power off-the-equilibrium-path (that is, after he abused rights or failed to punish those who did). The following discussion presents historical CCIs in which this condition has been satisfied to various degrees. It highlights the efficiency of CCIs in which there is no ruler with an independent military power.

32.3.2 A Ruler with an Independent Coercive Power

Motivated by a severe external military threat, the Genoese altered their political system in 1194 by introducing a *podestà* (literally, a "power") to create a balance of coercive power among their clans. The *podestà* was a specialist in violence, a non-Genoese hired by the city for a year to be its military leader, judge, and administrator. He was supported by the soldiers and judges he brought with him. The *podestà's* military strength was such that the threat of him possibly joining forces with a clan that was attacked deterred each of Genoa's main clans from instigating an attack. Yet, the *podestà* was too weak to become a dictator and abuse rights given the strength of Genoa's clans and its population. A set of pre-specified rules defined which actions by the *podestà* constituted an abuse.

Central to the *podesteria* system, however, was the way that the "threat" of the *podestà* retaliating against a transgressor by joining forces with another clan was made credible. The *podestà* was promised a high wage at the end of his term, but because he was weaker than each clan, if one clan took control of the city, there would be no reason to reward the *podestà*. Hence, he was motivated to act against a transgressor because otherwise his payoff would be lower. Furthermore, this reward scheme made it in the *podestà's* interest to not fundamentally alter the balance of power between the clans. Hence, he could credibly be impartial and retaliate against those who broke the law rather than turn against an entire clan.

For this incentive scheme to be effective, however, it was imperative to ensure that no clan would be able to credibly commit—using such means as marriage—to rewarding the *podestà* if he assisted that clan against another. More generally, it was imperative that the *podestà* could not acquire a power base within Genoa. This was accomplished through a series of regulations, supervised by a committee, which restricted his actions. E.g., the incoming *podestà* was selected by a council, whose

members were chosen to prevent control by any specific clan, and the selection process for the new *podestà* was governed by the outgoing *podestà*. The *podestà*—as well as his relatives to the third degree—was restricted from socializing with Genoese, buying property, marrying a local woman, or managing any commercial transactions for himself or others. Furthermore, the *podestà*, as well as the soldiers who came with him, had to leave the city at the end of his term and not return for several years. A *podestà*'s son could not replace him in office.

The Genoese promise to reward a *podestà* at the end of his term if no clan had initiated and won an inter-clan war was credible because renegeing would have impeded Genoa's ability to hire a high quality *podestà* in the future. Indeed, Genoese *podestàs* were recruited from a handful of Italian cities, in particular from Milan, and the contract between Genoa and its *podestà* was read in front of the "parliament" of the city from which the *podestà* was recruited. This does not imply that a *podestà* was given a free hand to mismanage the city's affairs. After the end of his term, he had to remain in the city for fifteen days while his conduct was assessed by auditors. Deviations from pre-specified rules were punished by fines that were subtracted from his payment.

Historically, other polities, similar to Genoa, contracted with an external ruler to govern them to mitigate internal conflicts over leadership. Indeed, the prevalence of this practice into the modern period seems to have facilitated colonization (Henley, 2004). As colonialism illustrates, rulers with independent coercive power were sometimes able to capture the polity. But even in this case, a balance of coercive powers often secured rights. Consider the Manchurian Qing Dynasty (1644–1911) which gained control over China after the Manchurians were invited to intervene in a civil war in China. The relative size of China and that period's military technology implied a high cost of subduing a revolt by the masses. The Emperors recognized "the people as a persistent potential threat to the Chinese state" (Wong, 1997, p. 93) while a Chinese tradition conferred legitimacy on a dynasty if it was able to foster economic prosperity. Emperors were therefore motivated to commit to protect property rights and foster prosperity.

An elaborate system of rules, organizations, and precedents was used to generate the correct beliefs that rights would not be abused by, ironically, making the threat of revolt credible. The Qing imposed a relatively low tax, coordinated expectations regarding taxation by fixing its amount and making it common knowledge at the village level. Additional customary taxes were collected by the bureaucrats to finance local public goods and to remunerate them. Yet, bureaucrats were disciplined not to tax the peasants beyond the official and customary level. A bureaucrat who increased taxation faced riots by the peasants, which subsequently triggered an investigation and possibly punishment by the central authority (Yang Li, 2002).

32.3.3 A Ruler with a Conditional Coercive Power

In the competition among the Italian maritime city states, it was Venice which overtook Genoa. Interestingly, it had CCIs with a ruler without an independent coercive power but with a conditional power: he was strong on-the-equilibrium

path (that is, as long as he did not abuse rights or failed to punish those who did) but was weak off-the-equilibrium-path (that is, after he abused rights or reneged on punishing those who did). Hence, Venice's CCIs entailed a low-cost protection of property rights.

Initially, however, Venice's polity was characterized by the absence of an effective rule. The Venetians were left to govern themselves from the eighth century due to the decline of the Byzantine Empire.²³ The city's main families and clans became engaged in lengthy and bloody competition over the position of the Doge (who previously was a Byzantine governor sent from Constantinople). Of the twenty-nine Dogi that governed between 742 and 1032, about three-quarters were either assassinated, blinded, resigned, or expelled due to internal conflict.

A stable polity was established, however, during the eleventh century, probably in response to the increasing gains from jointly mobilizing resources to promote long-distance trade. The Doge was made a magistrate, elected for life and responsible for establishing social order and providing public goods. Coercive power remained highly diffused among the prominent clans and families of Venice which were represented in a general council. The Doge himself was supervised by an elected council, was not allowed to go against its advice, and was subject to the law. Without a standing army, a Doge could be punished for breaking the law and abusing rights. Yet, the Doge was strong on-the-equilibrium-path as each clan was motivated to support the Doge if any other clan attempted to capture the position of the Dogeship or abuse others' rights.

Motivation for the clans to lend the Doge their coercive power was based on the way that Venice's institutions distributed gains from the city's wealth. Lucrative economic and political posts were distributed independently of a clan's relative coercive power. Posts, including that of the Doge, were allocated to members of the political elite through a mixture of deliberation and random selection. The random component implied that one's clan relative military power and patronage system had little influence on the outcome. Explicit rules and historical experience coordinated the clans' beliefs on the above behavior.

32.3.4 Balancing Coercive Power with Economic Power

The prospect of losing future economic gains following an abuse can constrain coercive power even in the absence of countervailing coercive power. In considering CCIs based on balancing coercive power with economic power it is useful to differentiate between two groups of institutions. First, institutions based on an inherent limit in the ability to gain economically from one's coercive power; second, institutions based on the need for administrative capacity to gain from abusing rights and effectively rule. The following discussion reflects that these institutions have not been extensively studied.

²³ See Lane, 1973; Norwich, 1977. The analysis here builds on Greif, 1995, Forthcoming.

CCIs Based on the Limited Reach of the Ruler's Grabbing Hand Confiscation of an asset can reduce its value due to complementarities between it and other assets, human capital, and expertise. If this reduction is high enough, providing a stream of rent to those with coercive power would be more profitable to them than capturing the asset. Rent can be provided, for example, by paying taxes or providing economically lucrative jobs. The cost of security is the distortions implied by mixing politics and economics in the absence of the rule of law. These costs will decrease the more value is lost from confiscation, which is the case when production requires high human capital, inventiveness, and complementary assets that are difficult to expropriate. Indeed, Mexico has failed for a long time to de facto nationalize its oil industry because the foreign oil companies had the expertise, organizational capacity, and the complementary assets required to render oil production profitable (Haber et al., 2003).

Abuses can also be deterred by the expectation that the economic agents will shift their activities following an abuse in a manner that would be costly to the abuser. Agents can shift their activities abroad or turn to produce or consume goods, such as leisure, which are more difficult to expropriate. Deterrence is enhanced by the mobility of assets, complementarities with other assets that cannot be captured, and the ability of the economic agents to overcome the collective action and free-rider problems associated with collectively responding to abuses. The following example highlights the general principles underpinning such institutions.

Specifically, consider the case of the medieval merchant guild (Greif et al., 1994). Long-distance trade in late medieval Europe was based upon exchanging goods in geographically favorable places. Medieval rulers who controlled these areas faced the temptation to abuse merchants' property rights using their coercive power. Furthermore, a ruler could abuse the rights of some merchants but not others implying that, when there are many merchants, the threat of one whose rights were abused to never trade again following an abuse was insufficient to enable a ruler to credibly commit to secure rights. Hence, without an institution making the ruler's pledge to provide protection credible, alien merchants were unlikely to frequent that trading center, a costly outcome for both ruler and merchants.

Theoretically, to surmount the ruler's commitment problem at the efficient level of trade, an organization with two abilities was needed: first, the ability to coordinate the responses of all (or enough) merchants if the rights of any merchant were abused; second, it had to have the ability to enforce its embargo decision on the merchants despite the fact that if an embargo was in force, an individual merchant had a lot to gain from trading. Indeed, the value of his trade to the ruler would be so high that the ruler could credibly commit to respect that merchant's rights. The late medieval period witnessed the rise of merchants' territorial associations with coordination and internal enforcement capacities that governed the relations between their members and rulers of other territorial areas. These organizations took many forms, such as the German Hansa, merchant guilds, and autonomous cities. Yet, all of them had the capacity to coordinate action and discipline their members. Together with the shared beliefs associated with the above multilateral reputation mechanism that they made possible; these organizations constituted the merchant guild institution.

CCIs Based on the State's Administrative Structure Sending soldiers on a rampage is a costly way for a ruler to materially benefit from his coercive power. In

contrast, state-controlled administration with the appropriate information, organizational capacity, and incentive reduces the cost and increases the benefits from confiscating wealth. It has information regarding who has wealth, where it is located, the capacity to take control of it, and the ability to dispose of or employ it in a way that benefits the ruler. Furthermore, administration is required for the state to function: the ruler and the court must be supplied, soldiers have to be recruited, trained, equipped, and paid, and the public-order benefitting the ruler has to be maintained. The size, capacity, and control over the administration thus influence the costs and benefits of abusing rights to a ruler.

The (limitedly) absent state: By not creating an effective administration to govern a particular economic sphere, a ruler can commit not to abuse rights in that sphere because limited administration increases his cost of confiscation. When a ruler stands to gain less from abuse, property is more secure. Security increases with the cost and time it takes to establish an administration with the capacity to abuse rights at low cost. Similarly, security increases the more the agents are able to consume, transfer, or hide their assets after observing the initial stages of establishing a more effective administration. The initial absence of an effective administration, in turn, fosters their ability to do this. By being absent from a particular economic sphere—in the sense of not having an effective administration—a state can better commit to respect rights. Note that a state committing to rights in this way can be and usually will be absent only to a limited extent. To survive it has to be able to raise revenues, have an effective military force, and be able to provide public goods. These must be supported by an administration confined to these tasks.

Because assets' mobility increases security in the presence of an absent state, CCIs based on such absence were particularly important historically in securing traders' rights. Indeed, a common feature of many past market economies was the small extent to which movable assets were taxed apart from customs and payment for services within the market area itself. This was the case in the medieval Muslim world whose market economy was perhaps second-to-none. Traders were not subject to wealth or income tax and were taxed only when they voluntarily brought their goods to a particular locality such as market places and ports (Goitein, 1971; Udovitch, 1988). The same situation prevailed in the prosperous market economy of Imperial China (Wong, 1997; Pomeranz, 2000; Yang Li, 2002). In both cases, however, the state was not absent from other spheres of economic activity. It provided public goods, including social order, personal security, and agricultural and commercial infrastructures.

Delegation of state administration to asset holders: Commitment to not abuse the rights of asset holders can also be achieved by delegating state administration to these asset holders. Instead of having an administration controlled by the state, public goods and services to the state are provided directly by the asset holders. Similar to an absent state, commitment to property rights is achieved by depriving the ruler of the information and organizational capacity required for low-cost abuse of rights. More commitment is possible, however, because the asset holders can respond to abuse of their rights by cutting off the flow of services to the state. When the state depends on financing or tax collection provided by the asset holders to sustain its courts or maintain its army, the cost of withdrawing services can be high.

The factors determining the extent of security through delegation are similar to those in the case of the absent state.²⁴

The large extent to which delegation can secure rights is reflected in its ability to constrain the most powerful rulers of sixteenth century Europe: the Hapsburg Kings of Spain. The kings borrowed heavily from Genoese financiers who had an international monopoly over paying the royal army outside Spain. The Genoese could therefore respond by withdrawing these services if the king refused to pay his debt. The famous bankruptcies of the Spanish kings were indeed periods of debt reorganization rather than abuse of rights per-se (Conklin, 1998).

The experience of the Genoese in Spain also illustrates how delegation fosters security by influencing the information structure and thus the cost of various actions. The Genoese collected the taxes used to pay Spaniards who held the Crown's domestic loans, each of which was linked to a particular tax revenue source. Because the king did not have information about who held the various loans, he could not repudiate them without possibly hurting those—such as his military elite—who could retaliate against him. Abuse risked hurting those upon whom the ruler depended to maintain his control.

Self-governance: Property rights are even more secure when delegation is done in the context of giving the asset holders an autonomy. As before, delegating the administration of the state renders it vulnerable to economic sanctions. But self-governance further fosters the asset holders' ability to commit to retaliating following an abuse of their rights. Self-governance entails having bodies of collective decision-making, mechanisms, such as judicial processes and police forces, to overcome the free-rider problem and motivate and induce members to participate in sanctions (The above discussion of the merchant guild is relevant here).

Hapsburg Spain illustrates the effectiveness of self-governance. During the sixteenth and seventeenth centuries, towns had administrative autonomy (Nader, 1990) that balanced the Crown's coercive power with their economic power. In 1571 the Crown decided to substantially increase the sales taxes collected in the towns. The administration of sales tax collection, however, was in the hands of the towns themselves. They informed the Crown that they would not farm out this higher tax, leaving the king with little option but to look for additional revenue elsewhere.

32.3.5 The Origin and Dynamics of CCIs and Political Institutions

Many factors that are exogenous to CCIs influence whether a particular CCI can be an equilibrium outcome in a given situation. Furthermore, generically, many CCIs can be equilibrium outcomes in a given situation. So, while exogenous factors

²⁴Delegation is different from farming out of state administration under which agents (who are not necessarily the asset holders) compete for the right to e.g., collect taxes. Property rights security thus depends on the state's interest and ability to prevent over-taxation. Ottoman history illustrates the fragility of such systems.

determine the set of feasible institutions, initial, possibly even transitory, historical factors influence which CCI will prevail.

CCIs based on either balance of coercive or economic powers are more likely to be an equilibrium the more those with coercive power value the future. CCIs based on a balance of coercive powers are also more likely to be an equilibrium as the production and military technologies reduce the per-period gain from abuse and increase the asset holders' potential military strength. E.g., the lower is the portion of the product that can be expropriated, the less mobile and more perishable is the product, the more defense is superior to offense, the easier it is to convert civilian production controlled by the asset holders to military production. CCIs based on balancing coercive power with economic power are more likely to be an equilibrium as the geography, production, monitoring, and information technologies increase the costs and reduce the benefits of expropriation. E.g., those with coercive power have limited independent economic resources, it is less costly to move assets away from their reach, and establishing a new administration is time consuming and costly. The distortions and hence costs entailed by the need to secure rights declines in the above parameters.

For a particular CCI to prevail the relevant actors should share the beliefs that the related strategies will be followed. They should share beliefs regarding what action constitutes an abuse and what the consequences of doing so will be. Which set of shared beliefs, out of the many that are generically possible as an equilibrium outcome, will prevail reflects initial conditions, such as the initial distribution of wealth, military might, and coordinating mechanisms. Non-economic and potentially temporary factors, such as cultural heritage, the legitimacy and interest of coordinating organizations, and charismatic leaders, therefore, play a role in institutional selection. Initial conditions, spontaneous evolution, learning, and intentional design influence institutional selection.

The CCIs of Genoa and Venice illustrate this argument. The Genoese and Venetians had similar military technology and similar initial endowments in the form of little arable land and locations favorable to trade. But they differed in their initial social structures, distributions of wealth and military ability, and tradition of centralized endogenous rule. The less concentrated social structure and more even and wide spread distribution of wealth and military might in Venice and the coordinating effect of the tradition of a Doge led to different institutional development.

Once a particular CCI establishes itself as an equilibrium outcome each actor's unilateral ability to change it is limited. An attempt to unilaterally change the CCI by taking such actions as raising a standing army or expanding the administration, entails military or economic retaliation. CCIs imply deterrence; one is prevented from taking an action that he would find beneficial in the absence of the institution. Each actor acts optimally given the constraints on his behavior implied by the institution, but he would have preferred, *ex-ante* or *ex-post*, not to be constrained. This is so even in the case when the CCI is Pareto-improving, enabling one to credibly commit not to abuse rights. E.g., a CCI can enable a ruler to *ex-ante* commit not to confiscate wealth *ex-post*, thereby promoting growth. *Ex-post*, however, the ruler would prefer to be able to confiscate it.

But CCIs do not last forever. They cease being effective, are changed unilaterally or multilaterally, intentionally, or spontaneously, in response to exogenous changes in factors rendering them equilibria, unexpected consequences, strategic innovations, external threats, and mutual gains. Unilateral institutional changes occurred in Europe, for example, following the fifteenth century Military Revolution which tilted the balance of coercive power in favor of the Crown and landed nobility and against peasants. Less able to constrain abuse, peasants throughout Europe were subject to increasing serfdom and taxation. This increase, however, was gradual, reflecting a learning process of the implied new balance of power (Pettengill, 1970).

Gain from cooperation and external threats are a main source of mutually agreed upon institutional change as the histories of Genoa and Venice illustrate. The cities' histories also illustrate that the heritage of previous CCIs influence the set of CCIs that can be established as a new equilibrium outcome. The distribution of military and economic resources, shared beliefs, social structures, cognitive structures, and organizations that were part of the previous CCIs constitute part of the initial conditions in processes of institutional change. Hence new CCIs include institutional elements inherited from previous ones (Greif, [Forthcoming](#)). Before and after the Genoese *podesteria*, a high concentration of military power and wealth and beliefs regarding the objectives and behavior of clans implied CCIs based on inter-clan mutual deterrence.

The set of feasible, consensual, institutional changes is also limited by the need to insure that each actor who can *ex-ante* block the change will not do so. This requires that those who will, *ex-post*, have more coercive or economic power and commit not to use this power to make others worse off. Changes that otherwise would be Pareto-improving would not be feasible without such commitment. For these changes to be agreed upon, those who will have less power should nevertheless expect to be better off by having a smaller share in a larger pie or by *ex-ante* devising *ex-post* safeguards, which take into account the *ex-post* new distribution of powers.²⁵ Strategic innovations motivated by the constraints implied by existing CCIs facilitated establishing new CCIs. The introduction of the *podestà* in Genoa represents an innovation leading to a Pareto-improving CCI. The innovation constitutes a response to the needs of the Genoese, but it reflects the introduction of a *podestà* in Italian cities that were controlled by the Emperor decades earlier and a learning process through which Italian cities experimented in various CCIs and administrative structures. Strategic innovations aimed at advancing one's relative power often take advantage of the ambiguity over which actions constitute a deviation. E.g., the rights of monopolies were ill defined under feudalism and European kings later attempted to take advantage of this by creating monopolies to gain resources and might. Gradual and unanticipated consequences implied by existing institutions lead to institutional change. They alter such factors as wealth, military power, and

²⁵ E.g., Greif, 1998, [Forthcoming](#) (Genoa), 2001 (contemporary Middle Eastern dictators fearing that fostering development will create a countervailing economic power). Acemoglu & Robinson, 2000 and Acemoglu, 2003 (the limits on political transitions due to the need to commit to ex-post compensate a ruler). Fearon, 1997 (bargaining over distribution when it determines ex-post coercive power and hence further bargaining power). Galor & Moav, 2003 (inter-state conflicts motivate capitalists to increase workers' human capital although this leads to their demise).

information that constitute the necessary conditions for the existing CCI to remain an equilibrium. The absolutist French kings of the late sixteenth century who were above the law had to pay a high interest rate for their loans. Aspiring to be able to better commit to pay their debts, they allowed the financiers to establish corporations. These were better able to balance the Crown's coercive power with an economic power. The long-run unexpected consequence was that these financiers became so economically powerful that they blocked institutional reforms that would have hurt them financially but might have saved the monarchy (Root, 1989).

Such unanticipated consequences imply that those who benefit from the existing CCIs would be wary of organizational, economic, and political changes that can undermine these CCIs. Rulers, in particular, would be wary of delegation, self-governance, and independent military organizations. The threat to rulers from allowing self-governance is well reflected in the history of Genoa where a large scale, designed private-order institution altered CCIs by aligning the incentives of many economic agents and increasing their organizational capacity. The Genoese Bank of San Giorgio was established, according to Niccolo Machiavelli (1532), when the republic conceded control over various revenue sources to its creditors after a military defeat. These creditors organized themselves as a self-governed entity, the Bank of San Giorgio. As Genoa's debts continued to accumulate, the Bank gained the administration of most of the towns and cities in the Genoese dominion. It became so powerful, according to Machiavelli, that whoever gained political control over Genoa, had to respect the rights of the Bank "as it possesses arms, money, and influence," and abusing its rights entailed "the certainty of a dangerous rebellion" (p. 352).

Although delegation, self-governance, independent military organizations, and large-scale, private-order institutions endanger rulers, they nevertheless will establish or tolerate them when their benefits are high relative to the risk they imply. This is more likely to be the case when the ruler's budget constraint is binding; when the ruler's legitimacy is high and hence the risk of revolt is low; when the organization can be abolished at will (as was true regarding the Merchant Guild); and when the institution provides services that are important for the ruler's control, such as feeding the capital city. Greif (2000), for example, argued that because the legitimacy of European rulers was high relative to that of Muslim rulers in the late medieval period, the former allowed more designed CEIs than the latter. Okazaki (2002) noted that despite initial resistance, the Tokugawa Shogunate (1603–1868) allowed food merchants to organize themselves, fearing that otherwise volatility in food prices would lead to riots that would undermine their control.

CCIs influence political development, particularly whether representative bodies would emerge or be established by rulers as means for collective, political, decision-making. Such bodies have been common in a variety of political systems, ranging from monarchies (e.g., the English Great Council), constitutional monarchies (e.g., the English parliament), and tribal societies (e.g., the Afghani Loya Jirga). Seemingly diverse, these representative bodies reflect the constraint implied by CCIs. CCIs imply that, fearing retaliation, each actor will be deterred from taking unilateral actions that can be considered abusing rights or attempting to undermine the existing CCIs in a manner that would leave them worse off. Hence, an actor wishing to take such an action without invoking retaliation, would seek the consent of those who can retaliate.

This account provides a rationale for a puzzling observation: why do rulers, as was common in pre-modern Europe, allow for representative organizations whose members are not hand-picked by them although such bodies provide an arena for revealing and coordinating opposition to the ruler? The French revolution transpired after the Crown summoned the Estates-General for the first time since 1614. But when a ruler faces effective CCIs, he also stands to gain from representative bodies. They enable him to take actions with lower risk of costly retaliation. Representative bodies are a means to design explicit rules increasing—temporarily or permanently—the range of acceptable actions. Representative bodies are therefore more likely to be established when the ruler's budget constraint is binding, and he is unable to provide the public goods he and his subjects desire.

We have no models of bargaining in the context of various CCIs. It is intuitive, however, that those who provide a ruler with their consent, will demand and receive concessions for doing so. As noted above, these concessions are likely to take the form of safeguards against future abuses. Providing the ruler with additional resources risks shifting the balance of power in his favor, implying that those who are in a position to authorize this transfer, given the existing CCIs, will seek to safeguard their position. They will demand various concessions—such as freedom from taxes, administrative control, military resources, legal rights, the right to supervise or authorize various actions—to increase their ability to constrain the ruler *ex-post*.

Ironically, representative bodies also facilitate the abuse of property rights. Representative bodies are populated by those who balance each other's coercive and economic powers, providing them with a means to coordinate the abuse of rights for those who are not represented. In Poland, after the European Military Revolution that began in the fifteenth century, the balance of military power shifted in favor of the landed, lower nobility. They gained dominance in the kingdom's national assembly and used expropriated rights. The Polish serfs lost many rights during that time. Their tax obligations, for example, increased several times.

Conversely, if the economic and coercive power of groups represented in the representative body no longer constrains the ruler's actions, these groups will become, at most, a rubber stamp, a mechanism for a ruler to coordinate collective actions he finds useful. Alternatively, when the risk to the ruler of coordinating opposition through these bodies outweighs their benefits, he will not summon them. When absolutism reached its peak in France during the seventeenth century, the Estates-General was not summoned.

32.4 The Dynamics of Market-Supporting Institution and the Implied Dynamics of Markets and Political Institutions

The dynamics of CEIs reflect preceding initial institutions. Initial organic, private-order CEIs influence the extent to which economic agents will find it profitable to use public-order and designed CEIs if available and establish them if possible. Initial

CCIs influence the extent to which these CEIs can be effectively provided or established. Public-order and designed CEIs that reveal wealth to those with coercive power will only be utilized if the CCIs are such that revelation of wealth does not undermine property rights security. Conversely, rulers will permit designed, large-scale CEIs only if the CCIs are such that they either cannot prevent their establishment, or these institutions do not undermine their control.

Initial institutions can lead market economies along distinct institutional trajectories. Furthermore, the same CCIs that are conducive to the emergence of wealth-revealing public and designed CEIs, and hence to market expansion, are also conducive to the emergence of political institutions in which the commercial sector will be represented and have influence. Members of the commercial sector need not fear revealing their wealth if the CCIs imply that a costly retaliation can be imposed if their rights are abused. These same CCIs imply that a ruler will find it beneficial to get permission from the commercial elite prior to taking action, particularly any action related to their property (e.g., taxation), to reduce the likelihood of costly retaliation. When market expansion and the development of political institutions reflect these kinds of CCIs, market expansion will further strengthen them and enhance their influence.

32.4.1 The Demand for Public-Order and Designed CEIs

Organic, private-order CEIs are universal, reflecting responses to gains from exchange. Although these organic CEIs reflect the need for trade they also reflect social and cultural factors. Initial social structures demarcate and verify membership and provide networks for information transmission. Cultural beliefs and behavioral norms coordinate expectations and provide a shared understanding of the meaning of various actions. *Ceteris paribus*, initial social structures and cultural features therefore influence which, among the many possible organic CEIs, will emerge, become an integral part of these institutions, and be reproduced by them (Greif, 1994a, 1996).

Empirical and theoretical work indicates the relevance of the extent to which a society is more “communalist” or more “individualist.”²⁶ In the former case, larger, innate social structures that are based on kin, place of birth, or religion (e.g., lineage, tribes, or religious sects) are prominent and members of those societies feel involved in the lives of other members of their group. In more individualist societies, the individual and family, rather than the larger, innate social structure, are prominent and individuals expect that others will interfere relatively little in their affairs. The more communalist a society is in the initial stages of market development, the more, *ceteris paribus*, its organic CEIs will be based on each intra-group’s economic and

²⁶This discussion is based on Greif (1994a) in which I used the term ‘collectivism’ instead of ‘communalism.’ ‘Communalism’ was suggested by Timur Kuran and Joel Mokyr. For works indicating the importance social and cultural factors see Granovetter, 1985; Clay, 1997a, b; McMillan & Woodruff, 2000; Moriguchi, 2003; Stulz & Williamson, 2003; Biggs et al., 2002; Fafchamps, 2004; Olds & Liu, 2000.

social sanctions among its members. The society will be more “segregated:” each individual will interact socially and economically mainly with members of his group. The more individualistic a society is in the initial stages of market development, the more its organic CEIs will be based, *ceteris paribus*, on bilateral economic and social sanctions among individuals and families. The society will be more “integrated:” economic transactions will be conducted among people from different groups. In either case, as discussed above, the symbiotic relationship between organic CEIs and their underlying cultural and social foundations will lead to their mutual reinforcement.

Organic CEIs that reflect communalism and imply segregation generate relatively weak demand for public-order and designed CEIs. The ability of each social group to punish its members reduces the relative cost of intra-group economic exchange while the thinness of intergroup exchange reduces the benefits to each individual from leaving his group and pursuing outside exchange. Even if economic efficiency calls for institutional development, it may nevertheless not be rational for each individual agent to pursue the change. In contrast, organic CEIs that reflect individualism and imply integration, imply relatively weak ability of each group to discipline its members. Such CEIs therefore generate a relatively strong demand for public-order and designed CEIs (Greif, 1994a; Kranton, 1996b).

Initial CEIs thus influence the extent to which the economic agents find it profitable to use, if available and establish, if possible, public-order and designed CEIs. The “demand” for these CEIs reflects more than non-institutional, environmental, and technological conditions.

32.4.2 The Supply of Public-Order and Designed, Private-Order CEIs

Demand for public and designed CEIs will not necessarily be met, however. Using public-order CEIs reveals wealth to those with coercive power. Appealing to the court, using a land registry, applying for a business license, or submitting to a regulatory agency generates valuable information that can be used to identify and capture a person’s wealth. Public-order institutions can be *effectively* supplied only if revealing wealth to those with coercive power does not undermine the effectiveness of the CCIs and hence the security of property rights. If this is not the case, even if public-order CEIs are established, they will be underutilized. Distinct CCIs imply different extents to which public-order CEIs can be *effectively* supplied.

Although the relationships between various CCIs and effectively supplying public-order CEIs have not been systematically studied, tentative conjectures can nevertheless be advanced. Consider first CCIs based on the state’s administrative structure. An absent state has a limited ability to provide wealth-revealing, public-order CEIs without undermining property rights’ security. It can provide weak public-order institutions which do not increase its ability to abuse rights supporting social order and personal safety. Delegating the state’s administration to those in need of wealth-revealing, public-order CEIs entails a greater ability to effectively supply them. The property rights of the administrators are secured and, if they either

make a living from providing these administrative services or benefit from commercial expansion, they can better commit than the ruler to respect others' property rights. For similar reasons, self-governance by those in need for wealth-revealing, public-order CEIs entails an even greater ability to provide such institutions, particularly on the local level.

Consider now coercion-constraining institutions based on balancing coercive power. When a balance of coercive power among social units constrains its use, there is no third party with coercive power to back the operation of the public-order institution.²⁷ A ruler with independent coercive power is able to more effectively provide public-order institutions although his ability is limited because revealing wealth increases the ruler's ability to gain from abusing rights. This increased ability to gain from abusing rights undermines the previous balance of coercive powers. A ruler whose conditional coercive power is provided by those in need of wealth-revealing public-order CEIs is able to even more effectively provide these institutions. Unlike a ruler with independent coercive power, one with conditional coercive power can be punished by the asset holders following an abuse.

Table 32.1 summarizes the inter-relationships between various organic CEIs and CCIs and the demand for and ability to effectively supply public-order CEIs.

Different initial combinations of organic CEIs and CCIs therefore imply distinct institutional dynamics. In particular, organic CEIs reflecting and entailing communalism and segregation combined with an absent state or a balance of coercive power among social units imply low demand and low ability to effectively supply wealth-revealing public-order CEIs. On the other hand, organic CEIs reflecting and entailing individualism and integration imply a high demand for public-order CEIs. These can be particularly effectively supplied if the CCIs are based on self-governance by those who stand to gain from public-order CEIs who also have a coercive power that balances the ruler's.

Similar analysis applies to designed CEIs. Demand for them reflects initial organic CEIs. Many designed CEIs reveal wealth to those with coercive power. The wealth of the guild, the lists of stock holders, the records of transactions in the stock exchange, and the information stored by the credit card companies or credit bureaus, reveal who has wealth and in what form. Furthermore, many designed CEIs operate in the shadow of the law or are quasi-private, as discussed above, and hence their operation depends on public-order CEIs. The ability to effectively supply wealth-revealing designed CEIs is a function of CCIs. Such CEIs will be established and used only by those with the countervailing economic or coercive power required to protect their assets. Others will only use CEIs that neither reveal wealth nor rely on public-order.

Table 32.1 does not reflect two more considerations. First, the demand for public-order and designed CEIs can be more effectively met, *ceteris paribus*, the more a

²⁷ On the demand side, the expectation that a business dispute among members of different social units can lead to a military conflict among them undermines incentives to enter into such transactions.

Table 32.1 Supply and demand for public-order CEIs

<i>Coercion-constraining institutions as determinants of ability to effectively supply public-order CEIs:</i>			<i>Organic, private-order CEIs (technological factors) as determinants of demand for public-order CEIs:</i>	
			<i>Low</i>	<i>High</i>
	CCIs based on administrative structure	CCIs based on balancing coercive powers	Communalism/segregation (lower demand if small, static economies and low gains from impersonal exchange)	Individualism/integration (higher demand if large, dynamic economies and higher gains from impersonal exchange)
<i>Low</i>	Absent state: Thin administration controlled by the state	Mutual deterrence: No ruler. Asset holders coercive power mutually deters abuse	Low demand, low ability to effectively supply. E.g., <i>China under the Qian and the first Muslim empire</i> <i>Communalism and an absent state</i>	High demand, low ability to effectively supply
<i>Medium</i>	Delegation: The state's administration is controlled by the asset holders	A ruler with independent coercive power: Balanced by the asset holders coercive power	Low demand, medium ability to effectively supply	High demand, medium ability to effectively supply
<i>High</i>	Self-governance: Administration provided by autonomous units controlled by the asset holders	A ruler with conditional coercive power: Asset holders provide the ruler with coercive power	Low demand, high ability to effectively supply	High demand, high ability to effectively supply. E.g., <i>England: Individualism. Autonomy with economic and coercive powers. A ruler with conditional coercive power</i>

ruler's coercive power is checked by the limited reach of his grabbing hand. E.g., the ability of the asset holders to move their assets elsewhere following an abuse enhances their ability to reveal wealth without fearing losing it. Second, those with coercive power will permit designed CEIs to be established only if their benefits outweigh the countervailing economic or coercive power that these CEIs entail.

The above discussion emphasizes that institutional dynamics reflect initial—organic, private-order and coercion-constraining—institutions. But technological and other non-institutional factors also influence these dynamics particularly by influencing the relative efficiency of various CEIs and hence their demand. This

demand would be, for example, higher in dynamic economies (where there is more to gain from an increased ability to quickly respond to changing needs) and when there is more to gain from impersonal exchange characterized by separation between the *quid* and the *quo* over time and space. We do not have good empirical evidence or theoretical understanding regarding the relative importance of institutional and non-institutional factors in determining demand for various CEIs.

32.4.3 Market Expansion and Political Development

For the reasons discussed above, the CCIs fostering the supply of wealth-revealing, designed and public-order CEIs also aid the development of political institutions in which the commercial sector has voice and influence. These political institutions, in turn, play an important independent role in market expansion by providing information and knowledge required to bring such CEIs about. In markets, discrepancies between demand and supply are filled as individuals respond to the information conveyed and the motivation provided by prices. In the absence of a price system, however, the demand for institutions has to be directly communicated to those with the ability to respond to it. Market development is thus fostered by a polity in which the commercial sector has a voice and influence on the function, policy, and organization of the state. This influence is a reflection of CCIs. The commercial sector will have political representation if it has the coercive and economic powers to balance that of the ruler. Also, as discussed above, these CCIs are also likely to make possible the exchange between the commercial sector and a ruler required to motivate him to provide such public-order CEIs.

Furthermore, when these CCIs exist, market expansion will lead, *ceteris paribus*, to further strengthening the economic and coercive power of the commercial sector, and the development of political institutions in which the commercial sector has voice and influence. When they do not exist, though, commercial expansion is not likely to lead to such political development. Indeed, as further discussed below, in societies such as the medieval Islamic empire or pre-modern China and Spain, the commercial expansion that occurred failed to lead to liberal governance. The opposite occurred, however, in England.

32.5 Reflections on the Historical Development of Markets and Political Institutions

A comparative analysis of the institutions that supported markets in various historical episodes, their extent, and dynamics has yet to be conducted. Yet, our current state of knowledge suggests the merit of the above conjecture. For example, it postulates that a market can thrive under a state that is absent from the commercial sphere, a ruler with independent coercive power constrained by the coercive power of the masses and coordinated by observing abuse, and communalist and segregated organic CEIs. Such a market, however, would only have a limited ability to extend as

there is weak demand and ability to effectively supply public-order and designed CEIs. This may very well have been true of Imperial China.

Indeed, pre-modern Chinese markets—at least as measured by market integration in grain—were no less developed than Europe's as late as the nineteenth century (Shiue & Keller, 2003) and standards of living in various areas within China were comparable to Europe's (Pomeranz, 2000). The Chinese state was active in providing public goods, such as defense, famine relief, commercial infrastructure, and distribution of knowledge regarding better agricultural techniques (e.g., Pomeranz, 2002). The Empire had an effective administration and a long tradition, dating back at least to the Zhou dynasty (1122–256 BC) of the legal enforcement of contracts, particularly those regarding assets relevant to the state's revenues like land (e.g., Zelin et al., 2004).

At the same time, many students of China's institutions have provided evidence suggesting that its organic CEIs reflect communalism and lead to segregation, and its CCIs were based on a state absent from the commercial sphere that provided only weak public-order CEIs. The administration of the state was heavily centralized yet thin and did not extend below the roughly 1300 county magistrates. Beneath each of these magistrates “were several towns and hundreds of villages and a population ranging from several tens of thousands to hundreds of thousands” (Wong, 1997, p. 108). In the commercial sector, property rights were secured by an absent state and commercial taxation was low. In the agrarian sector, property rights were secured by balancing the state's coercive power with that of the numerous peasants (Yang Li, 2002) and balancing the power of the local elite by that of the state (Wong, 1997; Yang Li, 2002). The state fostered the creation of shared beliefs regarding appropriate taxation by announcing that it would never be raised and distributing information regarding its level in the villages.

Even during most of the last dynasty, the Qing, there was no commercial code of law, and it was administered by magistrates, persons of literary and philosophical learning with multiple duties unrelated to the law. These magistrates were subject to heavy penalties if they made mistakes and hence, they sought compromises rather than legal rulings. Organic CEIs were central to the operation of the market. In comparing them to the European CEIs, Hamilton (1991) noted the large extent to which they reflected communalism, and that historically “the Chinese society consiste[d] of networks of people whose actions are oriented by normative social relationships” (Hamilton, 1994, p. 199). In particular, lineage was the social structure around which business organizations were formed.

As the economy grew, lineages responded to changing needs and opportunities by becoming more designed. Their economic organization was often based on contractual relationships, had centralized bureaucracies, and drew on outsiders' resources and talents (e.g., Herrmann-Pillath, 1999; Redding, 1991). In summarizing the related vast literature regarding late Imperial China, however, Herrmann-Pillath (1999) noted that it was the relative absence of the state from the commercial sphere that hindered further development and led to an institutional evolution that was different from Europe's. Similarly, although Pomeranz (2000) highlighted the role of distinct natural endowments in enabling Europe to economically overtake China, he noted that the Chinese state interfered much less than the European states in the operation of the market.

The Qing's responses to the military and economic conflict with the West during the nineteenth century lends support to the conjecture regarding the institutional foundations of its markets. Constrained from increasing the land tax, the Qing resorted to taxing goods in transit because merchants along the road, being few in number, did not pose a threat, promoting a "top-down" industrialization, and providing public-order CEIs. Departure from an absent state without creating a countervailing economic or coercive power, was counter-productive. Corruption prevailed, trade suffered, and designed CEIs, such as the stock exchange, were not established (Yang Li, 2002; Goetzmann & Köll, 2003). The state resorted to expanding the role of guilds, delegating to them such functions as commercial tax collection and provision of local public goods. Top-down industrialization and monopolistic guilds, however, hindered industrialization and commerce (Goetzmann & Köll, 2003; Ma, 2004). Even the famous Chinese silk industry fell behind Japan's.

Political development in China is consistent with the above theoretical conjecture. At least under the Qian (1644–1911), prior to the Opium War (1840), the state's budget constraint was not binding and there were no CCIs empowering the commercial sector. Indeed, representative bodies were not established in China. Furthermore, consistent with the argument that rulers fear large-scale economic organizations, economic corporations were not legal entities, and guilds, although known, were few and relatively weak until the late nineteenth century when the state delegated various functions to them in response to the fiscal pressure from the conflict with the West.

On the other end of the spectrum, the above conjecture implies that individualism and integration create demand for public-order and designed CEIs. These can be effectively supplied if there is a ruler constrained by a commercial sector with self-governance which, in addition, has coercive power to render the ruler one with a conditional coercive power. In this case, if the ruler's budget constraint is binding and his need for resources fluctuates, he would establish a representative body that includes the commercial sector. Arguably, this was the case in pre-modern England. Its most distinguishing features were individualism, the autonomy—self-governance and military ability—of its commercial sector, and its rulers' weak independent administrative capacity and lack of sufficient revenues.

While I return to the case of England below, it should be noted that its market economy was arguably not a match for China's circa 1000 AD. By the nineteenth century, however, it was the forerunner and symbol of the emergence of the modern market economy. England is well known to have been individualist, at least from the late medieval period (Macfarlane, 1978). The autonomy of its cities and parishes, which were controlled by either the commercial or landed sectors, implied both self-governance and a military ability to check the coercive power of rulers. At the same time, a ruler with conditional coercive power constrained the autonomous cities to compete with each other economically, but not militarily, while coordinating common policies. The budget constraint implied by these CCIs, and the inter-state European competition fostered the creation of an effective parliament that facilitated resolving collective action problems associated with providing public-order institutions and

other public goods. Public-order CEIs were established, used, and expanded. Markets expanded increasing the relative power of the commercial elite and hence fostering institutional trade. But prior to the rise of democracy, consistent with the argument advanced here, property rights' protection was not universal. Even after the Glorious Revolution of 1688, protection was afforded to the landed, commercial, and financial elite. Only those who had economic and military power received protection.

The rise of the modern market economy, however, was a European (or Western) phenomenon. It neither began in pre-modern England nor was it stopped from relatively quickly expanding in the rest of Europe. This observation is consistent with the claim regarding European individualism, which arguably fostered organic CEIs that generated demand for public-order and designed CEIs. Furthermore, the re-emergence of polities in Europe following the collapse of the Roman and the Carolinian Empires transpired in the context of rulers with binding budget constraints (due to inter-state competition) and self-governed units controlled by asset holders with coercive and economic powers. By the late medieval period representative bodies, constitutional and limited monarchies were the rule in Europe alongside republican polities.

Although every society has individualistic and communalist elements, and categorization is a matter of their relative importance, a long line of research has emphasized the relative pervasiveness of European individualism.²⁸ Individualism is considered a heritage of ancient Greece (e.g., Hsu, 1983; Gurevich, 1995) and early Christianity encouraged it by placing the individual rather than his social group at the center of its theology. It advanced the creation of "a new society, based not on the family but on the individual, whose salvation, like his original loss of innocence, was personal and private" (Hughes, 1974, p. 61). From as early as the fourth century, the Church was also systematically engaged in weakening kin-based organization of society by prohibiting marriages among kin (sometimes up to seventh degree!) (Goody, 1983). By the late medieval period, Western individualism manifested itself in such diverse ways as war tactics, the emergence of confession, nicknames, and landholding (e.g., Morris, 1972; Macfarlane, 1978).

It is important to emphasize that no society is composed of "atomistic" individuals and even in such individualistic contemporary societies like the USA, social and business networks are important. Similarly, European history provides many examples of institutions based on multilateral punishment among neighbors and business associates (e.g., Muldrew, 1998). By and large, however, these did not reflect innate, kin-based social structures. They either reflected economically motivated processes through which such organic CEIs emerge, as discussed above, or were designed CEIs.

Late medieval, European coercion-constraining institutions reflect a dispersed distribution of coercive and economic powers and states' meager administrative

²⁸ Various studies, however, used the term individualism differently. For communalism in China see Hamilton 1991. For segregation in the Islamic world, see Lapidus, 1984 and Hodgson, 1974. Goitein, 1955 noted that the organic CEIs among the Maghribis represent those of the Muslim world.

capacities. After the disintegration of the Roman and Carolinian Empires the rulers of the emerging polities had relatively weak coercive power and administrative capacities. The size of European armies was small in absolute and relative terms. Frederic Barbarossa (d. 1190), the Emperor of the Holy Roman Empire, the King of Germany, and the King of Italy sailed to the third Crusade with several hundred knights. As late as the early fifteenth century, European armies were only a few thousand strong, the largest one (France), numbering circa 40,000 in 1470 (Downing, 1992, p. 69). Balance of coercive power between rulers, great lords, and cities, and administrations based on delegation and self-governance—particularly in the form of autonomy—was the rule.

Indeed, the polity under the European Feudal system was one in which a ruler was a coordinator of coercive power, who was further constrained by the self-governance of its subordinated units, feudal lords, and cities alike. These subordinated units had the wealth, technology, organization, and manpower required to sustain military strength in a period in which, for a long time, defense was superior to offense. No wonder that the period's ideology considered the ruler's job to be merely the maintenance of social order—providing the balance of power among his vassals and coordinating their joint protection and other endeavors. As noted by the historian of Frederic Barbarossa, his “duty was merely to protect all the subjective rights everybody had. It was not his business to issue laws of his own. . . he was supposed to play a purely passive role as law protector.” (Munz, 1969, p. 100). Indeed, Barbarossa had little independent military might. He had to stop his military campaign against the Normans of southern Italy in 1155, for example, because his vassals declared that they served their time for that year.

Medieval England similarly reflects the essence of the feudal king as a coordinator of power but also the need for constantly refining the associated institution to maintain a balance of power. As noted by Tilly (1990, p. 154):

in the process of making war and intervening in dynastic rivalries, the barons on whom the English king relied for their wars acquired enough power to fight the king as well as each other, exacting chartered concessions—most dramatically the Magna Carta—from the monarch. The Great Charter of 1215 committed the king to cease squeezing feudal obligations for the wherewithal to conduct wars, to stop hiring mercenaries when barons would not fight, and to impose the major taxes only with the consent of the great council, representative of the magnates.

The importance of the underlying balance of coercive power is reflected in a clause in the charter delivering several castles to the barons and the prohibition on a mercenary army.

Increasing peace brought about by the feudal order, population growth, lords' desires to gain from trade, and attempts of kings and lords to strengthen their positions vis-à-vis each other, fostered urban growth.²⁹ Changes in military

²⁹The ‘Peace of God’ movement also played a role in bringing peace by coordinating the countervailing coercive power of economic agents. E.g., Head & Landes, 1992.

technology, particularly after the eleventh century, shifted the balance of coercive power to the masses and urban dwellers and away from the armored knights. Among these changes were the reintroduction of stone walls, the invention of the crossbow, the introduction of the longbow and the pike-based, heavy infantries. Once cities grew in population and wealth, they were able to gain military might comparable to that of lords and hence retain autonomy and gain rights.

These changes and the emergence of a new balance of coercive power are well reflected in the rise of representative bodies and republican movements throughout Europe by the twelfth century. By then republican movements swept European cities, particularly in Northern Italy, whose ruler, the Emperor, was weakened due to its military conflict against the Pope that was a part of the Investiture Controversy. But even the Pope had to confront a militant republican movement in Rome itself. Representative bodies were established throughout Europe, in England, Spain, France, and Flanders among other states. Even the Emperor of the Holy Roman Empire, Charles IV, had to issue a Golden Bull (1356) detailing the rules governing the empire as a constitutional monarchy.

Various rulers' relative military weakness during this time (prior to the fifteenth century Military Revolution) is well reflected in cases in which they overestimated their ability to extract taxes from their subjects. After the Empire attempted to tax the cities of Northern Italy, they broke away from it militarily gaining *de facto* independence and *de jure* freedom in the Peace of Constance (1181). The Swiss Confederation was established in 1291 in response to what residents of several cantons of the Holy Roman Empire considered inappropriate taxation.

In this context of self-governance, rulers with conditional coercive power provided by the economic agents, and economic agents with political representation, the necessary conditions for providing public-order and designed CEIs were met. Indeed, the late medieval period witnessed legal revival, the establishment of public-order and designed CEIs, and market expansion. This legal renewal also reflects other pan-European processes such as the conflict between the secular authorities and the Church (Berman, 1983); its details may have been shaped by the need to protect judges from intimidation (Glaeser & Shleifer, 2002); and it has been facilitated by the Roman legal tradition and the associated concept of designed, man-made laws.

Yet, the function—serving the economy—and implications—creating public-order and designed CEIs—of this legal revival are illustrated by the fact that the legal foundations of the modern business corporation were laid in this period. The modern corporation was created through the fusion of the late medieval joint-stock company, which was in fact a partnership, and the traditional legal form of the corporation as it was developed during the medieval period. More broadly, the contemporary European laws and practices regarding commerce, bankruptcy, insurance, apprenticeship, patents, and banking originated then. Designed CEIs, as well as hybrids between private and public ones, were established throughout Europe, taking such forms as guilds, municipalities, monasteries, universities, insurance fraternities, banks, and large-scale partnerships and family firms. This led to further innovations and practices, such as trading in shares, limited liability, auditing, and various accounting procedures. The invention of public debt that served Europeans

well for centuries to come also attests to the existence of effective CCIs. A necessary condition for public debt is that the state can commit to repay the wealth that was placed in its custody.

Northern Italy, free from any ruler, emerged as Europe's leader in institutional, organizational, and legal innovations that fostered commercial expansion. This freedom reflected the coercive power they were able to obtain based on their intra-cities CCIs, the Investiture Controversy between the Emperor and the Pope that weakened both, and the commercial opportunities in the Mediterranean sea. In the long run, however, Northern Italy declined not least because most cities, like Genoa, failed to establish institutions that were able to constrain violence for long. Internal violence plagued these cities and whoever gained control attempted to make the most of it through profitable but inefficient policies, such as providing guilds with monopoly rights and forestalling technological advances. Later, the Italian city-republics fell prey to the mercenaries they brought in to fight their inter-city wars and to the feudal agrarian lords which the cities never fully defeated.

By the fifteenth century, the Military Revolution had led to external domination over northern Italy. Cannons, introduced to the Italian battlefields during the 1494 French invasion, made the thin and tall medieval city walls a frail defense. Firearms enabled equipping and training larger armies than had been possible before, providing an advantage to larger and richer states. The Italian city states failed to coordinate responses against such invasions, arguably due to the rapidity of the Military Revolution and the history of military confrontations among themselves. In the absence of effective CCIs, markets and economic vitality subsided.

This was not true in Flanders, which, together with northern Italy, experienced large-scale urban growth, and autonomy during the late medieval period. It began with the fortifications of towns against the Norman invasion and gained momentum when these cities began to process wool exported from England. Several times during this period, the prosperous cities of this area conflicted with their feudal overlords to gain and retain independence but achieved only self-governance. CCIs based on self-governance and a ruler with independent coercive power prevailed. Unlike northern Italy, therefore, these cities were protected from an external invasion during the initial stages of the Military Revolution. Within Flanders itself, however, the prosperity of cities shifted the balance of coercive power from the ruler to their favor. In 1463 Philip the Good created a representative body, the States General, which enacted laws and had the authority to vote on taxation. European commerce shifted to Flanders which became the center of innovations in public-order and designed CEIs such as the first European bourse (in Bruges) the stock market in Amsterdam and increasing transferability of bills.

When Charles V, the King of Spain and the Holy Roman Emperor, inherited Flanders in the early sixteenth century, its administration was based on autonomous and well-coordinated cities. The military revolution implied that wealth could buy military might more than before while developments in fortification techniques restored the balance between defense and offense. The growth of the cities therefore shifted the balance of coercive power in their favor. As noted by the prominent historian of Flanders, Israel (1995), the very success of the economy of Flanders

posed a danger to the Hapsburg regime. Yet, given the CCIs that limited the ability of the Habsburgs to tax their German possessions (which were, as noted above a constitutional monarchy), the King gambled on pressing Flanders to pay more to finance his religious and other wars. The resulting tax revolt (1579) turned into a war of independence in which northern Flanders became the Dutch Republic and the center of European commerce during the seventeenth century.

The internal organization of this republic, however, was such that it was ill suited to mobilizing resources to wage effective wars elsewhere. In 1651 England passed the first of the Navigation Acts that were directed at undermining Dutch commercial dominance by shifting trade and freight to England and its shipping industry. The chief provisions were that no goods grown or manufactured in Asia, Africa, or America could be transported to England except in English vessels, and that the goods of any European country imported into England must be brought in British vessels, or in those of the country producing them. In the subsequent wars, the Dutch Republic failed to reverse these acts.

While the institutional history of England was not a linear progression toward institutions favorable to market extension, it nevertheless can be characterized as having a ruler who was a coordinator of others' coercive power and an administration provided by autonomous, particularly commercial units controlled by economic agents who had actual or potential military power. Strong feudal lords, autonomous cities, a ruler without a standing army, and a parliament that both approved taxation and coordinated political actions are the manifestations of these coercion-constraining institutions.

Initially, William the Conqueror and his immediate successors faced the challenge of restraining the coercive power of the lords. The military weakness of the Crown is reflected in the Magna Carta (1215) as noted above and the king's obligation not to have a standing royal army, not to recruit mercenaries, and not to tax without consent by the Great Council. That Council, in which the nobles were represented, reflected the prevailing CCIs: a balance of military power between the Crown and the lords.

In the context of the conflict between the Crown and the lords, particularly during the thirteenth century, the former gave charters to numerous English cities. By the end of the thirteenth century, there were about 500 such autonomous, self-governed, towns (boroughs) that became an integral part of the kingdom's administration. Furthermore, they had the wealth, manpower, organizational infrastructure, and production capacity to have potential and actual military power. Alongside the self-governed parishes, these towns changed England's CCIs. Indeed, chartering cities had arguably been a strategic response by the Crown, aimed at diverting tax revenues away from the lords and creating a countervailing power.

That these towns became part of England's CCIs is well reflected in the events surrounding the transformation of the Great Council into a Parliament in which the towns' dwellers were represented. In 1265 the King Henry III dissolved the Great Council and levied unapproved taxes. The effectiveness of the CCIs balancing the Crown's power with that of the lords was challenged. Earl Simon de Montfort responded with a revolt, during which he called a meeting in which the nobles, the clergy and representatives of the counties and towns were present.

Although de Montfort and his army were eventually defeated, the event reflects the increasing importance of the counties and towns' support in national conflicts. Recognizing that they were part of the CCIs he was facing, King Edward I summoned the so-called "model parliament" (1295), which included, for the first time the representation of the commercial sector: two burgesses from each borough and two citizens from each city. It was during this parliament that the Crown issued a charter seceding the right of approving new taxes to the parliament. In 1297 Edward I confirmed the Magna Carta, asserting that it should be observed as common law, and declared that on no account were aids and taxes to be taken without the common assent of the whole kingdom and for the common benefit. Yet, during that period the Crown abused the rights of Italians traders. Only the rights of those with a countervailing power were respected.

CCIs that constrain the power of the state based on the administrative and military power of the commercial sector enable effective provision of public-order CEIs and establish designed CEIs. Indeed, the latter half of the thirteenth century was a period of reform and expansion of English law and the legal system. Edward I is known as the "lawyer-king" (Hogue, 1996, p. 69) and his legislation directly influenced the extent of the markets.

Recall, for example, that by the thirteenth century, the Community Responsibility System (CRS), based on the legal autonomy of the English towns, enabled impersonal exchange characterized by separation between the *quid* and the *quo*. During that century, the CRS began to decline due to the commercial expansion and the growth in the size, number, and economic and social heterogeneity of towns. The CCIs that restricted the Crown's power, at that time, however, were such that it was possible to replace the CRS with a state wide legal process for placing collateral and its collection. This, however, was not the case in other parts of Europe. In Germany, for example, CCIs that balanced the central authority, and local lords were no longer an equilibrium due to the Investiture Controversy. Although the CRS declined, a suitable alternative was not provided (Greif, 2004; Volckart, 2001).

The development of the Common Law courts further constrained the Crown with these courts perceived rights for independence in areas where they acquired customary jurisdiction. Infringing on the jurisdiction of these courts was considered an abuse of rights.³⁰ By the fifteenth century the ability to effectively provide public-order CEIs was fostered by these semi-independent courts. This situation and its implications regarding the provision of partial public-order CEIs is illustrated in the first known court case in England regarding negotiable credit instruments (1436). The London Mayor's court at Guildhall had customary rights in cases involving merchants, but one of the parties approached the King's Bench to transfer the case to its jurisdiction. The Mayor of London, however refused to consent to the Bench's demand, arguing that "according to the Law Merchant and the ancient liberties and

³⁰Exactly how corruption was prevented in these courts is not clear. Arguably, internalization of values, social pressure, compensation by the litigants, and competition among courts over cases played a role.

free customs of the city itself. . . the mayor. . . have the power and use of hearing” such cases (Munro, 1990, p. 74). The king withdrew his demand, and the negotiated credit instrument became legal. By the 16th and 17th centuries such public-order CEIs enabled the expansion of credit and bonds beyond that possible based on reputation alone (Muldrew, 1998).

Subsequent events further reduced the power of the lords and increased the relative ability of the commercial sector to constrain the king’s power. The increasing financial needs of the Crown due to the Military Revolution compelled it to sell its landed properties making it more dependent on tax contributions. The War of the Roses (1455–85) decimated the ranks of the great lords making the Crown more dependent than ever on the administrative capacity and other resources of its autonomous towns and the local unpaid Justices of Peace. The Crown confiscated the Church’s large land holdings in the sixteenth century in the context of establishing the Church of England, but its coercive power was sufficiently constrained at this point that it was unable to use this resource to undermine the existing CCIs. On the contrary, the Crown’s binding budget constraint compelled it to sell this land to the gentry, thereby further strengthening them. Moreover, the greater efficiency in which the gentry utilized the land probably further constrained abuse by the larger implied loss of value (Rajan & Luigi, 2003). The flow of wealth from the emerging Atlantic trade may have had a similar impact (Acemoglu et al., 2002).

The Civil War of the seventeenth century was another step in this process of institutional evolution. The war made it evident that the crown neither had the independent military ability nor the administrative capacity required to rule without the consent of the economic elite which was also the military elite. The Crown was a ruler with conditional coercive power and the supremacy of the Parliament was firmly established. The zenith of formalizing this situation occurred during the Glorious Revolution (1688). New rules coordinating on appropriate behavior by the Crown, such as the Bill of Rights, a better separation between the judiciary and the executive, and new organizations, such as the Bank of England, formalized and fostered this situation (North & Weingast, 1989). This essence of the Glorious Revolution—that reflected a *de facto* prior situation—accounts for the puzzling observation that the historical evidence does not indicate that the Glorious Revolution altered the security of property rights.³¹

Indeed, once the parliament gained supremacy, it was not in the business of protecting property rights *per-se*. Its policy reflected the interests of those who controlled it, namely, the landed, commercial, and financially elite. The subsequent history is thus marked by gross abuses of property rights through the radical increase in taxation, monopolies, parliamentary enclosures of the open fields, and colonial expansion.³² (Harris, 2004. For economic analyses of the great English trading

³¹ Clark, 1996; O’Brien, 2001; Quinn, 2001; Sussman & Yefeh, 2000, 2004; Harris, 2004.

³² Harris, 2004. For economic analyses of the great English trading companies which were monopolies. See Irwin, 1988; Carlos & Nicholas, 1996; Carlos, 1992.

companies which were monopolies. See Irwin, 1988; Carlos & Nicholas, 1996; Carlos, 1992.) Yet, a state controlled by its landed, commercial, and financially elite and later empowered by the Industrial Revolution was a boon for the extension of markets. The evolution of the modern market reached its zenith.

Other European states began, under feudalism, with initial conditions similar to England's. Their subsequent institutional evolution, however, differed due to such factors as the greater risk of invasions (France), discoveries that provided resources for the Crown (Spain), and feudal lords too strong for the Crown to constrain in the absence of autonomous cities to do so (Germany). These distinct experiences illustrate how precarious England's institutional evolution was. But Europeans shared a common heritage of individualism, self-governance, a broad distribution of coercive powers, and man-made laws. Reversing their institutional developments and enabling market extension was relatively easy.

32.6 Concluding Comments

Contract-enforcement institutions provide the foundations of markets, and their details, expansion, or contraction determine the market's extent and dynamics. Contract-enforcement institutions changed due to environmental changes that influenced various institutions' relative economic efficiency and profitability. In addition, however, as this chapter has argued theoretically and demonstrated historically, the dynamics of contract-enforcement institutions are also a function of initial, organic, contract-enforcement institutions and coercion-constraining institutions. New institutions emerge or are established in the context of existing ones.

Distinct, initial, organic, contract-enforcement institutions generate different demands for additional, public-order and designed, private-order, contract-enforcement institutions that can further extend the market. Distinct, initial, organic institutions can emerge in the initial stages of market formation due to different economic conditions and social and cultural factors, such as individualism and communalism.

Hence, efficiency-promoting, contract-enforcement institutions will not be utilized or have the expected impact even if they are introduced, because they are not compatible with existing, organic, private-order institutions. These institutions create a wedge between the institutions that, if utilized by everyone, will extend the market, and the institutions that each individual, given the existing organic institutions, will find optimal to utilize. Initial, organic, contract-enforcement institutions generate different demands for additional institutions by influencing the extent to which economic agents will find these institutions profitable to use if available and establish if possible.

But even if there is a demand for public-order and designed, contract-enforcement institutions, their effective supply will not necessarily be forthcoming. Utilizing these institutions implies revealing wealth to those with coercive power, implying an increase in the risk that it will be expropriated. The extent of this increased risk, however, depends on the existing coercion-constraining institutions. For markets to function, property has to be protected from those with coercive power. When this is

not the case and economic agents are subject to predation, markets are confined to the exchanges that are possible based on private-order, contract-enforcement institutions that are more successful in mitigating the threat posed by coercive power.

More generally, public-order and pragmatic, private-order, contract-enforcement institutions can be supplied only if particular coercion-constraining institutions already prevail—specifically, those whose effectiveness is not undermined by revealing wealth. In particular, coercion-constraining institutions based on the absence of the state from the commercial sphere have a limited ability to effectively supply public-order and pragmatic, public-order institutions. Furthermore, when a ruler's legitimacy is weak and his budget constraints are not binding, he is likely to consider large-scale, private-order institutions more as a threat than a benefit.

In contrast, coercion-constraining institutions based on self-governance by the market participants, and a ruler with conditional coercive power provided by these participants are favorable for effectively supplying public-order institutions and establishing pragmatic, private-order institutions. Where such coercion-constraining institutions prevail, and the ruler's budget constraint is binding, political development associated with the rise of a liberal state will transpire. A state effectively constrained by the commercial sector will have to provide it with political representation, respect property rights, and will pursue market-enhancing policies.

Historical evidence suggests the merits of this analysis. The pre-modern market leaders—China and the Muslim world—had organic, contract-enforcement institution and coercion-constraining institutions that were neither conducive to the rise of public-order and designed, private-order, contract-enforcement institutions, nor giving political representation to the commercial sector. In contrast, England, considered by many as the initial role model for the modern market, had organic, contract-enforcement institutions and coercion-constraining institutions conducive to the rise of public-order and designed, private-order institutions, as well as representative bodies and a constitutional monarchy. This observation regarding the co-evolution of economic and political institutions is consistent with the argument that at least the security of property initially reflected “might rather than right.” England was no stranger to abusing rights through, for example, Parliamentary enclosures and the exploitation of colonies.

The public-order and designed contract-enforcement institutions that were initially developed in Europe are considered necessary for market development. It is also perceived that markets follow the creation of a constitutional state with effective administration and public-order institutions. Markets are assumed to follow the creation of public-order. Historically, however, it seems that limited government, representative bodies, and modern markets co-evolved in a process reflecting deeper institutional variables. These variables included particular, organic, contract-enforcement institutions and coercion-constraining institutions that revealed and embodied different cultural and social factors, and the particularities of pre-modern production, communication, transportation, and military technology.

If this is the case, development strategies need to be reconsidered. The same factors, such as communalism and segregation that led developing economies on distinct institutional trajectories may still prevail, implying that their institutional

needs are distinct from those that were developed in the West. Indeed, contemporary social psychologists have found that most of the developing countries are communalist, whereas the developed West is individualist (e.g., Bellah et al., 1985; Reynolds & Norman, 1988; and Triandis, 1990).

The details of the optimal contract-enforcement institutions in a communalist/segregated society, however, are arguably distinct from those developed in an individualist/integrated society and some empirical work suggests this is the case (Ensminger, 1997; Goldstein & Udry, 2002). Furthermore, market expansion, economic growth, and the development of more costly contract-enforcement institutions in the West has been a process rather than an event. The attempt to duplicate these costly institutions in poor economies may very well be like placing the wagon in front of the horses. Indeed, past European public-order institutions were very distinct from contemporary ones. The experience of the Community Responsibility System suggests, for example, that the optimal unit of non-contractual legal liability is not necessarily the individual. Levinson (2003) has advocated changing legal concepts accordingly.

Similarly, development policy has been predicated on the assumption that growth requires creating a western-style state with extensive administrative capacity. In the West, however, such a polity has been the end result of institutional and economic co-evolution rather than its beginning. In particular, central to this co-evolution were coercion-constraining institutions based on self-governance by the economic agents and widespread, yet locally organized, distribution of coercive power. The related constitutional polities had representative bodies that could effectively restrict the coercive power of rulers and each other by drawing on these resources. The creation of a western-style state with extensive administrative capacity and representative bodies in other parts of the world has been done in a different context. It therefore led, more often than not, to limited market expansion, crony capitalism, a high concentration of wealth among the politically well-connected, corruption, and predatory states.

Theory and history indicate the challenge of promoting welfare-enhancing market extension. Future research, however, will have to point to a better way to confront it. It is not merely the protection of property rights that matters. What matters to market expansion, and therefore to economic growth, are the details of the institutions that secure rights—contract-enforcement institutions and coercion-constraining institutions—whose rights they secure, in what products, and how these institutions are mapped into political institutions and policy.

Acknowledgment In writing this chapter I benefitted from discussions and comments from Yarone Greif, Saumitra Jha, Kivanc Karaman, Timur Kuran, John McMillan, and Yingyi Qian. Participants in seminars at FASID, Tokyo, CIAR, Canada, and Stanford Institute for International Studies provided helpful comments. The chapter has been greatly improved, however, by the detailed comments on earlier drafts provided by the editors, Claude Ménard and Mary M. Shirley, an anonymous referee, and, as usual, Joel Mokyr.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D. (2003). Why not a political Coase theorem? Social conflict, commitment, and politics. *Journal of Comparative Economics*, 31(4), 620–652.
- Acemoglu, D., & Robinson, J. A. (2000). Political losers as a barrier to economic development. *American Economic Review*, 90(2), 126–130.
- Acemoglu, D., Johnson, S., & Robinson, J. (2002). *The rise of Europe: Atlantic trade, institutional change and economic growth*. Working paper, MIT.
- Annen, K. (2003). Social capital, inclusive networks, and economic performance. *Journal of Economic Behavior and Organization*, 50(4), 449–463.
- Aoki, M. (2001). *Toward a comparative institutional analysis*. MIT Press.
- Arruñada, B. (2003). Property enforcement as organized consent. *Journal of Law, Economics, and Organization*, 19(2), 401–444.
- Banner, S. (1998). The origin of the New York stock exchange. *The Journal of Legal Studies*, 27, 113.
- Barzel, Y. (2002). *A theory of the state*. Cambridge University Press.
- Bates, R. H. (2001). *Prosperity and violence: The political economy of development*. Norton.
- Bates, R., Greif, A., & Singh, S. (2002). Organizing Violence. *The Journal of Conflict Resolution*, 46(5), 599–628.
- Bellah, R. N., Madsen, R., Sullivan, W. M., Swidler, A., & Tipton, S. M. (1985). *Habits of the heart: Individualism and commitment in American life*. University of California Press.
- Berman, H. J. (1983). *Law and revolution*. Harvard University Press.
- Bernstein, L. (1992). Opting out of the legal system: Extralegal contractual relations in the diamond industry. *The Journal of Legal Studies*, 21, 115–157.
- Bernstein, L. (1996). Merchant law in a merchant court: Rethinking the Code's secret search of immanent business norms. *University of Pennsylvania Law Review*, 144, 1765.
- Biggs, T., Raturi, M., & Srivastava, P. (2002). Ethnic networks and access to credit: Evidence from the manufacturing sector in Kenya. *Journal of Economic Behavior and Organization*, 49(4), 473–486.
- Boot, A. W., Greenbaum, S. I., & Thakor, A. V. (1993). Reputation and discretion in financial contracting. *American Economic Review*, 83(5), 1165–1183.
- Carlos, M. A. (1992). Principal-agent problems in early chartered companies: A tale of two firms. *American Economic Review*, 82(2), 140–145.
- Carlos, M. A., & Nicholas, S. (1996). Theory and history: Seventeenth century joint stock chartered trading companies. *Journal of Economic History*, 56, 916–924.
- Clark, G. (1996). The political foundations of modern economic growth: Britain, 1540–1800. *Journal of Interdisciplinary History*, 26(4), 563–588.
- Clay, K. (1997a). Trade without law: Self-enforcing institutions in Mexican California. *The Journal of Law, Economics, & Organization*, 13(1), 202–231.
- Clay, K. (1997b). Trade, institutions, and credit. *Explorations in Economic History*, 34(4), 495–452.
- Conklin, J. (1998). The theory of sovereign debt and Spain under Philip II. *Journal of Political Economy*, 106(3), 483–513.
- De Soto, H. (1989). *The other path*. Harper & Row.
- Dixit, A. (2003a). On modes of economic governance. *Econometrica*, 71(2), 449–482.
- Dixit, A. (2003b). *Two-tier market institutions* (Working paper). Princeton University.
- Dixit, A. (Forthcoming). *Lawlessness and economics: Alternative modes of governance*. Princeton University Press.
- Downing, B. M. (1992). *The military revolution and political change*. Princeton University Press.
- Ellickson, R. C. (1991). *Order without law*. Harvard University Press.
- Ellison, G. (1994). Cooperation in the prisoner's dilemma with anonymous random matching. *Review of Economic Studies*, 61(3), 567–588.
- Ensminger, J. (1997). Changing property rights: Reconciling formal and informal rights to land in Africa. In J. N. Drobak & J. V. C. Nye (Eds.), *The Frontiers of the new institutional economics*. Academic.

- Epstein, S. A. (1996). *Genoa & the Genoese, 958–1528*. The University of North Carolina Press.
- Fafchamps, M. (2004). *Market institutions in sub-Saharan Africa: Theory and evidence*. MIT Press.
- Fafchamps, M., Biggs, T., Conning, J., & Srivastava, P. (1993). Enterprise finance in Kenya. In *World bank, Africa region, regional program on enterprise development*. World Bank.
- Fearon, J. D. (1997). *Bargaining over objects that influence future bargaining power* (Working paper). University of Chicago.
- Fehr, E., & Fischacher, U. (2004). *Third party punishment and social norms* (Working paper 106). Institute for Empirical Research in Economics, The University of Zurich.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public good experiments. *American Economic Review*, 90, 980–994.
- Furubotn, E. G., & Richter, R. (1997). *Institutions and economic theory*. The University of Michigan Press.
- Galor, O. & Moav, O. (2003). *Das Human Kapital* (Working paper 2000–17). Brown University.
- Gambetta, D. (1993). *The Sicilian mafia: The business of private protection*. Harvard University Press.
- Ghatak, M., & Guinnane, T. W. (1999). The economics of lending with joint liability: Theory and practice. *Journal of Development Economics*, 60(1), 195–228.
- Ghosh, P., & Ray, D. (1996). Cooperation in community interaction without information flows. *The Review of Economic Studies*, 63(3), 491–519.
- Glaeser, L. E., & Shleifer, A. (2002). Legal origi. *Quarterly Journal of Economics*, 117, 1193–1230.
- Glaeser, L. E., & Shleifer, A. (2003). The rise of the regulatory state. *Journal of Economic Literature*, 41(2), 401–425.
- Goetzmann, W., & Köll, E. (2003). *The history of corporate ownership in China*. Yale University.
- Goitein, S. D. (1955). The Cairo Geniza as a source for the history of the Muslim civilization. *Studia Islamica*, 3, 75–91.
- Goitein, S. D. (1971). *The community. A Mediterranean society* (Vol. 2). University of California Press.
- Goldstein, M., & Udry, C. (2002). *Gender, land rights, and agriculture in Ghana*. Yale University.
- Gonzalez de Lara, Y. (2002). *Institutions for contract enforcement and risk-sharing: From debt to equity in late medieval Venice* (Vol. 6, p. 257). Memo, Ente Einaudi, Bank of Italy.
- Goody, J. (1983). *The development of the family and marriage in Europe*. Cambridge University Press.
- Granovetter, M. S. (1985). Economic action, social structure, and embeddedness. *American Journal of Sociology*, 91(3), 481–510.
- Greif, A. (1989). Reputation and coalitions in medieval trade: Evidence on the Maghribi traders. *Journal of Economic History*, XLIX, 857–882.
- Greif, A. (1992). Institutions and international trade: Lessons from the commercial revolution. *American Economic Review*, 82(2), 128–133.
- Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *American Economic Review*, 83(3), 525–548.
- Greif, A. (1994a). Cultural beliefs and the Organization of Society: A historical and theoretical reflection on collectivist and individualist societies. *The Journal of Political Economy*, 102(5), 912–950.
- Greif, A. (1994b). On the political foundations of the late medieval commercial revolution: Genoa during the twelfth and thirteenth centuries. *The Journal of Economic History*, 54(4), 271–287.
- Greif, A. (1995). Political organizations, social structure, and institutional success: Reflections from Genoa and Venice during the commercial revolution. *The Journal of Institutional and Theoretical Economics*, 151(4), 734–740.
- Greif, A. (1996). On the inter-relations and economic implications of economic, social, political, and normative factors: Reflections from two late medieval societies. In J. N. Drobak & J. Nye (Eds.), *Frontiers of the new institutional economics*. Academic.
- Greif, A. (1997). Contracting, enforcement, and efficiency: Economics beyond the law. In M. Bruno & B. Pleskovic (Eds.), *Annual World Bank conference on development economics*. The World Bank.

- Greif, A. (1998). Self-enforcing political systems and economic growth: Late medieval Genoa. In co-authored by Robert H. Bates, A. G., Levi, M., Rosenthal, J.-L., & Weingast, B. R. (Eds.), *Analytic narratives*. Princeton University Press.
- Greif, A. (2000). The fundamental problem of exchange: A research agenda in historical institutional analysis. *European Review of Economic History*, 4(3), 251–284.
- Greif, A. (2001). *The Islamic equilibrium: Legitimacy and political, social, and economic outcomes* (Working paper). Stanford University.
- Greif, A. (2002). Institutions and impersonal exchange: From communal to individual responsibility. *The Journal of Institutional and Theoretical Economics*, 158(1), 168.
- Greif, A. (2004). Impersonal exchange without impartial law: The community responsibility system. *Chicago Journal of International Law*, 5, 109–138.
- Greif, A. (Forthcoming). *Institutions: Theory and history*. Cambridge University Press.
- Greif, A., & Kandel, E. (1995). Contract enforcement institutions: Historical perspective and current status in Russia. In E. P. Lazear (Ed.), *Economic transition in Eastern Europe and Russia: Realities of reform*. Hoover Institution Press.
- Greif, A., Milgrom, P. R., & Weingast, B. R. (1994). Coordination, commitment and enforcement: The case of the merchant guild. *The Journal of Political Economy*, 102(4), 745–776.
- Grossman, H. I. (1997). *Make us a king: Anarchy, predation, and the state* (Working paper w6289, NBER). Cambridge, MA.
- Grossman, H. I., & Kim, M. (1995). Swords or ploughshares? A theory of the security claims to property. *Journal of Political Economy*, 103, 1275–1288.
- Guinnane, T. W. (1994). A failed institutional transplant – Raiffeisens credit cooperatives in Ireland, 1894–1914. *Explorations in Economic History*, 31(1), 38–61.
- Gurevich, A. (1995). *The origins of European individualism*. Blackwell.
- Haber, S., Maurer, N., & Razo, A. (2003). When the law does not matter: The rise and decline of the Mexican oil industry. *Journal of Economic History*, 63(1), 1–32.
- Hamilton, G. G. (1991). The organizational foundations of Western and Chinese commerce: A historical and comparative analysis. In G. G. Hamilton (Ed.), *Business networks and economic development in east and Southeast Asia*. Centre for Asian Studies.
- Hamilton, G. G. (1994). Civilizations and the Organization of Economies. In N. Smelser & R. Swedberg (Eds.), *The handbook of economic sociology*. Princeton University Press/Russell Sage Foundation.
- Hardin, R. (1989). Why a constitution. In B. Grofman & D. Wittman (Eds.), *The federalist papers and the new institutionalism*. Agathon Press.
- Harris, R. (2004). Government and the economy, 1688–1850. In R. Floud & P. Johnson (Eds.), *Cambridge economic history of Britain*. Cambridge University Press.
- Hart, O., & Moore, J. (1999). Foundations of incomplete contracts. *Review of Economic Studies*, 66, 115–138.
- Head, T. F., & Landes, R. (Eds.). (1992). *The peace of god: Social violence and religious response in France around the year 1000*. Cornell University Press.
- Henley, D. (2004). Conflict, justice, and the stranger-king: Indigenous roots of colonial rule in Indonesia and elsewhere. *Modern Asian Studies*, 38(1), 85–144.
- Herrmann-Pillath, C. (1999). On the importance of studying late Qing economic and social history for the analysis of contemporary China, or: Protecting sinology against social science (Duisburg Working Papers in East Asian Studies, Vol. 3).
- Hodgson, M. G. S. (1974). *The venture of Islam* (Vol. 2). Chicago University Press.
- Hodgson, G. M. (1998). The approach of institutional economics. *Journal of Economic Literature*, 36(1), 166–192.
- Hoffman, P. T., Postel-Vinay, G., & Rosenthal, J.-L. (2000). *Priceless markets: The political economy of credit in Paris, 1660–1870*. University of Chicago Press.
- Hogue, A. R. (1996). *Origins of the common law*. Liberty Press.
- Hsu, F. L. K. (1983). *Rugged individualism reconsidered*. University of Tennessee Press.
- Hughes, D. O. (1974). Toward historical ethnography: Notarial records and family history in the middle ages. *Historical Methods Newsletter*, 7(2), 61–71.

- Ingram, P. (1996). Organizational forms as a solution to the problem of credible commitment: The evolution of naming strategies among U.S. hotel chains, 1896–1980. *Strategic Management Journal*, 17, 85–98.
- Irwin, D. A. (1988). Welfare effects of British free trade: Debate and evidence from the 1840s. *Journal of Political Economy*, 96, 1142–1164.
- Israel, J. I. (1995). *The Dutch Republic. Its rise, greatness, and fall, 1477–1806*. Clarendon Press.
- Johnson, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2000). Tunneling. *American Economic Review*, 90(2), 22–27.
- Johnson, S., McMillan, J., & Woodruff, C. (2002). Courts and relational contracts. *Journal of Law, Economics, and Organization*, 18, 221–277.
- Kali, R. (1999). Endogenous business networks. *Journal of Law, Economics, and Organization*, 15(3), 615.
- Kambayash, R. (2002). *The registration system and the grade wage system, coordination and relative performance evaluation*. Tokyo University.
- Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies*, 59, 63–80.
- Klein, D. B. (Ed.). (1997). *Reputation*. The University of Michigan Press.
- Klerman, D. (2007). Jurisdictional competition and the evolution of the common law. *The University of Chicago Law Review*, 74(4), 1179–1226.
- Klerman, D., & Mahoney, P. (2004). *The value of judicial Independence: Evidence from eighteenth-century England* (Working paper, USC).
- Konrad, K., & Skaperdas, S. (1996). The market for protection and the origin of the state. In *Conference on political violence at Princeton University*, Princeton, NJ.
- Kranton, R. E. (1996a). The formation of cooperative relationships. *Journal of Law, Economics, and Organization*, 12(18), 214–233.
- Kranton, R. E. (1996b). Reciprocal exchange: A self-sustaining system. *The American Economic Review*, 86(4), 830–851.
- Kreps, D. (1990). Corporate culture and economic theory. In J. E. Alt & K. A. Shepsle (Eds.), *Perspectives on positive political economy*. Cambridge University Press.
- Lacher, J. M., & Weinberg, J. A. (1989). Optimal contracts under costly state falsification. *Journal of Political Economy*, 97(6), 1345–1363.
- Landa, J. T. (1994). *Trust, ethnicity, and identity*. University of Michigan Press.
- Lane, F. C. (1973). *Venice: A maritime republic*. Johns Hopkins University Press.
- Lapidus, I. M. (1984). *Muslim cities in the later middle ages*. Cambridge University Press.
- Levinson, D. J. (2003). Collective sanctions. *Stanford Law Review*, 56, 345–428.
- Li, S. (1999). *The benefits and costs of relation-based governance: An explanation of the east Asian miracle and crisis* (Working paper). City University of Hong Kong).
- Lipset, S. M. (1959). Some social requisite of democracy: Economic development and political legitimacy. *American Political Science Review*, 53, 69–105.
- Ma, D. (2004). Why Japan, not China, was the first to develop in East Asia: Lesson from sericulture, 1850–1937. Memo, Foundation for Advanced Studies in International Development. Forthcoming in *Economic Development and Cultural Change*, 52, 369.
- Macaulay, S. (1963). Noncontractual relations in business: A preliminary study. *American Sociological Review*, 23, 55–70.
- Macfarlane, A. (1978). *The origins of English individualism* (Vol. 6, p. 255). Basil Blackwell.
- Machiavelli, N. (1990 [1532]). *Florentine histories* (also known as *history of Florence*). Introduction by Harvey Mansfield (Banfield Jr., L. F. & Mansfield, H. C., Trans.). Princeton University Press.
- MacLeod, W. B., & Malcomson, J. M. (1989). Implicit contracts, incentive compatibility, and involuntary unemployment. *Econometrica*, 57(2), 447–480.
- Maskin, E. S., & Tirole, J. (1999). Unforeseen contingencies and incomplete contracts. *Review of Economic Studies*, 66, 83–114.
- McMillan, J. (2002). *Reinventing the bazaar: A natural history of markets*. W.W. Norton & Company.

- McMillan, J., & Woodruff, C. (2000). Private order under dysfunctional public order. *Michigan Law Review*, 98, 101–138.
- Menger, C. (1963 [1883]). *Problems of economics and sociology* (F. J. Nock, Trans.). University of Illinois Press.
- Milgrom, P. R., North, D. C., & Weingast, B. R. (1990). The role of institutions in the revival of trade: The law merchant, private judges, and the champagne fairs. *Economics and Politics*, 2, 1–23.
- Moriguchi, C. (2003). Implicit contracts, the great depression, and institutional change: A comparative analysis of U.S. and Japanese employment relations, 1920–1940. *Journal of Economic History*, 63, 1–41.
- Morris, C. (1972). *The discovery of the individual 1050–1200*. S.P.C.K. for the Church Historical Society.
- Moselle, B., & Polak, B. (2001). A model of the predatory state. *Journal of Law, Economics, and Organization*, 17(1), 1–33.
- Muldrew, C. (1998). *The economy of obligation*. Macmillan.
- Munro, J. H. (1990). The international law merchant and the evolution of negotiable credit in late-medieval England and the low countries. In D. Puncuh (Ed.), *Banchi Pubblici, Banchi Privati e Monti di Pietà nell'Europa Preindustriale*. Società Ligure de Storia Patria.
- Munz, P. (1969). *Frederick Barbarossa*. Cornell University Press.
- Muthoo, A. (2000). *On the foundations of property rights, part I: A model of the state-of-nature with two players*. In typescript, Department of Economics, Essex, UK.
- Nader, H. (1990). *Liberty in absolutist Spain. The Habsburg Sale of towns, 1516–1700*. Johns Hopkins University Press.
- Neal, L., & Quinn, S. (2001). Networks of information, markets, and institutions in the rise of London as a Financial Center in the Seventeenth Century. *Financial History Review*, 7, 117–140.
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- North, D. C., & Thomas, R. P. (1973). *The rise of the Western world*. Cambridge University Press.
- North, D. C., & Weingast, B. R. (1989). Constitutions and commitment: Evolution of institutions governing public choice. *Journal of Economic History*, XLIX, 803–832.
- Norwich, J. J. (1989 [1977]). *A history of Venice*. Vintage Books. (1989 ed. New York: Random House).
- O'Brien, P. K. (2001). *Fiscal exceptionalism: Great Britain and its European rivals – From civil war to triumph at Trafalgar and Waterloo* (Working paper 65/01). Department of Economic History, London School of Economics.
- Okazaki, T. (2002). *The role of the merchant Coalition in pre-modern Japanese economic development: An historical institutional analysis*. Faculty of Economics, University of Tokyo.
- Olds, K. B., & Liu, R. H. (2000). Economic cooperation in 19th-century Taiwan: Religion and informal enforcement. *Journal of Institutional and Theoretical Economics*, 156(2), 404–430.
- Olson, M. (1993). Dictatorship, democracy, and development. *American Political Science Review*, 87(3), 567–576.
- Pettengill, J. S. (1970). The impact of military technology on European income distribution. *Journal of Interdisciplinary History*, X(2), 201–225.
- Polinsky, A. M., & Shavell, S. (2000). The economic theory of public enforcement of law. *Journal of Economic Literature*, 38(1), 45–76.
- Pomeranz, K. (2000). *The great divergence, China, Europe and the making of the modern world economy*. Princeton University Press.
- Pomeranz, K. (2002). *Is there an east Asian development path? Long-term comparisons, constraints, and continuities*. AEA Annual Meeting.
- Portes, A. (1994). The informal sector and its paradoxes. In A. Portes (Ed.), *The economic sociology of immigration*. Russell Sage Foundation.
- Przeworski, A., Alvarez, M. E., Cheibub, J. A., & Limongi, F. (2000). *Democracy and development*. Cambridge University Press.

- Quinn, S. (1997). Goldsmith-banking: Mutual acceptance and Interbanker clearing in restoration London. *Explorations in Economic History*, 34, 411–432.
- Quinn, S. (2001). The glorious Revolution's effect on British private finance: A microhistory, 1680–1705. *Journal of Economic History*, 61(3), 593–615.
- Rajan, R. G., & Luigi, Z. (2003). *Saving capitalism from the capitalists*. Random House.
- Ramseyer, J. M. (1991). Legal rules in repeated deals: Banking in the shadow of defection in Japan. *The Journal of Legal Studies*, 10, 91–117.
- Rauch, J. E. (2001). Business and social networks in international trade. *Journal of Economic Literature*, 39, 1177–1203.
- Redding, S. G. (1991). Weak organization and strong linkages: Managerial ideology and family business networks. In G. Hamilton (Ed.), *Business networks and economic development in east and Southeast Asia*. Center of Asian Studies. University of Hong Kong.
- Reynolds, C. H., & Norman, R. V. (Eds.). (1988). *Community in America*. University of California Press.
- Richardson, G. (2002). *Christianity and craft guilds in late-medieval England* (Working paper 211). University of California at Irvine.
- Root, H. L. (1989). Tying the King's hands: Credible commitments and royal fiscal policy during the old regime. *Rationality and Society*, 1(2), 240–158.
- Rose-Ackerman, S. (1999). *Corruption and government*. Cambridge University Press.
- Ryser, M. (1997). Sanctions without law: The Japanese financial clearinghouse guillotine and its impact on default rate. In D. B. Klein (Ed.), *Reputation*. The University of Michigan Press.
- Schaefer, U. (1989). Forwards and futures in Tokugawa-period Japan: A new perspective on the Dojima rice market. *Journal of Banking and Finance*, 13(4–5), 487–513.
- Shiue, C. H., & Keller, W. (2003). *Markets in China and Europe on the eve of the industrial revolution*. University of Texas.
- Shleifer, A., & Vishny, R. W. (1993). Corruption. *The Quarterly Journal of Economics*, 108(3), 599–617.
- Skaperdas, S. (1992). Cooperation, conflict, and power in the absence of property rights. *American Economic Review*, 82(4), 720–738.
- Skaperdas, S., & Syropoulos, C. (1996). *Insecure property rights and the stability of exchange* (Technical report). Institute for Mathematical Behavioral Science, Pennsylvania State University.
- Stringham, E. (2003). The extralegal development of securities trading in seventeenth century Amsterdam. *The Quarterly Review of Economics and Finance*, 43, 321–344.
- Stulz, R., & Williamson, R. (2003). Culture, openness, and finance. *Journal of Financial Economics*, 70(3), 313–349.
- Sussman, N., & Yefeh, Y. (2000). Institutions, reforms, and country risk: Lessons from Japanese government debt in the Meiji period. *Journal of Economic History*, 60(2), 442–467.
- Sussman, N., & Yefeh, Y. (2004). *Constitutions and commitment: Evidence on the relation between institutions and the cost of capital*. The Hebrew University.
- Tadelis, S. (2002). The market for reputations as an incentive mechanism. *Journal of Political Economy*, 110(4), 854–882.
- Tilly, C. (1990). *Coercion, capital, and European states, AD 990–1992*. Blackwell.
- Tirole, J. (2001). Corporate governance. *Econometrica*, 69(1), 1–35.
- Townsend, R. M. (1979). Optimal contracts and competitive markets with costly state verification. *Journal of Economic Theory*, 21(2), 265–293.
- Triandis, H. C. (1990). Cross-cultural studies of individualism and collectivism. In J. Berman (Ed.), *Nebraska symposium on motivation, 1989*. University of Nebraska Press.
- Udovitch, A. L. (1988). Merchants and Amirs: Government and trade in eleventh-century Egypt. *Asian and African Studies*, 22, 53–72.
- Volckart, O. (2001). *The economics of feuding in late medieval Germany*. (Working paper). Institut für Wirtschaftsgeschichte, Berlin.
- Watson, J. (2002). Starting small and commitment. *Games and Economic Behavior*, 38, 176–199.

- Weingast, B. R. (1996). Political institutions: Rational choice perspectives. In R. Goodin & H.-D. Klingemann (Eds.), *A new handbook of political science*. Oxford University Press.
- Weingast, B. R. (1997). The political foundations of democracy and the rule of law. *American Political Science Review*, 91(2), 245–263.
- Williamson, O. E. (1985). *The economic institutions of capitalism*. The Free Press.
- Williamson, D. V. (2002). *Transparency and contract selection: Evidence from the financing of trade in venetian Crete: 1303–1351*. US Department of Justice.
- Wong, R. B. (1997). *China transformed. Historical change and the limits of the European experience*. Cornell University Press.
- Woodruff, C. (1998). Contract enforcement and trader liberalization in Mexico's footwear industry. *World Development*, 26(6), 979–991.
- Yang Li, M. (2002). *Essays on public finance and economic development in a historical institutional perspective*. Ph.D. dissertation, Stanford University.
- Yarbrough, B. V., & Yarbrough, R. M. (2003). The contractual role of boundaries: Law and economics meets international organization. *European Journal of International Relations*, 9(4), 543–590.
- Zelin, M., Ocko, J. K., & Gardella, R. (2004). *Contract and property in early modern China*. Stanford University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part VII

Culture and Informal Institutions

Institutions Before States: Aggressive Egalitarianism and Private Property

33

Samuel Bowles and Jung-Kyoo Choi

Contents

33.1	Introduction	872
33.2	Institutions Before States as Conventions	874
33.3	Aggressive Egalitarianism as a Convention	877
33.3.1	Food Sharing Conventions Before States	877
33.3.2	Modelling Aggressive Egalitarianism: The Civic-Grabber-Sharer Game	880
33.4	Institutional Persistence and Change: Private Property Before States	884
33.4.1	Possession-Based Individual Property Before States	884
33.4.2	Modeling the Origin of Private Property: The Civic-Bourgeois-Sharer Game	886
33.5	Egalitarian Wealth Distributions Prior to States	889
33.6	Discussion	891
	References	891

Abstract

Institutions are sometimes represented as a state-imposed set of rules of the game regulating social interactions in a population. But the cooperative and egalitarian practices and other enduring social structures observed among foragers in the ethnographic record almost certainly governed social interactions since the appearance of biologically modern humans hundreds of thousands of years ago. The subsequent emergence of private property as a fundamental institution pre-dates states by many millennia. In the absence of state rule, institutions such as sharing the necessities of life beyond the family and respecting the property rights

Thanks to Max Greenberg, Caroline Seigel, Elisabeth Wood, the editors, an anonymous referee, and the Santa Fe Institute, for their contributions to this research.

S. Bowles (✉)

Santa Fe Institute and CORE Econ, Santa Fe, NM, USA

J.-K. Choi

Kyungpook National University, Daegu, South Korea

of others resemble linguistic conventions, conformity with which is a mutual best response conditional on most others conforming. Recent archaeological evidence along with models of sharing and private property as conventions suggest that what the first states accomplished was not to have put an end to Hobbes' chaotic state of nature and its "warre of every man against every man," but instead, more in line with Engels than Hobbes, to have expanded the scope of coordinated social interactions and to have inaugurated unprecedented and enduring levels of inequality of wealth.

Keywords

Private property · Civil society · Archaic proto-state · Evolutionary game theory · Convention · Thomas Hobbes · Wealth inequality · Friedrich Engels

33.1 Introduction

Thomas Hobbes envisioned an epoch without a coercive state as "the time when men live without a common Power to keep them all in awe," resulting in "this warre of every man against every man" in which "there is . . . no society, . . . nothing can be unjust" and consequently the "life of man [is] solitary, poor, nasty, brutish, and short" (Hobbes, 2005) pp. 101–103). His characterization can be defended as a philosopher's thought experiment, not as social science or history. But to Hobbes this was not just an illuminating hypothetical; it was a foundation of his critique of classical and Renaissance political theory and its advocacy of free states and the agency of civic-minded autonomous citizens as a foundation of good governance (Skinner, 2002).

Hobbes' idea, in modern language, is that, in the absence of a coercive state, members of a population would be incapable of ensuring public order and providing other public goods. The argument still resonates despite contrary evidence from across the social sciences (Boehm, 2000; Graeber & Wengrow, 2021; Ostrom, 1990). An example is the enduring interest in the concept of the tragedy of the commons offered by Garrett Hardin, who wrote: "If ruin is to be avoided in a crowded world, people must be responsive to a coercive force . . . a 'Leviathan'" (Hardin, 1978, p. 314) which would take the form of "mutual coercion, mutually agreed upon" (Hardin, 1968, p. 1247).

A consequence of the ongoing hold of this Hobbesian framework has been insufficient attention to forms of successful cooperation and collective action not directly involving states. Exemplary of this lacuna is how little we know about the governance of human societies in the two or three hundred thousand years of the prehistory of biologically modern humans prior to the emergence of the first proto-states perhaps 6000 years ago. Hobbes' vision of the state of nature has also (at least until recently) fostered a societal equivalent of the *tabula rasa* assumption among economists, facilitating an account of the evolution of institutions consistent with methodological individualism. Here is a passage from a pioneering work on the economic theory of institutions:

If economics . . . is going to study the rise and evolution of social institutions, a very simple methodological approach is suggested. We should start our analysis in a Lockean state of nature in which there are no social institutions at all, only agents, their preferences, and the technology they have at their disposal. . . . The next step would be to study when, during the evolution of this economy, such institutions as money, banks, property rights, competitive markets, insurance contracts, and the state would evolve. (Schotter, 1981, p. 20)

But such an institutional *tabula rasa* never existed. Hobbes thought the same: In the passage above describing his state of nature, he wrote: “I believe it was never generally so” (Hobbes, 2005, p. 101).

Early humans—“the hunting ape”—surely had rules regulating how the meat of large, infrequently killed prey would be shared, as do contemporary hunter-gatherers that rely on the occasional acquisition of large nutritional packages (Kaplan & Hill, 1985; Stanford, 1999). Possession-based property rights—meaning a secure claim on an asset based on individual possession—are observed not only among humans living in Neolithic societies well before states (Bogaard et al., 2009), but also in other animals—spiders, long-tailed macaques and hamadryas baboons, for example (Hammerstein & Reichert, 1988; Kummer, 1991; Kummer & Cords, 1991; Sigg & Falett, 1985).

We will use these two cases—the sharing of the necessities of life among group members beyond the family and the emergence of private property—to illustrate the not-very-Hobbesian workings of institutions before states, including the processes by which institutional evolution may have occurred. Because we do not have written records prior to the emergence of proto-states six millennia ago, most of this evidence is archaeological and subject to substantial uncertainties, which we will seek to mitigate using data from a set of recent small-scale societies in the ethnographic record that may be similar in the relevant respects to pre-state communities.¹

In the next section we will introduce some terms—institutions and states (including a variant that we term the archaic proto-states (Bowles & Fochesato, 2024))—as well as key dates—particularly the emergence of the first archaic proto-state. Using these definitions we also provide evidence that institutions structured social interactions before states, permitting the coordination of human activity well beyond the family, resulting in improved societal well-being. In Sect. 33.3, we describe a set of egalitarian institutions in Neolithic western Eurasia—such as the redistribution of food and limitations on the intergenerational transmission of wealth—and propose a model of how, despite the presence of self-aggrandizing individuals, these institutions appear to have persisted over the very long run. Section 33.4 presents evidence on early forms of private property and proposes an explanation of how the domain of this institution was enlarged to include stored harvested crops, domesticated animals, dwellings, and eventually land following the introduction of farming. These developments occurred well before the concentration of political power in proto-states and

¹ The archaeologist Robert Kelly rightly cautions against “viewing living foragers as our Paleolithic ancestors” or as “relic populations” (Kelly (1995) p. 334). See also Wobst (1978). We address these concerns in Sect. 33.3 below, where we identify practices—particularly sharing meat and other food beyond the family—among recent foragers that, on ecological and other grounds, appear likely to have been similarly widespread in prehistory.

apparently without substantial collateral increases in the degree of economic inequality. In Sect. 33.5, we show that, beginning around 5000 years ago, the rise of the archaic proto-state and later proper states and slavery inaugurated modern levels of wealth inequality.

The striking characteristic of societies prior to states, then, is not disorder but instead the rare and ephemeral nature of political and economic disparities, which seems likely to be a result of what the archaeologist Ian Hodder calls aggressive egalitarianism (Hodder, 2014). The novelty introduced by the first states was not to have curbed chaos arising from the pursuit of self-interest but to have vastly expanded the scope of social interactions and to have introduced enduring modern levels of wealth inequality. This is less in line with Hobbes' view of the state and more with Friedrich Engels', who saw the early concentration of political power as the emergence of a "state of the most powerful, economically ruling class, which by its means becomes also the politically ruling class, and so acquires new means of holding down and exploiting the oppressed class" (Engels, 2010 [1884], Chap. 9).²

33.2 Institutions Before States as Conventions

By institutions we mean the set of laws, informal rules, or conventions that give a durable structure to social interactions among members of a population.³ Conformity to the behaviors prescribed by institutions may be secured by a combination of centrally deployed coercion (laws), social sanction (informal rules), and mutual expectations (conventions) that make conformity a best response for virtually all members of the population conditional on most others also conforming. Institutions before states are marked by the absence of laws enforced by a centralized body.

Thus we follow Max Weber by defining a state as an entity that can "monopolize the use of force," that is, a "system of order [that] claims binding authority . . . over all action taking place in the area of its jurisdiction" in which "the use of force is regarded as legitimate only so far as it is either permitted by the state or prescribed by it" (Weber, 1978 [1921], p. 56).⁴ Many archaeologists have adopted a variant of

²The passage continues: "The ancient state was, above all, the state of the slave-owners for holding down the slaves, just as the feudal state was the organ of the nobility for holding down the peasant serfs and bondsmen, and the modern representative state is the instrument for exploiting wage-labor by capital."

³Douglass North offers this definition: "Institutions are a set of rules, compliance procedures and moral and ethical behavioral norms designed to constrain the behavior of individuals in the interest of maximizing wealth or utility of principals." North (1981, p. 201). If read carefully, and taking account of the fact that the principals that North refers to in the private economy would be lenders in credit markets and employers in labor markets, North's definition of the state is remarkably close to Engels. We concur with North in taking widely accepted social norms to be institutions but differ in not requiring that institutions be "designed" to maximize the wealth of principals or for any other purpose. We use the term organization rather than institution to refer to entities with proper names, like the Bank of England, or the Ford Motor Company.

⁴Weber (1978 [1921]) pointed out that, defined in this way "the state has only in modern times reached its full development" (p. 56).

Weber's definition of the state (Kirch, 2010; Service, 1975). Bruce Trigger, for example, defines a state as "a politically organized society that is regarded by those who live in it as sovereign . . . and has leaders who control its social political, legal, economics, and cultural activities" and adds that "central governments possessed ultimate control over justice and the use of force" (Trigger, 2003, pp. 47, 92).

States in this Weberian usage are thus distinct from other forms of political leadership termed "chiefdoms" and "big man" systems, in which leaders coordinate the activities of a group, constituting (in the words of the anthropologist Elman Service) "a permanent agency of coordination," but lack coercive powers (Service, 1962). Referring to the literature in archaeology, Norman Yoffee makes this distinction: "In political terms chiefdoms contain hereditary . . . leaders (sometimes called a nobility) and centralized direction . . . but they have no formal machinery for forceful repression" (Yoffee, 2005, p. 23). The archaeologists Gary Feinman and Joyce Marcus distinguish archaic states "as having more power than the rank societies . . . particularly in the areas of waging war, exacting tribute, controlling information, drafting soldiers and regulating manpower and labor" (Feinman & Marcus, 1998, p. 15).

Many of the polities that are central to our interpretation appear to have been neither chiefdoms nor states. What is striking about what some have termed the "first states" or "archaic states" in Mesopotamia is their apparent lack of a monopoly on the legitimate use of force.⁵ Kin-based lineages and the exercise of what Weber called patrimonial authority "not enforced by physical coercion" were essential to these systems of rule, constituting the "patrimonial state," in which "the political realm as a whole is approximately identical with a huge princely manor" (Weber, 1978 [1921], p. 1013). To capture this aspect of early political centralization, we use the term "archaic proto-state" to describe a system of elite governance with substantial kin-based or other personalized lines of authority rather than a centralized anonymous bureaucratic system of governance. It appears that the earliest archaic proto-states emerged in Mesopotamia a bit less than 6000 years ago (Bowles & Fochesato, 2024).

The nature and scope of institutions before states is suggested by archaeological findings indicating the coordinated activities of substantial numbers of people, in some cases involving multiple communities. Examples are nonresidential public buildings (Roux et al., 2000), massive constructions ranging from the tower at Jericho to the monumental enclosures at Göbekli Tepe (Celik, 2015; Dietrich et al., 2012; Kuijt & Goring-Morris, 2002) and evidence of long-distance trade in obsidian as early as the African Middle Stone Age (McBrearty & Brooks, 2000).

The inhabitants of Abu Hureyra on the banks of the Euphrates at the end of the Paleolithic (11,700 years ago) and early Neolithic constructed massive traps whose funnel-shaped designs consisted of high stone walls extending over a kilometer in two directions that were capable of capturing and killing large numbers of migrating gazelle (estimates are 50–500) in a single day (Moore et al., 2000, pp. 442–449). Similar constructions dating even earlier are found in the region. We can only

⁵Diakonoff (1969), Gelb (1967), Yoffee (1995), and Stasavage (2020).

speculate about the conventions, social norms, and political structures that were required to coordinate a society-wide activity of this scope (the village may have grown to more than five thousand residents, p. 494). Moore and his coauthors (the leading archaeologists of the site) observe:

The village was so large that authority to take decisions affecting the whole community had to be vested in an individual or small group... on a relatively formal basis, ... acting with community consent ... settling disputes ... and organizing collective activities [including] the gazelle ... hunts. It was the community leaders who probably arranged the distribution of land and who controlled the use of surface water. (Moore et al., 2000, p. 505)

Before the imposition of state rule, aboriginal Australians likewise engaged in communal hunting, involving in some cases a human circle 10–30 km in diameter, as well as fish and duck hunting nets 90 meters long (Lourandos, 1997). Communal infrastructure included evidence of a complex of fish traps covering an estimated total area of 21,600 square meters and a drainage and eel trapping system that is estimated to have required 13,000 h of labor to construct. Consumption of the massive quantity of eels trapped was immediate and involved 1000 or more people.

In California, the Yurok people from ten different downstream villages annually constructed a fish weir entirely blocking the Klamath River at Kepel (a span of 70 m) to capture migrating salmon (Kroeber, 1960, pp. 11–18). The project was directed by an engineer and was associated with 2 months of elaborate ceremony before, during, and after the annual construction and dismantling of the weir. People from upstream villages (who did not participate in the construction of the weir) had the right to harvest salmon from the weir during the period that the salmon migration at their village was interrupted. An account from 1860 describes weir similar to the construction at Kepel, along with associated ceremonies and similar rights of upstream villages.

These and similar projects—all based on coordinated actions of large numbers of people without a centralized monopoly of coercive powers—suggest that the persistence and change in institutions before states might be modeled in a manner similar to languages, that is, as conventions (Bowles et al., 2017; Lewis, 1969; Young, 1995). These are social practices common to a population that endure because conforming to their prescribed behaviors is the best response for virtually all participants, conditional on most others also conforming. Infrequent transitions from one convention to another may take place if a sufficient fraction of the population shifts (either by chance or intentionally) to an alternative convention.

A feature of modeling institutions as conventions is that it provides a unified account of the observed punctuated equilibria pattern of institutional evolution, capturing both long-term persistence and occasional transitions to novel configurations (Eldredge & Gould, 1972). Daron Acemoglu and James Robinson remark: “Approaches positing institutional persistence as a matter of fact, and then thinking of institutional changes as unusual events will not be satisfactory” (Acemoglu & Robinson, 2009). The emergence and spread of private property in the primary means of subsistence during the Neolithic poses this challenge particularly starkly,

because for perhaps hundreds of millennia during the Pleistocene, mobile bands of hunter-gatherers had organized their economies on a quite different set of rules in which food was widely shared when it was acquired.

33.3 Aggressive Egalitarianism as a Convention

Drawing on methods in evolutionary game theory pioneered by Robert Sugden, Peyton Young, Jorgen Weibull, and others, we represent sharing some kinds of food—thought to be widely practiced among both hunter-gatherers and farmers prior to the introduction of animal traction—as a convention (Binmore et al., 2003; Sugden, 1986, 1989; Weibull, 1995; Young, 1998). We explain how food sharing as an institution may have persisted over long periods even when individuals could pursue self-aggrandizing strategies such as refusing to share.

33.3.1 Food Sharing Conventions Before States

Archaeological evidence of food-sharing practices among farmers in Neolithic western Eurasia include communal storage facilities (Willcox & Storedeur, 2012), building features such as the absence of kitchens in some dwellings—suggesting communal processing and possibly consumption of food (Weide et al., 2022)—and the decoration of public spaces in homes with the heads of hunted animals, thought to be associated with sharing meat beyond the family (Bogaard et al., 2009).

Available ethnographic data on food sharing in hunter-gatherer populations indicate that the fraction of acquired food that is consumed by other households is substantial. The fact that particular individuals often were assigned the role of distributing the food acquired by hunting or gathering is not evidence for private property, even though, confusingly, the term “ownership” is sometimes used to describe these roles. John Dowling, for example, writes: “...the slain animal is shared with others. The rights and prerogatives entailed in ownership are primarily those of performing the distribution, not of deciding whether or not the animal will be distributed ... ownership carried no exclusive rights.” (Dowling, 1968, p. 505). Polly Wiessner writes: “Ownership, however, means little more than having the right to distribute the meat or to decide who should do so, for all foragers have rules stipulating that meat must be widely shared” (Wiessner, 1982, p. 177).

Particularly striking are the Aka Pygmies in Central Africa among whom, Bahuchet writes, the “owner” is “forbidden to eat the animal” but instead is simply “in charge of sharing. ... Thus the notion of ownership among the Aka is quite different from ours, theirs being based on each individual’s responsibility in assuring the prosperity of the community.” He concludes that “the underlying concept conveyed” by the term ownership is not “possession” but “responsibility” (Bahuchet, 1990).

Kristin Hawkes and her coauthors report that sharing large meat packages is “like appropriation from the public domain.” And, describing the distribution procedures among the three most studied groups of foragers (the Ache, !Kung, and Hadza) they

write: “in none of them can the hunter or anyone else exclude others from shares of the animal. This is the common pattern among ethnographically known foragers” (Hawkes et al., 2001, p. 132). Thus, an individual hunter may have the right to distribute their prey to others but monopolizing it for their family would typically not be possible because it would be a serious transgression of the norms of co-insurance based on food sharing, a transgression which, if widely accepted, would undermine the survival of the group.

Some of the best evidence on the extent of food sharing concern the Ache, living in eastern Paraguay as foragers subsisting on wild resources at contact (which occurred in 1970). Data are available on the distribution of foods between procurer and consumer among Ache groups foraging in the forest between 1980 and 1982 (Gurven et al., 2002). A detailed study of four foraging trips shows that the mean amount of food given away to members of other nuclear families was very close to (and not statistically different from) the amount that would have resulted in an exactly equal distribution of foods by each family to each family (including the acquiring family). Among Ache living in a farming settlement, about half of the cultivated foods are given to other nuclear families; by contrast three quarters of forest meat and other products are given to other families (Gurven et al., 2001).

Detailed data across a number of populations on flows of food from those who acquired it (by either hunting-gathering or farming) to those outside the nuclear family of the acquirer are consistent with the above accounts. Figure 33.1 gives the fraction of food acquired by one family that is redistributed to other families,

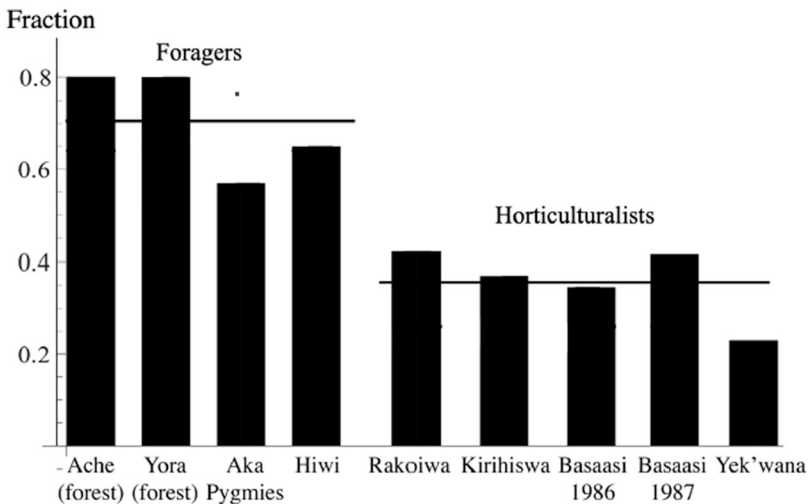


Fig. 33.1 The fraction of food acquired by one family that is redistributed to other families (effective tax rate). (Source. Ache (forest): (Gurven et al., 2002); Yora: (Hill & Kaplan, 1989); Aka: (Bahuchet, 1990, 1991); Hiwi: (Gurven et al., 2000); Rakoiwa, Kirihiswa and Basaasi (1986 and 1987): (Hames, 2000); Ye'kwana: (Hames & McCabe, 2007). If direct reciprocation of food transfers is practiced (as is common), these estimates of retained amounts understate the true retention rate)

analogous to the effective tax rate financing redistribution, though it is implemented by social norms, not by a government. The mean amount retained among foragers is less than half of that among horticulturalists; but even among the latter, “the effective tax rate” is on a par with the rate of income tax paid by a family of four in today’s Nordic countries.

What can we infer from this ethnographic evidence about food sharing in prehistory? James Woodburn, Bruce Knauff, and others have provided distinctions essential to identifying characteristics of hunter-gatherer groups likely to have predated the Holocene, all stressing mobility, the lack of political hierarchy, and pervasive egalitarianism (Knauff, 1991; Woodburn, 1982). The most systematic effort of this type is that of Christopher Boehm (2012). Beginning with Lewis Binford’s comprehensive list of 339 hunter-gatherer societies (Binford, 2001), Boehm selected those which were at least partially mobile and both economically and politically independent during the period studied by ethnographers, and whose livelihood did not depend on modern innovations (thereby eliminating equestrian hunters and those commercially involved in trapping, for example).

After this culling, Boehm’s 154 “Late Pleistocene plausible” societies were nutritionally dependent on hunting and fishing (obtaining 65% of their calories from these activities, rather than from gathering), and moved on average 11 times per year for a total migration distance of 262 km. Some groups were semi-sedentary, like the 22 groups in Boehm’s sample from the NW coast of North America who moved on average just 4.5 times per year. Boehm estimated that 52% of the calories obtained by the members of these 154 “Late Pleistocene plausible” groups came from “large packet” flesh sources (large animals and marine resources). Boehm’s evidence is replicated using detailed caloric measurements for diet in nine foraging groups. Kaplan, Hill, Hurtado, and Lancaster found that 63% of the calories derived from hunting (including honey collection) and fishing (Kaplan et al., 2000).

Boehm concludes: “In the Upper Paleolithic, the overall percentage reliance on large-packet flesh might have been greater because of the proliferation of megafauna. . . . There is general agreement that mobile hunter-gatherers share their large game at the band level whenever it is acquired sporadically and that this likely held for the Upper Paleolithic.”

From the archaeological evidence it appears that Neolithic populations probably engaged in widespread sharing of food. Moreover, it seems likely that sharing was sustained because people actively resisted the concentration of wealth and power, deploying a set of practices that appear designed to thwart self-aggrandizement. These include evidence of deliberate destruction of the dwellings and other assets of the wealthy (Hodder, 2014; Twiss et al., 2008; Verhoeven, 2000; Wright, 2014), secondary burial of skulls in collective graves which appears to have been a way to emphasize group rather than individual identity (Kuijt, 1996), and likely shared ownership of some animals (Hodder, 2014).

Referring to early farmers, Boehm writes: “their egalitarianism, like that of hunters, is the product of constant vigilance and sometimes very harsh sanctioning” (Boehm, 2000, p. 13). Bruce Trigger as well as David Graeber and David Wengrow provide detailed accounts of this “aggressive egalitarianism” among the Iroquois at

the time of contact with Europeans (Graeber & Wengrow, 2021; Trigger, 1990). To stress the deliberate and aggressive enforcement of food distribution norms among the Hadza in Tanzania half a century ago, Nicolas Blurton Jones referred to transfers from the well fed to the hungry as “tolerated theft” (Blurton-Jones, 1984).

33.3.2 Modelling Aggressive Egalitarianism: The Civic-Grabber-Sharer Game

To better understand the economic and social structures that might have supported the observed pattern of food sharing, we now present an evolutionary model of a hypothetical population similar in size and composition to a Late Paleolithic ethnolinguistic unit (600 individuals per generation), made up of many subpopulations of twenty individuals per generation, about the size of forager bands or small villages, among which there is occasional migration.⁶ The model and its parameterization were developed to understand the Early Neolithic transition from hunting and gathering to farming; when simulated along with climate data over the period 40,000 BP to the present, it closely replicates the observed archaeological record.⁷

Here, we use the model for a different purpose: to explore the dynamics of wealth inequality in a Late Paleolithic or Neolithic population in which both aggressive egalitarianism and self-aggrandizement are behavioral options, reflecting patterns of self-interest, sharing, and collective discipline found in ethnographic studies of hunter-gatherers and horticulturalists (Boehm, 1993, 2000; Hayden, 2001; Kaplan & Gurven, 2005; Kaplan & Hill, 1985; Wiessner, 2005). We will show that in simulations of the model, societies sustain the practice of extensive food sharing resulting in egalitarian distributions of wealth over long periods sometimes. However, they occasionally make substantial but short-lived excursions to population distributions in which sharing breaks down, with concomitant increases in inequality.

The simulated agents may either farm or hunt and gather, and when they interact with others in their subpopulation, the strategy that each follows determines how the goods they have acquired are divided up. There are four stages in the model: production, distribution, cultural updating, and intergroup competition. In the production stage, individuals adopt one of two technologies: farming or hunting-gathering, acquiring a product as a result. In the distribution stage, each member

⁶The substantial size of the subpopulations makes it unlikely that widespread sharing and other forms of cooperation could be sustained by self-interested, repeated interactions, as can be readily demonstrated for dyadic games. It is widely but mistakenly thought that “our forebearers lived in small nomadic bands of a few dozen individuals . . . each of our ancestors was, in effect, on a camping trip that lasted an entire lifetime, and this way of life endured for most of the last 10 million years.” Cosmides and Tooby (2010, p. 11) Genetic and ethnographic evidence on the size of groups and a demonstration of the difficulty of sustaining cooperation by folk theorem-type reasoning in repeated game theory in groups of any substantial size are both provided in Bowles and Gintis (2012).

⁷The model and its simulations are described in the text and online supplementary materials of Bowles and Choi (2013) and in Chap. 11 of Bowles (2004).

of the population is paired randomly with another member of their subpopulation. In these paired interactions, individuals may adopt one of three behavioral strategies.

The first behavioral type, called the Sharer, is similar to the Dove in the Hawk-Dove Game, a variant of the Chicken Game which represents sharing and self-aggrandizement. In an interaction, a Sharer concedes half of the product to the other, or the whole product if the other claims it. When two Sharers are paired, they exchange equal shares of their product and as a result have equal payoffs. A second behavioral type, the Grabber (similar to the Hawk), attempts to claim the entire product; if paired with a Sharer, they are successful, but if paired with another Grabber, a costly conflict results, resulting in negative expected payoffs.

The third (novel) behavioral type, the Civic, acts exactly like a Sharer when paired with a fellow Civic or a Sharer (they share), but when paired with an individual who refuses to share (a Grabber), they join with any other Civics in the subpopulation to contest the claims of the Grabber, succeeding with a probability that increases with their numbers. If they succeed, they distribute the product equally among all the Civics, whereas the loser—the Grabber—bears a cost. If, instead, the Civics fail, they all bear a share of the losers' cost and the Grabber gets the product.⁸ The Civics are the aggressive egalitarians.

Because the payoffs of each behavioral type differ according to the type with which they are paired, expected payoffs depend on the distribution of types in each subpopulation. Agents occasionally update their strategies, switching from lower to higher payoff behaviors while occasionally experimenting by choosing a strategy randomly. As a result, the distribution of payoffs in each subpopulation evolves over time as the distribution of behavioral types changes.

The above process calibrated with an empirically plausible set of model parameters is depicted in Fig. 33.2, in which the coordinates of a point in the simplex (which sum to one) give the fraction of the three behavioral types in the population. Thus, points closer to a vertex indicate a larger share of the type indicated at the vertex, and each vertex represents a homogeneous population of just a single type. The vector field given by the arrows represents the direction of change of the population distribution induced by the payoff differences in each region of the figure.

The final stage of the model is between-group competition, in which the subpopulations are randomly paired and the one with the higher average payoffs wins the inter-group contest, gaining some of the loser's wealth. Members of the losing subpopulation then differentially copy the strategies of the higher-payoff winning subpopulation. These four stages are then repeated over many periods.

An equilibrium in this model is characterized by a stationary distribution of behavioral types. One of these is similar to the Hawk-Dove equilibrium, in which there are no Civics. Grabbers gain high incomes by exploiting Sharers; and they

⁸The Civics' cooperative punishment strategy is based on ethnographic studies of the coordinated maintenance of social order in stateless societies (for example, Mahdi (1986) and Wiessner (2005)). Coordinated punishment contrasts with the noncooperative (individual) punishment strategies in the conventional public goods with punishment game and has been shown to be evolutionary viable under plausible parameters in Boyd et al. (2010).

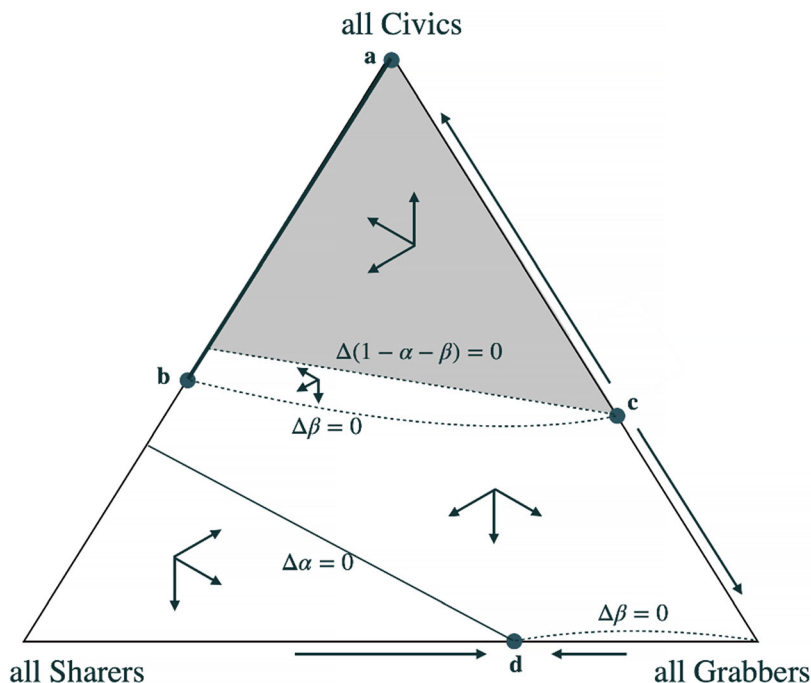


Fig. 33.2 Within sub-population dynamics for the three behavioral types in the Civic-Grabber-Sharer Game. The solid line on the left edge is the set of neutrally stable states without Grabbers in the population. The line labeled $\Delta\alpha = 0$ gives the population distributions for which the fraction of Sharers in the population is unchanging, and analogously for $\Delta\beta = 0$ (fraction of grabbers unchanging) and for $\Delta(1-\alpha-\beta) = 0$ (fraction of Civics unchanging). Point c is an unstable stationary state composed of just Civics and grabbers. The shaded area is the basin of attraction of the “no Grabbers” set of equilibria

persist because the occasional Civic that appears due to migration or behavioral experimentation will have to engage in frequent contests with aggrandizing Grabbers without the assistance of other Civics, therefore losing many conflicts and paying the loser’s cost. The resulting lower payoffs of the Civics will prevent them from proliferating in the subpopulation. Because Grabbers also interact with each other, bearing the negative expected payoff as a result of the conflict, average payoffs in the Grabber-Sharer equilibrium are low, and payoff differences between Grabbers and Sharers are substantial.

The all-Civic population is also an equilibrium, from which an occasional greatly outnumbered Grabber (introduced by migration or experimentation) will readily be eliminated by the Civics, united against them. Because within-subpopulation conflicts are as a result rare, average payoffs are close to the maximum possible. And, because Civics and Sharers are behaviorally identical and have the same payoffs if Grabbers are absent, payoff inequality is limited. As a result, a subpopulation without Grabbers or a mixed Civic-Sharer subpopulation will tend to prevail in

contests with mixed Sharer-Grabber subpopulations due to their superior average payoffs.

But the all-Civic equilibrium is evolutionarily vulnerable because it is only neutrally (not asymptotically) stable: small displacements from the all-Civic population distribution (due to immigration of Sharers or Civics randomly switching to Sharer) will not be amplified by the within-subpopulation selection process, but they will not be dampened either. So, the all-Civic equilibrium is subject to drift (along the solid line on the left edge of the simplex in the figure), which occasionally will result in a large fraction of the subpopulation becoming Sharers. This mostly Sharer subpopulation is then readily invaded by Grabbers, moving the subpopulation to the Sharer-Grabber equilibrium at point **d**. In the neighborhood of this stationary state, Grabbers seize the goods of Sharers and bear the costs of conflicts with each other, not exactly a “warre of every man against every man” (when Sharers meet, they peacefully share) but close enough for us to call **d** the Hobbesian equilibrium.

The mathematics of the resulting dynamical system, including inter-group competition, do not yield any informative analytical results (the interactions are too complex). But we can simulate a population like the one depicted in the model, allowing us to trace the persistence and occasional breakdown of cooperative sharing by the degree of inequality in income in each sub-population. Figure 33.3 shows the Gini coefficient for payoffs among the 20 members (of a single generation) of one of the subpopulations over the period from 15,000 to 6000 years ago, prior to the emergence of archaic proto-states (they were predominantly hunter-gatherers, with only 18% of members engaging in farming over the course of this simulation).

While the mean and most common level of inequality is quite limited, the subpopulation does occasionally experience breakdowns of cooperation and substantial levels of inequality (a Gini coefficient greater than 0.5 occurs three times in the run presented).

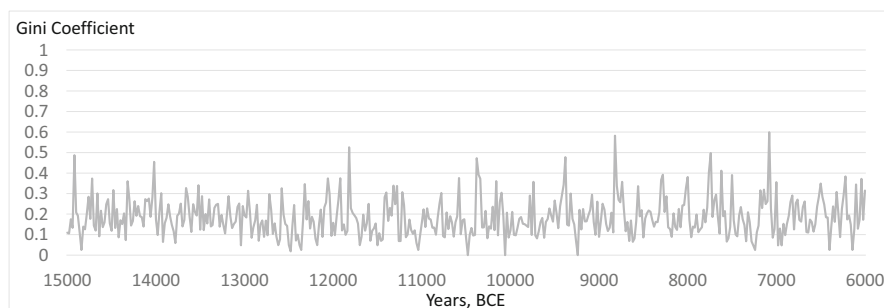


Fig. 33.3 Simulated income inequality in a subpopulation during the late Pleistocene and early Holocene. The excursions to occasionally high levels of inequality result from the breakdown in sharing resulting from a movement from the upper left in the simplex in Fig. 33.2 along the left edge and then down to the Hobbesian equilibrium at point **d**. (There is nothing unusual about the particular subpopulation shown in the figure; other subpopulations exhibit very similar results.) (Source: Own computations described in text based on the model in Bowles and Choi (2013))

33.4 Institutional Persistence and Change: Private Property Before States

The above simulation suggests that sharing and limited economic inequality during and before the Early Neolithic may have been sustained by two processes: first, mutually policed sharing similar to what occurs in the neighborhood of a mostly-Civic subpopulation, exemplifying the egalitarian social norms described by the term aggressive egalitarianism; and second, the process of intergroup competition in which the more cooperative and egalitarian (and hence higher-payoff) subpopulations tended to prevail.

The collectivist and egalitarian social order sustained by these processes was eventually replaced by a different institutional environment. This new order was one in which private property played a larger role, limiting within-group conflicts and thereby allowing a more individualistic and less egalitarian social order to prevail in inter-group competition. We now provide archaeological evidence and a modification of the Civic-Grabber-Sharer Game suggesting how this might have occurred.

33.4.1 Possession-Based Individual Property Before States

Hunting and gathering populations exploiting limited sites of unusual productivity—such as the migratory routes of salmon, eels, or gazelles mentioned above—routinely established group property rights and occasionally distinct rights for particular lineages or even households. Similar to these cases is what the archaeologist Curtis Marean terms the “coastal adaptation” that is based on dense, immobile, and hence defensible resources (Marean, 2014). Among sedentary hunter-gatherers occupying such sites, a shift from group-based territoriality to family-held rights of possession may have occurred.

An example is the community of maritime hunter-gatherers at Huaca Prieta on the northern coast of Peru who exploited the rich local open ocean fisheries using nets, floats, and other gear.⁹ It seems likely that the fishing gear and stores were owned by individuals or families. Subsequently, intensely cultivated lima beans and avocado became part of their diet and still other cultigens (including arboreal fruit and cotton) were added.

Raised farming platforms demarcated by ridges dating to about 5000 years ago have also been found, along with evidence of irrigation, indicating investment-intensive farming that would have been facilitated by private property rights. The people of Huaca Prieta appear to have first developed a system of private property rights as part of their fishing economy, and then extended these rights to crops, trees, and improved farmland. As Thomas Dillehay, one of the foremost archaeologists of the period puts it: “fishing configured early farming.”

⁹This account is based on Dillehay et al. (2012), Moseley (2006), and personal communication from Dillehay.

But aside from these exceptional cases, private ownership was not technically feasible for the most important sources of households' livelihoods prior to the advent of farming. Private ownership requires that the resource be both demarcated (delimited) and defended, and most foraged resources were so dispersed that demarcating and defending them was simply not cost-effective.

A typical foraging group, reports Bruce Winterhalder, might exploit a territory of 500 km² while a farming community an order of magnitude larger in population might exploit a quantity of land one-tenth this size (Winterhalder, 1990). What is important for the feasibility of privately held property is the size of the area exploited by a family or individual, rather than by a group. The sheer impossibility of demarcating and defending the usual territories exploited by foragers is suggested by the fact that the perimeters of the annual ranges of four Dobe Ju/'haoansi forager families averaged 325 km.¹⁰ Moreover, these family ranges were substantially overlapping, indicating that these foraging territories were not exclusive.

Not surprisingly, then, one does not get evidence of private storage or ownership of particular animals by individuals (or nuclear families) or field boundaries, until those resources are sufficiently valuable to justify the costs of establishing possession and defending ownership. This occurs (as pointed out initially by Dyson-Hudson and Smith (1978)) exceptionally in hunter-gatherer societies but on a grand scale only with farming. Cultivation of crops and animal husbandry greatly increased the productivity of land and animal assets, making them worth both demarcating and defending.

But where sedentism was possible, the appearance of private property in some of the essentials of life may have predated not only the first archaic proto-states, but also farming. Illustrative are the hunter-gatherers who initially settled at Abu Hureyra. Sedentism there was made possible by proximity to rich concentrations of resources (including the gazelle migrations mentioned above), allowing for the accumulation of valued objects (including dwellings), the possession of which would have been unambiguous, with access granted by permission of the owner and not by any collective right.

During the farming period at that settlement (termed Abu Hureyra 2) individuals were buried "under the floors of their houses." There is remarkable evidence that this occurred over successive generations of individuals within a given family at a particular dwelling (Moore et al., 2000, p. 506). Moore and his coauthors conclude that the evidence indicates that "the houses were family dwellings. It also indicates that a family could lay claim to the space its house occupied in the village, and its descendants could build a new house on the same spot in which to live. In other words, family rights to property were firmly established in Abu Hureyra 2" (p. 505).

At Catalhoyuk in central Anatolia (10,500–10,100 BP), according to the archaeologist Amy Bogaard and her coauthors, "families stored their own produce of grain, fruit, nuts and condiments in special bins deep inside the house. . ." (Bogaard et al., 2009, p. 649). Archaeologists Ian Kuijt and Bill Finlayson report that "Starting at

¹⁰Calculated from Kelly (1995) Fig. 5–3.

10,500 BP [Before the present, prior to plant domestication], food storage starts to be located inside houses, and by 9500 BP dedicated storage rooms appear in Neolithic villages. This transition from extra-mural to intra-mural storage system may reflect evolving systems of ownership and property . . . with food storage systems becoming part of household or individual based systems” (Kuijt & Finlayson, 2009, p. 10966). Much earlier indications of private storage come from Dobranichkevka and other sites on the Russian Plain as early as 18,000 years ago Russia (Soffer, 1985, 1989).

Further evidence of early private property is provided by the so-called *Hofplatz* model of settlement, in which European farmers more than seven millennia ago intensively tilled small plots adjacent to their homes, implying “a notion of ‘use-rights’ to areas of settlement based on membership in households” (Bogaard et al., 2011, p. 396). Raised-bed cultivation, fence posts, and other indicators of intensive cultivation and spatial demarcation (for example in New Guinea at least 6440 BP (Denham et al., 2004)) are also consistent with the emergence of private property along with farming, well before the first archaic proto-states.

33.4.2 Modeling the Origin of Private Property: The Civic-Bourgeois-Sharer Game

We now propose answers to two questions about institutional dynamics that come to mind in light of this evidence of early property rights: How could private property be sustained in the absence of some entity capable of enforcing the associated rights? And how could possession-based private property displace the cooperative sharing norms of distribution, given that this new convention (which, recall, is a mutual best response, conditional on most others doing the same) could not be adopted piecemeal.

The answer to the first question—how could it work without a state?—was provided by the biologist John Maynard Smith and his Hawk-Dove-Bourgeois Game. Maynard Smith introduced possession of a site (like a spider’s web or an animal’s foraging territory) as a means of coordinating social interactions. An actor playing his novel Bourgeois strategy would fight to defend a site if they were in possession of it but would not contest possession of another’s site. Maynard Smith showed that in a population in which most actors play the Bourgeois strategy, doing the same is a best response, thereby confirming the status of possession-based individual property—whether it be the spider’s web or stored grain—as a convention that could persist, once established without any enforcing authority.

But there is an important condition for this to be the case: possession must be sufficiently unambiguous. If possession is difficult or impossible to establish (as with the dispersed and mobile elements in a forager’s livelihood) Bourgeois actors will frequently engage in conflicts, in the extreme, their behavior approximates the situation (and low payoffs) of Grabbers in the previous model (Bowles, 2004). A key fact—the more readily established possession of the means of production in a farming than a foraging economy—then explains why private property may have been of more limited importance before the Neolithic Agricultural Revolution (Bowles & Choi, 2013, 2019).

The second question—how could this convention be introduced *de novo*—is more difficult to answer, as illustrated by the case of two would-be farmers, members of a group of foragers in Malaysia, the Batek: “The traditional Batek notions that all natural resources are unowned until collected and that any food obtained in excess of the needs of the procurer’s family must be shared with other families seem well suited to a nomadic foraging life, but wholly unsuited to . . . peasant farming. . .” (Endicott, 1988). The two Batek men who had discovered cultivated rice tried planting some. But their fellow group members simply harvested it (and, of course, felt obliged to share the harvest with the entire group). The hapless Batek would-be rice farmers went back to foraging, having discovered that, unlike a new crop or method of cultivation, a new system of property rights could not be introduced piecemeal.

To explore how the private property convention – modeled by the Bourgeois strategy—might have emerged, we modify the game by substituting the Bourgeois for the Grabber. There are three effects of this change. First, as can be seen in the right panel of Fig. 33.4, a population composed entirely of Bourgeois types (the lower right vertex) is an equilibrium. Second, the shaded basin of attraction of the Civics cum Sharers equilibria (the solid line on the left edge of the simplex) is now smaller. This means that the likelihood of drift or some exogenous shock displacing the Civics cum Sharers equilibria and propelling the population to the Bourgeois equilibrium is greater.

The third consequence, not shown in the figure, is that unlike the Hobbesian Grabber-Sharer equilibrium, there are no conflicts among members of a population at the Bourgeois equilibrium (again conditional on possession being unambiguous). The result is to nullify the previous advantage in superior average payoffs that before let the Civics cum Sharers populations prevail in their contests with the Hobbesian groups.

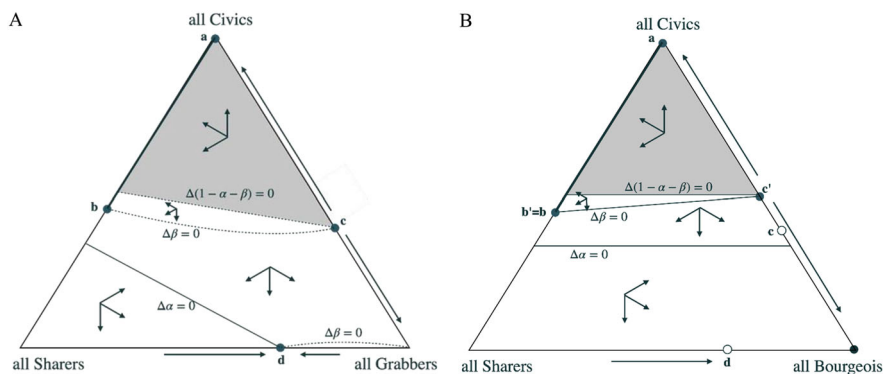


Fig. 33.4 The emergence of possession-based private property in the Civic-Sharer-Bourgeois Game (right panel). The left panel, shown here for ease of comparison, is identical to Fig. 33.2. As in Fig. 33.2, the line labeled $\Delta\alpha = 0$ gives the population distributions for which the fraction of Sharers in the population is unchanging, and analogously for $\Delta\beta = 0$ (fraction of Grabbers on the left, Bourgeois on the right is unchanging) and for $\Delta(1-\alpha-\beta) = 0$ (fraction of Civics unchanging). Point **c** is an unstable stationary state composed of just Civics and grabbers. The shaded area is the basin of attraction of the “no Grabbers” set of equilibria. In the right panel the lower right vertex is the Bourgeois equilibrium and the former Grabber-Sharer equilibrium at point **d** no longer exists. (a) Aggressive (all Civic) egalitarianism. (b) A private property (all Bourgeois) alternative

This third consequence is decisive. Recall that, prior to the advent of farming (and the concomitant ability to define possession unambiguously), a population initially composed mostly of Civics could drift into the basin of attraction of the Hobbesian equilibrium. But this within-subpopulation evolutionary vulnerability was more than offset by the superior average payoffs of mostly Civic populations, resulting in the defeat of Hobbesian populations in between sub-population contests, restoring a predominantly Civic sub-population. Now, by exploiting possession (or not) of an asset as a coordinating device, the Bourgeois convention is able to sustain average payoffs on a par with the Civic equilibrium, and so eliminates the Civic's success in between-group contests as compensation for their vulnerability to drift.

To determine if this dynamic could provide an explanation of the increasing importance of private property, we simulated the agent-based model described with two key modifications. The first is that possession of hunted or gathered resources is ambiguous while for farmed resources this is not the case. The second is that we used estimates of the volatility of temperature (and other aspects of climate) over the past 40,000 years (shown in Fig. 33.5 as the black line) to capture the degree to which these favored a mobile hunting and gathering livelihood or a sedentary farming life.

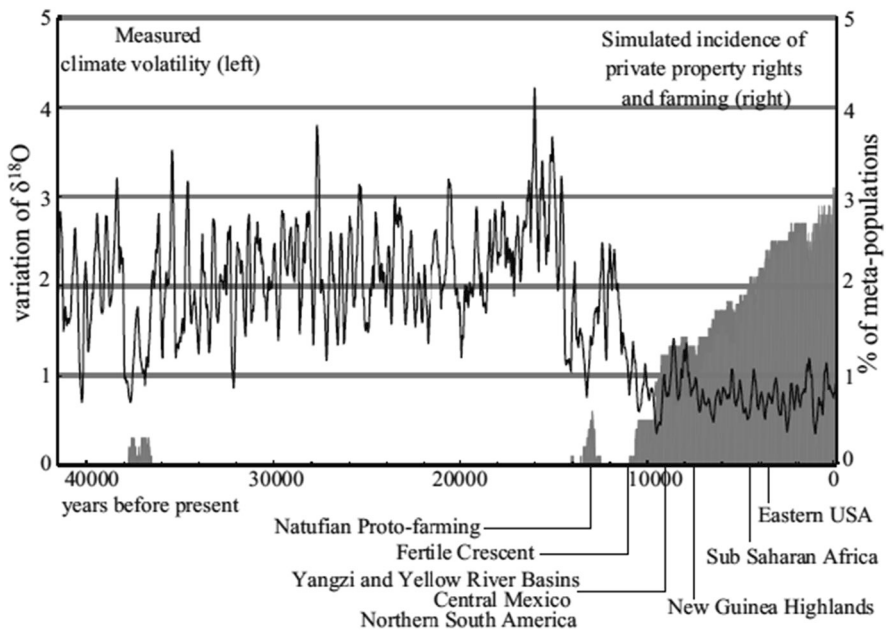


Fig. 33.5 Climate variability and the prehistoric emergence of agriculture. Dates on the horizontal axis are years before the present. The grey bars (right axis) indicate the number of simulated populations (of 1000) in which more than 50% of the entire population were Bourgeois Agriculturalists. Estimated dates of some well-studied cases of the emergence of agriculture are on the horizontal axis (Bellwood, 2005). Climate variability (the line, left axis) is the 100-year maximum difference in surface temperature measured by levels of $\delta^{18}\text{O}$ from Greenland ice cores backward smoothed over 200 years to minimize spurious variation due to measurement error. (Source: Bowles and Choi (2013))

As in the previous simulation, actors could choose to farm or forage and to behave according to three strategies: Civic, Sharer, and Bourgeois. Figure 33.5 shows that in the simulated populations with the amelioration of climate, private property and farming co-evolved. We do not have sufficient archaeological evidence to track the growing importance of private property at each of the sites where farming was first taken up, but it is notable that the appearance of the first farming sites and the presumably concomitant emergence of private property in stores, animals, and eventually land, all occurred well before the first proto-states in the various parts of the world where farming was independently developed.

33.5 Egalitarian Wealth Distributions Prior to States

The two institutions before states that we have described and modeled—the mostly-Civic convention that policed sharing resources and the Bourgeois convention that used possession of property as a coordinating device—were both associated with limited levels of wealth inequality in western Eurasia prior to the Bronze Age (beginning about 5300 years ago). We know this from a new data set of comparable wealth inequality data based (for the Neolithic, Bronze, and Iron Ages) on house sizes, storage areas, burial goods, and in a few cases acreage of land cultivated (Bowles & Fochesato, 2024).

The relevant evidence is in Fig. 33.6, where the first instance of what we consider to be an archaic proto-state is the Mesopotamian site at Tepe Gawra Level XI in the 4th millennium BCE. The most reliable date for the emergence of farming (meaning annual cultivation, often prior to domestication) in the region is the 9th millennium BCE (Bowles & Bogaard, 2024). Thus it appears that, for many millennia, farming economies with extensive private property rights were extraordinarily egalitarian, at least by modern standards. We do not have sufficient archaeological data on wealth inequalities among prehistoric hunter-gatherers in the region, to make any informed inferences about the degree of wealth inequality in economies that might have been governed by the mostly-Civic convention; the data that are available are for sedentary populations (mobile hunter-gatherers—foragers—leave few archaeological footprints). On the basis of what we know, it seems unlikely that farming economies were markedly more unequal than sedentary hunting and gathering economies prior to the development of animal traction and other capital-intensive means of production. What can be determined with some certainty is that wealth inequality was substantially less in the western Eurasian economies, that were not governed by states than in those governed by the first archaic proto-states.¹¹ But in describing the farming and hunting and gathering economies before states, we see no reason to adopt the utopian overtones of Engels' term "primitive communism." The limited disparities of material wealth notwithstanding, experienced inequalities may

¹¹ The evidence is summarized in Bowles and Fochesato (2024).

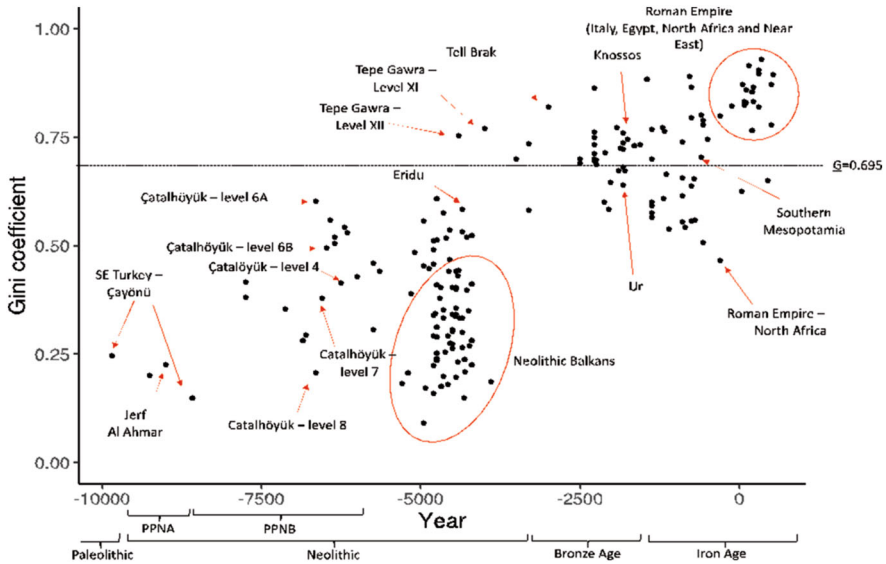


Fig. 33.6 Inequality in material wealth in western Eurasia from the end of the Paleolithic to the Roman Empire. Dates on the horizontal axis are years before the Common Era (BCE). PPNA and PPNB refer to the western Asian Pre-Pottery Neolithic periods A and B periodization used by archaeologists. Our database is $n = 185$ estimates of the Gini coefficient (including three from the late Paleolithic, not shown in this figure). The line labeled $G = 0.695$ indicates the mean of 182 comparable estimates of wealth inequality in state-governed societies in which slavery did not play a major role in production for the period from 300 BCE to 2000 CE. (Source: Bowles and Fochesato (2024))

nonetheless been substantial. In both hunter-gatherer and early farming populations, people may have cared more about the size of their family than the size of their dwelling or the other indicators of material wealth that we measure with archaeological data.

By this measure (Gini coefficients of reproductive success), ethnographic hunter-gatherer populations (even foragers, that is, mobile hunter-gatherers) exhibit substantial levels of inequality, certainly on a par with material wealth inequality among the Early Neolithic western Eurasian farmers in our data set (Borgerhoff Mulder et al., 2009; Hill & Hurtado, 1996; Howell, 2000; Ross et al., 2023). Moreover, there may be a genetic footprint of the emergence of substantial levels of polygyny—contributing to inequality in reproductive success among men—during the Neolithic in western Eurasia, that is prior to the emergence of the archaic proto-state, followed by its demise during the Bronze Age (Bowles et al., 2023; Karmin et al., 2015).

Moreover, mortal conflicts between hunter-gatherer groups appear to have been frequent. Archaeological and ethnographic evidence suggest that the fraction of deaths due to intergroup conflict among hunter-gatherer groups may have rivaled that associated with the interstate rivalries of the twentieth century (Bowles, 2006).

33.6 Discussion

What the emergence of the first archaic proto-states and the subsequent consolidation of more nearly Weberian states accomplished, then, was not to have instituted order in a previously institution-free and chaotic Hobbesian state of nature. Instead the state established an entirely new social order, one under which modern levels of wealth inequality could endure over long periods of time. Our reasoning and models have provided a possible explanation for both of the long-term persistence of two conventions that appears to have supported substantially egalitarian wealth distributions—mutually policed sharing and mutually recognized possession-based property—as well as an account of how a transition from the former to the latter may have occurred, all of this taking place prior to the concentration of coercive political power in archaic proto-states.

What we still lack is an explanation of how this concentration of coercive power in elite hands could have been imposed on a population still gaining some of their livelihood from hunting and hence heavily armed and skilled at killing. The puzzle may be resolved in light of the (by the early Bronze Age) substantial preexisting wealth inequality and the collective benefits of the state as a coordinating device including the provision of public goods and irrigation on a substantial scale (Allen et al., 2022; Fargher & Blanton, 2017; Fargher et al., 2011).¹² The evolutionary success of state-based governance (in its various forms) is also almost certainly due to the military advantages associated with the larger demographic units that states were capable of governing.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., & Robinson, J. A. (2009). Foundations of societal inequality. *Science*, 326, 678–679.
- Allen, R., Bertazzini, M., & Heldring, L. (2022). The economic origins of government. *American Economic Review*, 113(10), 2507–2545.
- Bahuchet, S. (1990). Food sharing among the pygmies of central africa. *African Study Monographs*, 11, 27–53.
- Bahuchet, S. (1991). Les pygmées d’aujourd’hui en Afrique centrale. *Journal des Africanistes*, 61, 5–35.
- Bellwood, P. (2005). *First farmers: The origins of agricultural societies*. Blackwell.
- Binford, L. (2001). *Constructing frames of reference: An analytical method for archeological theory using hunter-gatherer and environmental data sets*. University of California Press.
- Binmore, K., Samuelson, L., & Peyton Young, H. (2003). Equilibrium selection in bargaining models. *Games and Economic Behavior*, 45, 296–328.

¹²Norman Yoffee and Gojko Barjamovic suggest that a degree of concentrated political power was crucial to providing the institutional and physical infrastructure of private trade at Assur. But they write that “traders moved goods for centuries before centralized states appeared in Mesopotamia” Yoffee and Barjamovic (2018) p. 22.

- Blurton-Jones, N. G. (1984). A selfish origin for human food sharing. *Ethology and Sociobiology*, 5(1), 1–3.
- Boehm, C. (1993). Egalitarian behavior and reverse dominance hierarchy. *Current Anthropology*, 34(3), 227–254.
- Boehm, C. (2000). *Hierarchy in the forest*. Harvard University Press.
- Boehm, C. (2012). Variance reduction and the evolution of social control. *Santa Fe Institute working paper*.
- Bogaard, A., Charles, M., Twiss, K. C., Fairbairn, A., Yalman, N., Filipovic, D., Demirergi, G. A., Ertug, F., Russell, N., & Henecke, J. (2009). Private pantries and celebrated surplus: Storing and sharing food at neolithic catalhoyuk, central anatolia. *Antiquity*, 83, 649–668.
- Bogaard, A., Krause, R., & Stroem, H.-C. (2011). Towards a social geography of cultivation and plant use in an early farming community: Vaihingen an der enz, south-West Germany. *Antiquity*, 85, 395–416.
- Borgerhoff Mulder, M., Bowles, S., Hertz, T., Bell, A., Beise, J., Clark, G., Fazzio, I., Gurven, M., Hill, K., Hooper, P. L., Irons, W., Kaplan, H., Leonetti, D., Low, B., Marlowe, F., Mcelreath, R., Naidu, S., Nolin, D., Piraino, P., Quinlan, R., Schniter, E., Sear, R., Shenk, M., Smith, E. A., Von Rueden, C., & Wiessner, P. (2009). Intergenerational wealth transmission and the dynamics of inequality in small-scale societies. *Science*, 326(5953), 682.
- Bowles, S. (2004). *Microeconomics: Behavior, institutions, and evolution*. Princeton University Press.
- Bowles, S. (2006). Group competition, reproductive leveling and the evolution of human altruism. *Science*, 314, 1569–1572.
- Bowles, S., & Bogaard, A. (2024). Can marshall plus malthus explain the evolution of ancient societies? A review essay on *economic prehistory* by dow and reed. *Journal of Economic Literature*, 62(3), 1213–1229.
- Bowles, S., & Choi, J.-K. (2013). The co-evolution of farming and private property in the early holocene. *Proceedings. National Academy of Sciences. United States of America*, 110(22), 8830–8835.
- Bowles, S., & Choi, J.-K. (2019). The neolithic agricultural revolution and the origin of private property. *Journal of Political Economy*, 127(5), 2186–2228.
- Bowles, S., & Fochesato, M. (2024). The origins of enduring economic inequality. *Journal of Economic Literature*, 62(4), 1475–1537.
- Bowles, S., & Gintis, H. (2012). *A cooperative species: Human reciprocity and its evolution*. Princeton University Press.
- Bowles, S., Hwang, S. H., & Naidu, S. (2017). The evolution of egalitarian socio-linguistic conventions. *American Economic Review*, 107(5), 572–577.
- Bowles, S., Ross, C. T., Mulder, M. B., & Seung-Yun, O. (2023). The decline of polygyny: An interpretation. *Unpublished manuscript*.
- Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, 328, 617–620.
- Celik, B. (2015). New neolithic cult centers and domestic settlements in light of the urfa region surveys. *Documenta Praehistorica*, XLII, 353–364.
- Cosmides, L., & Tooby, J. (2010). *Evolutionary psychology: A primer*. University of California at Santa Barbara.
- Denham, T., Haberle, S., & Lentfer, C. (2004). New evidence and revised interpretations of early agriculture in highland new Guinea. *Antiquity*, 78(302), 839–857.
- Diakonoff, I. M. (1969). The rise of the despotic state in ancient Mesopotamia. In I. M. Diakonoff (Ed.), *Ancient mesopotamia* (pp. 173–203). Nauka. (translation of the 1956 Russian version).
- Dietrich, O., Heun, M., Notroff, J., Schmidt, K., & Zarnkow, M. (2012). The role of cult and feasting in the emergence of neolithic communities. New evidence from göbekli tepe, South-Eastern Turkey. *Antiquity*, 86, 674–695.
- Dillehay, T. D., Bonavia, D., & Others. (2012). Chronology, mound-building and environment at huaca prieta, coastal Peru, from 13,700 to 4000 years ago. *Antiquity*, 86, 48–70.
- Dowling, J. (1968). Individual ownership and the sharing of game in hunting socieites. *American Anthropologist*, 70, 502–507.

- Dyson-Hudson, R., & Smith, E. A. (1978). Human territoriality: An ecological assessment. *American Anthropologist*, 80, 21–41.
- Eldredge, N., & Gould, S. J. (1972). Punctuated equilibria: An alternative to phyletic gradualism. In T. J. M. Schopf & J. M. Thomas (Eds.), *Models in paleobiology* (pp. 82–115). Freeman, Cooper.
- Endicott, K. (1988). Property, power and conflict among the batek of Malaysia. In T. Ingold, D. Riches, & J. Woodburn (Eds.), *Hunters and gatherers* (pp. 110–127). St. Martin's Press.
- Engels, F. (2010 [1884]). *The origin of the family, private property, and the state*. Penguin.
- Fargher, L., & Blanton, R. (2017). The independent republic of tlaxcallan. In D. Nichols & E. Rodriguez-Alegria (Eds.), *The oxford handbook of the aztecs* (pp. 535–542). Oxford University Press.
- Fargher, L., Heredia, V. Y., & Espinoza and Richard Blanton. (2011). Alternative pathways to power in late postclassic highland mesoamerica. *Journal of Anthropological Archaeology*, 30, 306–326.
- Feinman, G. M., & Marcus, J. (1998). Introduction. In G. M. Feinman & J. Marcus (Eds.), *Archaic states* (pp. 3–15). School of American Research Press.
- Gelb, I. J. (1967). Approaches to the study of ancient society. *Journal of the American Oriental Society*, 87, 1–8.
- Graeber, D., & Wengrow, D. (2021). *The dawn of everything: A new history of humanity*. Farrar, Straus and Giroux.
- Gurven, M., Hill, K., Kaplan, H., Hurtado, A. M., & Lyles, R. (2000). Food transfers among hiwi foragers of Venezuela: Tests of reciprocity. *Human Ecology*, 28(2), 171–218.
- Gurven, M., Allen-Arave, W., Hill, K., & Hurtado, A. M. (2001). Reservation food sharing among the ache of Paraguay. *Human Nature*, 12, 273–297.
- Gurven, M., Hill, K., & Kaplan, H. (2002). From forest to reservation: Transitions in food sharing among the ache of Paraguay. *Journal of Anthropological Research*, 58, 93–120.
- Hames, R. (2000). Reciprocal altruism in yanomamo food exchange. In L. Cronk, N. Chagnon, & W. Irons (Eds.), *Adaptation and human behavior: An anthropological perspective*. Aldine Gruyter.
- Hames, R., & McCabe, C. (2007). Meal sharing among the ye'kwana. *Human Nature*, 18, 1–21.
- Hammerstein, P., & Reichert, S. (1988). Payoffs and strategies in spider territorial contests: Ess analysis of two ecotypes. *Evolutionary Ecology*, 2, 115–138.
- Hardin, G. (1968). The tragedy of the commons. *Science*, 162(3859), 1243–1248.
- Hardin, G. (1978). Political requirements for preserving our common heritage. In H. P. Bokaw (Ed.), *Wildlife and America* (pp. 310–317). Washington Council on Environmental Quality.
- Hawkes, K., O'connell, J. F., & Blurton-Jones, N. G. (2001). Hadza meat sharing. *Evolution and Human Behavior*, 22, 113–142.
- Hayden, B. (2001). Richman, poorman, beggarman, chief: The dynamics of social inequality. In G. Feinman & T. Price (Eds.), *Archaeology at the millennium* (pp. 231–272). Kluwer/Plenum.
- Hill, K., & Hurtado, A. M. (1996). *Ache life history: The ecology and demography of a foraging people*. Aldine de Gruyter.
- Hill, K., & Kaplan, H. (1989). Population and dry season subsistence strategies of the recently contacted yora of Peru. *National Geographic Research Report*, 5, 317–334.
- Hobbes, T. (2005). *Leviathan*. Continuum. [1651].
- Hodder, I. (2014). Çatalhöyük: The leopard changes its spots. *Anatolian Studies*, 64, 1–22.
- Howell, N. (2000). *The demography of the dobe!Kung* (2nd ed.). Aldine de Gruyter.
- Kaplan, H., & Gurven, M. (2005). The natural history of human food sharing and cooperation: A review and a new multi-individual approach to the negotiation of norms. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral sentiments and material interests: The foundations of cooperation in economic life*. MIT Press.
- Kaplan, H., & Hill, K. (1985). Food sharing among ache foragers: Tests of explanatory hypotheses. *Current Anthropology*, 26(2), 223–246.
- Kaplan, H., Hill, K., Lancaster, J., & Hurtado, A. M. (2000). A theory of human life history evolution: Diet, intelligence, and longevity. *Evolutionary Anthropology*, 9(4), 156–184.
- Karmin, M., Saag, L., Vicente, M., et al. (2015). A recent bottleneck of y chromosome diversity coincides with a global change in culture. *Genome Research*, 25, 459–466.

- Kelly, R. L. (1995). *The foraging spectrum: Diversity in hunter-gatherer lifeways*. Smithsonian Institution Press.
- Kirch, P. (2010). *How chiefs became kings: Divine kingship and the rise of archaic states in ancient Hawai'i*. University of California Press.
- Knauff, B. M. (1991). Violence and sociality in human evolution. *Current Anthropology*, 32(4), 391–428.
- Kroeber, A. L. (1960). Fishing among the Indians of northwestern califoria. *Anthropological Records*, 21(1), 135–141.
- Kuijt, I. (1996). Negotiating equality through ritual: A consideration of late natufian and perpotery neolithic a period mortuary practices. *Journal of Anthropological Archaeology*, 15, 313–336.
- Kuijt, I., & Finlayson, B. (2009). Evidence for food storage and predomestication granaries 11,000 years ago in the Jordan Valley. *PNAS*, 106(27), 10966–10970.
- Kuijt, I., & Goring-Morris, N. (2002). Foraging, farming and social complexity in the pre-pottery neolithic of the southern levant: A review and synthesis. *Journal of World Prehistory*, 16, 361–440.
- Kummer, H. (1991). Evolutionary transformations of possessive behavior. *Journal of Social Behavior and Personality*, 6(6), 75–83.
- Kummer, H., & Cords, M. (1991). Cues of ownership in long-tailed macaques, *macacca fascicularis*. *Animal Behavior*, 42, 529–549.
- Lewis, D. K. (1969). *Convention: A philosophical study*. Harvard University Press.
- Lourandos, H. (1997). *Continent of hunter-gatherers*. Cambrdige University Press.
- Mahdi, N. Q. (1986). Pukhutunwali: Ostracism and honor among pathan hill tribes. *Ethology and Sociobiology*, 7(3–4), 295–304.
- Marean, C. (2014). The origins and significance of coastal resource use in africa and western eurasia. *Journal of Human Evolution*, 77, 17–40.
- McBrearty, S., & Brooks, A. (2000). The revolution that wasn't: A new interpretation of the origin of modern human behavior. *Journal of Human Evolution*, 39, 453–563.
- Moore, A. M. T., Hillman, G. C., & Legge, A. J. (2000). *Village on the euphrates: From foraging to farming at abu hureyra*. Oxford University Press.
- Moseley, M. (2006). Preceramic civilization in coastal Peru: Revolutionary data, pioneering interpretations. *The Review of Archaeology*, 27, 72–81.
- North, D. C. (1981). *Structure and change in economic history*. W. W. Norton.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
- Ross, C. T., Hooper, P., Smith, J., Smith, E. A., Bowles, S., Jaeggi, A., et al. (2023). Reproductive inequality in humans and other mammals. *Proceedings. National Academy of Sciences. United States of America*, 120(22).
- Roux, J. C., Gerard Der A., Michel, B., & Danielle, S. (2000). Les batiments communautaires de Jerf el Ahmar e Mureybet. *Paleoorient*, 26(1), 29–44.
- Schotter, A. (1981). *Economic theory of social institutions*. Cambridge University Press.
- Service, E. R. (1962). *Primitive social organization: An evolutinary perspective*. Random House.
- Service, E. R. (1975). *The origin of the state and civilization: The process of cultural evolution*. Norton.
- Sigg, H., & Falett, J. (1985). Experiments on respect of possession in hamadryas baboons (*papio hamadryas*). *Animal Behavior*, 33, 978–984.
- Skinner, Q. (2002). *Visions of politics. Volume 3: Hobbes and civil science*. Cambridge University Press.
- Soffer, O. (1985). Subsistence practices and durationof occupation in the late valdai. In *The upper paleolithic of the central russian plain* (pp. 249–353). Academic.
- Soffer, O. (1989). Storage, sedentism and the eurasian palaeolithic record. *Antiquity*, 63, 719–732.
- Stanford, C. (1999). *The hunting apes: Meat eating and the origins of human behavior*. Princeton University Press.
- Stasavage, D. (2020). *The decline and rise of democracy: A global history from antiquity to today*. Princeton University Press.
- Sugden, R. (1986). *The economics of rights, co-operation and welfare*. Basil Blackwell.
- Sugden, R. (1989). Spontaneous order. *Journal of Economic Perspectives*, 3(4), 85–97.

- Trigger, B. (1990). Maintaining economic equality in opposition to complexity: An iroquoian case study. In S. Upham (Ed.), *The evolution of political systems* (pp. 119–145). Cambridge University Press.
- Trigger, B. (2003). *Understanding early civilizations*. Cambridge University Press.
- Twiss, K., Bogaard, A., Bogdan, D., Carter, T., Charles, M., Farid, S., Russell, N., Stevanovic, M., Yalman, E. N., & Yeomans, L. (2008). Arson or accident? The burning of a neolithic house at catalhoyuk, Turkey. *Journal of Field Archaeology*, 33(1), 41–57.
- Verhoeven, M. (2000). Death, fire, and abandonmnet. Ritual practice at late neolithic tel sabi abyad. *Archaeological Dialogues*, 7(1).
- Weber, M. (1978 [1921]). *Economy and society: An outline of interpretive sociology*. University of California Press.
- Weibull, J. W. (1995). *Evolutionary game theory*. MIT Press.
- Weide, A., Green, L., Hodgson, J., Douche, C., Tengberg, M., Whitlam, J., Dovrat, G., Osem, Y., & Bogaard, A. (2022). A functional ecological model reveals the nature of early plant magement in southwest asia. *Nature Plants*, 8, 623–634.
- Wiessner, P. (1982). Leveling the hunter: Constraints on the status quest in foraging societies. In P. Wiessner & W. Schiefenhovel (Eds.), *Food and the status quest: An interdisciplinary perspective* (pp. 171–191). Berghahn.
- Wiessner, P. (2005). Norm enforcement among the ju/'hoansi bushmen: A case of strong reciprocity? *Human Nature*, 16(2), 115–145.
- Willcox, G., & Storedeur, D. (2012). Large scale cereal processing before domestication during the tenth millennium cal bc in northern Syria. *Antiquity*, 86, 99–114.
- Winterhalder, B. (1990). Open field, common pot: Harvest variability and risk avoidance in agricultural and foraging societies. In E. A. Cashdan (Ed.), *Risk and uncertainty in tribal and peasant economiies* (pp. 67–87). Westview Press.
- Wobst, H. M. (1978). The archaeo-ethnology of hunter-gatherers or the tyranny of the ethnographic record in archaeology. *American Antiquity*, 43(2), 303–309.
- Woodburn, J. (1982). Egalitarian societies. *Man*, 17, 431–451.
- Wright, K. (2014). Domestication and inequality? Households, corporate groups and food processing tools at neolithic catalhoyuk. *Journal of Anthropological Archaeology*, 33, 1–33.
- Yoffee, N. (1995). Political economy in early mesopotamian states. *Annual Review of Anthropology*, 24, 281–311.
- Yoffee, N. (2005). *Myths of the archaic state: Evolution of the earliest cities, states, and civilizations*. Cambridge University Press.
- Yoffee, N., & Barjamovic, G. (2018). Old assyrian trade and economic history. In K. Kleber, G. Neumann, & S. Paulus (Eds.), *Grenzüberschreitungen: Studien zur kulturgeschichte des alten orients* (pp. 815–824). Zaphon.
- Young, H. P. (1995). Economics of conventions. *Journal of Economic Perspectives*, 10(2), 105–122.
- Young, H. P. (1998). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Culture Versus Institutions in the Great Enrichment

34

Joel Mokyr

Contents

34.1	Introduction: An Institutional Puzzle	898
34.2	Cultural Drivers of Institutional Change: The Industrial Enlightenment	902
34.3	High-Powered and Low-Powered Incentives	910
34.4	Conclusions	919
	References	921

Abstract

Is there a connection between the Industrial Revolution and the eighteenth century Enlightenment? The Enlightenment was a very broad cultural movement with many moving parts, but the one most relevant to economic history is what is known as the Industrial Enlightenment, a movement that directly advocated institutions that supported and promoted the accumulation and dissemination of technological progress and economic growth. These institutions included those that directly helped incentivize the accumulation of useful knowledge such as patronage and reputation effects, as well as patents, prizes, and other rewards to successful inventors. The movement also supported dissemination of knowledge through scholarly societies, publications, and various informal networks. The chapter concludes that both culture and the institutions it supported created a fertile ground for the rapid economic growth ignited by the Industrial Revolution.

This chapter is partially based on the keynote lecture delivered to the Associazione di Storia Economica (ASE) on September 17, 2021. I am indebted to the editors of this *handbook* and two anonymous referees for excellent comments and suggestions. Franco Malpassi and Chris Sims provided excellent research assistance.

J. Mokyr (✉)

Departments of Economics and History, Northwestern University, Evanston, IL, USA

Berglas School of Economics, Tel Aviv University, Tel Aviv, Israel

e-mail: J-Mokyr@northwestern.edu

© The Author(s) 2025

C. Ménard, M. M. Shirley (eds.), *Handbook of New Institutional Economics*,
https://doi.org/10.1007/978-3-031-50810-3_34

897

Keywords

Technological change · Enlightenment · Industrial Revolution · Economic progress · Innovation · Incentives · Dissemination

34.1 Introduction: An Institutional Puzzle

Despite some eminent dissenters who have yet to be persuaded, the growing consensus among economic historians today is that “institutions matter” even if they do not necessarily “rule.”¹ Following the pioneering work of Douglass North in the 1980s, economists have come to see institutions as the rules of the economic game, determining the incentives and constraints that economic players (individuals, firms, and organizations) face, and that they matter to the economic outcomes and performances. Institutions thus matter both to economic performance and to the distribution of resources and power. They determine the extent to which society can overcome opportunistic behavior, reduce transactions costs, and thus make markets work. In addition, they make collective action possible that creates public goods and organizations. These include the enforcement of property rights and contracts and more generally a rule of law and order necessary to facilitate exchange and a more efficient allocation of resources. These arrangements are critical to what is known as Smithian growth, the kind of economic growth that is driven by deepening and widening of markets, specialization and hence gains from trade, and better allocations of factors.²

The historical puzzle is that institutions supporting growth through gains from trade and better allocation of resources do not explain modern economic growth as it emerged during and after the British Industrial Revolution. While specialization and more efficient allocations continued to be an important driver of the rising economic performances of Western countries, the *primum movens* of economic growth increasingly became what eighteenth century writers called “useful knowledge,” leading to both process and product innovation and hence to rising productivity known as Schumpeterian growth. Useful knowledge included what we would call “science,” but it was much more than that, such as engineering (Mokyr et al., 2021; Hanlon, 2022), practical mathematics (Kelly & Ó Gráda, 2021), and industrial chemistry (Clow & Clow, 1952; Christie, 2018). Beyond that, there was a large accumulation of what we would call today artisanal tacit knowledge. Although such tacit knowledge would not qualify by any stretch as “science,” it stood Britain in good service as exemplified by the achievements of relatively poorly educated unscientific

¹For a recent critique of institutions as a prime mover in economic history, see especially McCloskey (2010, 2016, 2021b) and Clark (2007). The term “institutions rule” is borrowed from Rodrik et al. (2004).

²Institutions have also been placed at the center of the literature on the distribution of income and political power (and indirectly on economic power). This aspect of institutions has been discussed at length by Acemoglu et al. (2005) and Acemoglu and Robinson (2012).

“tinkerers” such as George Stephenson, Joseph Bramah, Charles Tennant, and Richard Roberts.

Useful knowledge, both innovation and diffusion, were at the heart of the Industrial Revolution and everything that came after. To summarize the main message of this chapter: institutions played a crucial role in the Industrial Revolution, just as they had in Smithian growth in the centuries before. But these were different institutions, encouraging innovation in addition to well-functioning markets and better allocations. Some of them were specifically innovation-friendly, while others created a more general environment that made progress possible. But to understand such institutions, we need to examine the beliefs and attitudes that supported them and that were part of the Enlightenment.

Why Europe? The Enlightenment was a European-world phenomenon. To be sure, the deeper roots go back further in time. The technologies that constituted the Industrial Revolution were new, but Europe’s advantage in generating them had started before the Industrial Revolution and should be therefore seen as causal and not endogenous to the Industrial Revolution. Whether or not we wish to call it a “Scientific Revolution” or not, in the centuries we associate with the Renaissance and the Baroque eras, Europeans increased their knowledge of the phenomena and regularities that govern the natural world at an accelerating rate, pulling ahead of the rest of the world. Progress continued apace in the decades after the publication of Newton’s *Principia* in 1687. On the eve of the Industrial Revolution, Dr. Johnson in his *Rasselas, Prince of Abyssinia* wrote a famous paragraph explaining why Europeans could easily visit Africa and Asia “for trade and conquest” but not the reverse. The imaginary African philosopher replied, “they are more powerful than we, sir, because they are wiser; knowledge will always predominate over ignorance. But why their knowledge is more than ours I know not” (Johnson, 1759, p. 49).

Within Europe, there were differences as well. Useful knowledge should have been spread fairly evenly across Western Europe, but the capability to deploy it in a profitable way surely was not. The main reason was the sharp inequality in tacit, artisanal knowledge, taught through personal contact and often unmodified. In that area, Britain held an advantage that was, for a while, unassailable and secured it the leadership role in technological progress in the first Industrial Revolution. The large literature explaining British precocity has provided a number of reasons that complement one another, among them geography such as the advantages of being an Island and the presence of minerals and coal; institutions such as the growth of the rule of law, constraints on the monarchy, and the norms that supported the existence of social capital and resulted in what Judge Blackstone famously called “a polite and commercial people” (Blackstone, (1765–69), Book III, Chap. 22; for details, see Kelly et al., 2014, 2023).

To repeat: this not to say that other factors did not play an important role. But the institutions that supported the growth and diffusion of useful knowledge were crucial. In the absence of continuously expanding knowledge, the other forces that could generate economic growth, including a growing international trade sector and rising labor mobility would have run into diminishing returns and eventually ground to a halt. In a world of purely Smithian growth, there would have been no Industrial

Revolution; the growth of prosperity due to the division of labor, regional specialization, more efficient markets, and other elements in Smith's idea of what drove growth would have mercilessly slowed down and fizzled out and the world would have ended up in the stationary state envisaged by classical economics, driven by the curse of concavity.

While I will be largely concerned with the institutional roots of useful knowledge and technological progress, it should be stressed that the institutional origins of Britain's Industrial Revolution were larger than just their effect on innovation (Mokyr, 2008). By 1700, Britain's institutions had developed features that set it apart from much of the rest of Europe. The seminal paper that related political institutions to the Industrial Revolution, published in 1989 by North and Weingast (1989), is well-known and requires little discussion. It established the view that emphasized the Glorious Revolution and the rise of Parliament as the central locus of power in England (later Britain) and the creation of effective constraints on the executive. In their view, these were crucial to secure property rights and the rule of law, and hence were the keys to subsequent economic progress. Parliament became "the place where absolute despotic power, which must in all governments reside somewhere, is entrusted," as Blackstone noted (1765–69, Book 1, Chap. 2, section III). It became what may be called a "meta-institution" that is, an institution that had enough legitimacy to create other institutions.

There were at least three other institutions that set Britain apart and that have been argued to be factors in bringing about economic progress. One was the English Poor Law. Beginning in the sixteenth century with the dismantlement of Church institutions, the English state gradually displaced the Church and voluntary associations as the main provider of poor relief. The "Old Poor Law" of 1601 formalized this system. In a seminal paper, Solar (1995) suggested a variety of ways in which the Poor Law contributed to British economic development and industrialization. As a form of insurance, he argued, it was relatively free from the usual banes of insurance markets, moral hazard, and adverse selection. It may have reduced the need to own land and encouraged the emergence of wage labor and cottage industries, which were riskier and more volatile sources of income. Thus, the Poor Law helped create a rural proletariat force necessary to create the capitalist, market-oriented farming system that made Britain's agriculture the most productive and efficient in Europe. Solar presented his claim that the security provided by the Old Poor Law fostered economic development as a hypothesis.³ A second institution that played an important role was the apprenticeship system, which in Britain had been largely taken out of the control of England's rather weak guilds and was organized more as a market-driven system, with contract enforcement increasingly the responsibility of local authorities and courts (Mokyr, 2019; Wallis, 2025). It has been argued that the effectiveness of its apprenticeship system provided Britain with a high-quality skilled artisanal class that was famous across Europe and provided it with the crucial

³Further arguments for the importance of the Poor Law in economic development of Britain are furnished in Greif and Iyigun (2013) and Greif et al. (2013).

technological prowess and competence to make the key inventions of the Industrial Revolution into a reality.⁴ A third advantage was the relative integration of the British economy due an important institutional difference between Britain and its continental competitors: the absence of internal tariff barriers and the high development of coastal shipping (driven by market forces) made regional specialization far more advanced than elsewhere, something that was fully recognized by the Continental nations and eventually corrected. But there was more: Britain had developed the institutions to make a privately financed transport infrastructure of highways, canals, and harbors feasible (Szostak, 1991; Bogart, 2011, 2019). At first glance, the transportation system would seem to have encouraged Smithian rather than Schumpeterian growth, but as argued in Kelly et al. (2023), the integration of markets and the regional specialization it supported may also have been an important factor in the advent of the Industrial Revolution.

Those institutional features clearly help in explaining the Industrial Revolution. They may be seen as necessary but insufficient conditions. All the same, more was needed. Institutions could be seen as the fertile British soils in which the seeds of the Industrial Revolution were planted. However, the Industrial Revolution was not a case of spontaneous generation. The seeds for the many advances in power technology, metals and other materials, chemistry, textiles, food processing, lighting, and other areas had to be produced by a system that made innovation happen at a much higher rate and larger scale than ever before, and that created economic modernity as we know it.

In what follows, I will raise the question of which institutional elements were behind the growth of useful knowledge in Europe during the British Industrial Revolution and analyze some of the cultural changes that drove them. This is not an issue on which either North or Acemoglu-Robinson spent much time. Somehow they assumed that if incentives were aligned, property rights were enforced, and a country was ruled by an inclusive government that was neither too anarchic nor too despotic, innovations would flow automatically.⁵ Such an assumption is rarely well-supported, much less documented.

What drove the institutional developments that took place before the Industrial Revolution that supported and encouraged the creation of useful knowledge? It is widely agreed that behind institutional change, there was cultural change, although scholars recognize that the coevolution of culture and institutions is subtle, complex, and bidirectional (Alesina & Giuliano, 2015). A simple causal connection between the two should be eschewed. To avoid any confusions, I will define precisely what I mean by culture and by institutions and adopt the definitions I used in Mokyr (2016, pp. 9–10). *Culture* is entirely of the mind: it includes preferences, beliefs, values, that are shared by reasonably large groups in society and passed on with modifications between generations. Because of those modifications, they are the subject of choice

⁴For extensive details on the British artisanal advantage, see Kelly et al. (2014, 2023, 2024) and Mokyr (2021).

⁵For a recent restatement in the same spirit, see Aghion et al., 2021.

as well as drift. *Institutions* are (following North's classic definition) the rules of the game (formal laws, customs, norms, and so on), and thus set the incentives (rewards and penalties) for behavior. They are set by society at large and hence are normally given parametrically to each individual much like prices are given in a competitive market. Below I will take a new look at the role that both culture and institutions played in the Industrial Revolution and hence in the Great Enrichment.

34.2 Cultural Drivers of Institutional Change: The Industrial Enlightenment

It is hard to think of the cultural and institutional changes in Europe in the eighteenth century without coming to grips with one of the most important cultural movements since the invention of monotheism: the Enlightenment. The Industrial Revolution started in Britain at a time when the Enlightenment was arguably at its peak, when the best minds in Europe were writing the works that are still regarded as the most influential sources of Enlightenment thought.⁶ The importance of the Enlightenment stands undiminished despite its many critics and enemies. To quote one author—no friend of the Enlightenment—“what is this event that is called the *Aufklärung* and that has determined, at least in part, what we are, what we think, and what we do today?” (Foucault, 1984, p. 21). More recently, Pagden (2013, p. 5) has opined that “no topic of historical debate...has exercised anything like the hold which the Enlightenment does over the ideological divisions of the modern world.”

What was the connection between the Enlightenment and the Industrial Revolution? It would seem a coincidence almost beyond belief that two events this momentous, occurring roughly at the same time and the same place, would be unrelated. Moreover, the early days of the Enlightenment predated the Industrial Revolution by many decades, so it seems unlikely to be concerned about reverse causation, i.e., that the Great Enrichment drove the Enlightenment. That said, the Enlightenment as a whole may be too large a tent and too broad an intellectual movement to be “causal” in this context. After all, for many years, historians convinced themselves that the Enlightenment was primarily a French affair and the Industrial Revolution a British one. Neither are true, but this view illustrates the difficulty in drawing a causal link here.

To sharpen the causal connection between the Enlightenment and the Industrial Revolution, I have proposed the concept of an *Industrial Enlightenment*, to denote the specific segment of the Enlightenment most relevant to the Industrial Revolution (Mokyr, 2002). It pertains to a body of thought concerned with applying knowledge of natural phenomena and regularities to production to achieve material progress. It thus excludes much of the philosophical writings of the most celebrated canonical

⁶Traditionally, economic historians date the Industrial Revolution between 1760 and 1830, though some of the most important breakthroughs, such as steam power, crucible steelmaking, and the use of coke in iron smelting, took place in the first half of the eighteenth century.

Enlightenment *philosophers* such as Hume, Kant, Montesquieu, and Condorcet. It is deeply connected to the belief in progress that came to life in Europe in the sixteenth and seventeenth centuries (Slack, 2015). Indeed, progress was at the very heart of the agenda of the Industrial Enlightenment. In the area of science and technology, perhaps more than anywhere else, did the majority of thinkers replace “fortune’s wheel” (a cyclical view of long-run history) with “time’s arrow”—a directional historical trend supporting meliorism (Outram, 1995, p. 124). Progress was at the heart of the Industrial Enlightenment.⁷

There is little evidence of a widespread belief in progress before the early modern age in Europe (Bury, 1955 [1920]; Terpstra, 2024). By 1700, however, the idea that material progress had taken place in the past and more of it could and should take place in the future had ripened in large parts of Europe, as much of the useful knowledge of the ancient world was being replaced in a wide range of fields. Most Enlightenment thinkers believed firmly in some kind of historical progress, although their exact notions of what it precisely consisted of differed. Very few thinkers at the time took part in the glum “counter-enlightenment” enunciated for example by Jean-Jacques Rousseau and the Königsberg philosopher Johann Georg Hamann (1730–1788) (Garrard, 2006).

The Industrial Enlightenment did more than just hoping for progress. It formulated a practical program as to how to bring it about. Astonishingly, much of this program was implemented as part of the two great political revolutions of the late eighteenth century and the more peaceful institutional reforms elsewhere. To be sure, unavoidable setbacks and misfirings occurred, but material progress continued apace in later centuries even if the term “the enlightenment” seems to have faded for describing what happened after 1815. The age of Enlightenment was perhaps not an enlightened age and much of its recommendations did not carry its full load of fruit until the nineteenth century. Yet continuing technological progress based on the accumulation of useful knowledge in the nineteenth century explains why the Industrial Revolution as a cluster of innovations did not fizzle out and became a sustained movement of economic growth.

The “Baconian program,” the ideology that formed the backbone of the Industrial Enlightenment, consisted of two components. First, research in fields such as physics, chemistry, and botany should *expand* the understanding of the material world and hence useful knowledge. As Boyle and other Baconians insisted, by directing research into natural phenomena that had been of practical interest for a long time, armed with better research equipment and scientific methods, nature could be harnessed to serve people better. The research agenda should thus be directed to areas where there was a high chance of solving practical technological problems—in medicine, manufacturing, shipbuilding, navigation, and so on. Secondly, the *access*

⁷Even a skeptical and hard-nosed intellectual such as Edward Gibbon could gush in the closing paragraph of Vol. IV of his *Rise and Decline* that we could “acquiesce in the pleasing conclusion that every age of the world has increased, and still increases, the real wealth, the happiness, the knowledge, and perhaps the virtue, of the human race” because the inestimable gifts of the arts and commerce could never be lost.

to this knowledge should be made as easy and cheap as possible, not only by dissemination but also by organizing and classifying what was known, so as to maximize the chance of it being applied in productive manner.

Hence, the practical program of the Industrial Enlightenment was founded on the firm belief that economic growth would be driven first and foremost by the growth and diffusion of useful knowledge.⁸ While most economists still saw Smithian growth as the main source of progress, it was becoming increasingly clear that technology was marching ahead at an ever-faster pace. In sum, the Industrial Enlightenment's agenda was about creating institutions that stimulated and incentivized the creation of new useful knowledge, and formed organizations that would help diffuse and disseminate existing knowledge by streamlining and cheapening access to it.

What role was played by what we call today "science"—essentially formal knowledge ("natural philosophy")? In England, it was practiced largely by well-educated intellectuals, most of them with a knack for practical questions. Enlightenment writers expressed more hope than certainty about the role of science in bringing about the progress that the *philosophes* were debating. There may be some truth to Outram's assessment that "The intellectual status of science was contested, its institutional organizations often weak and certainly thin on the ground, and the nature of its relations with the economy and with government often tenuous" (Outram, 1995, p. 48), but the Industrial Enlightenment was not just about the situation on the ground but about what the thinkers were hoping to bring about in the future.⁹

Historically, the exact role of science in the Industrial Revolution has been contested since the 1960s until today.¹⁰ Furthermore, the term "science," as Outram stresses, is something of an anachronism. The French term "science" and the German term "Wissenschaft" meant "knowledge" in general and not just and inquiry into natural phenomena and regularities. The only statement on which there seems to be a consensus is that the glass was partially full in 1750 and had gotten fuller (but far from wholly full) in 1850. Either way, natural philosophy in Britain tended to have a practical side. The vast bulk of important scientists of the

⁸The notion of the Enlightenment as a movement toward the practical improvement of material life through useful knowledge—the essence of modern economic growth—has found its expression in titles of studies such as such as Peter Jones (2016) and earlier in Porter (1982), Cunningham and French, eds. (2006), and Bertucci (2017).

⁹Joseph Priestley, one of the most articulate and distinguished English proponent of Baconian optimism, felt that "all things (and particularly whatever depends on science) have of late years been in a quicker progress toward perfection than ever. . . The wisdom of one generation will ever be the folly of the next" (1771, pp. 253, 265). In this rosy view, the ability of people to command and control nature would not only lead to them having a life that was more easy and comfortable but also mean that they would "grow daily more happy" (ibid., p. 6).

¹⁰Among the scholars strongly supportive of the importance of science and its impact on the Industrial Revolution are Musson and Robinson (1969), Margaret Jacob (1988, 1997, 1998, 2000), Wootton (2015), and Koschnick (2023). Among the early critics were Peter Mathias (1979) and Rupert Hall (1974). More recently, Ó Gráda (2016).

eighteenth century were concerned with questions that had the potential to have technological applications, none more so than John T. Desaguliers, Stephen Hales, Joseph Black, William Irvine, and Joseph Priestley. Some professional engineers were scientifically informed, such as John Smeaton or William Strutt. A few industrialists, similarly, had good scientific knowledge or kept close connections with scientists.¹¹ Furthermore, a market for scientific consultants developed in the eighteenth century.¹² Whether in the end the advice formally trained scientists dispensed was of much use or not, clearly quite a few industrialists thought so. Eighteenth century science still had not cracked many of the most pertinent questions needing answers and hence failed to have much impact on production techniques in use. Phenomena such as combustion, electricity, chemical reactions, the nature of steel, and much more were just being explored. Once science had made the hoped-for progress, Europe was ready for it and applications followed. The institutional structure inspired by the Industrial Enlightenment thus consisted of a host of organizations and networks, such as scientific societies, masonic lodges, libraries, and informal clubs, whose function was, among others, to disseminate useful knowledge from those who possessed it to those who believed they could use it (Koschnick, 2023). Dissemination of this type was the purpose of what Larry Stewart (1992) has called “public science.” For every successful attempt to use scientific insights and methods, there were many failures and disappointments. But progress was assured in the long term.

While at its base the Industrial Enlightenment was a cultural phenomenon reflecting a belief in material progress and ideas on how to bring it about, it had deep institutional roots. The “associational society” that Peter Clark (2000) has described so well included meetings of formal and informal networks, in which scientific and philosophical lectures and public experiments took place. Clark added (p. ix) that “If a British Enlightenment did exist, then one of its principal engines was the Georgian voluntary society. Fanning out across the English-speaking world, clubs and societies may have served as a vector for new ideas, new values, new kinds of social alignment, and forms of national, regional, and local identity.” As Roy Porter (1981, pp. 5–8) argued forcefully, the British Enlightenment—far more than elsewhere—was aimed at the solution of practical, material issues. England may not have produced the *Critique of Pure Reason*, he noted, “the

¹¹ For example, the Manchester cotton spinner George A. Lee, the owner of the first industrial mill to introduce gas lighting, was described by none other than Robert Owen as “one of the most scientific men of his age” (cited by Musson & Robinson, 1969, p. 99). Josiah Wedgwood, of course, was a Fellow of the Royal Society, an amateur geologist, and a close friend of both Priestley and Erasmus Darwin. Thomas Bentley, Wedgwood’s partner, was a genuine intellectual who spoke fluent French and Italian and was a founder and trustee of the celebrated Warrington dissenting academy.

¹² Among the most successful of these consultants were the clockmaker John Whitehurst, a member of the Lunar Society, and the Cornish applied mathematician Davies Giddy (Gilbert). Whether this activity, while lucrative, actually yielded on average any tangible net benefits to those who hired consultants is no more obvious in the eighteenth century than it is today. For more details, see Mokyr (2009a), pp. 57–59.

real intelligentsia was not chairbound but worked in the market place {...} English thought was concrete, practical {...} the Enlightenment in England was marked by its pragmatism.”¹³ The interests of British Enlightenment thinkers, unlike many of the great *philosophes* on the Continent, were not primarily in deep political and metaphysical issues, but in practical and pragmatic ones such as pistons, pumps, and pulleys.¹⁴

It was in these networks that much of the interaction between producers and natural philosophers occurred. For the complementarity between ideas and practice to occur, the practitioners on both sides had to connect. The famed “Lunar Society” and similar famed clubs such as the Manchester Literary and Philosophical Society and its successor in London, the London Chapter Coffee House were only the tip of the iceberg here: less regular (and less high-powered) meetings occurred in private salons, masonic lodges, and provincial societies (Stewart, 1992).

Did it work? Recent research has shown that these networks had a major impact on technological advances and industrialization (Dowey, 2016; Galofré-Vilà, 2023). When taken to the data, it turns out that the networks and other organizations designed to disseminate information are clearly capable of explaining the spatial pattern of technological progress as measured (imperfectly) by patents and prizes in industrial exhibitions. Moreover, the estimates that establish a causal connection between such forms of social capital and the Industrial Revolution may tend to underestimate the importance of the networks, because many of them were private and informal, taking the form of informal lectures and meetings in coffeehouses and taverns (Cowan, 2005).

Why did this matter?¹⁵ Historically, one of the sources of technological stagnation had been the social divide between those who knew things (*savants*) and those who made things (*fabricants*).¹⁶ The importance of this connection for technological progress was first pointed out by the sociologist Edward Zilsel (1942), who emphasized the importance of the relationship between artisans and intellectuals in early modern Europe for subsequent economic development. To construct pipelines or *passerelles* as Hilaire-Pérez (2000) has called them, through which those two groups could communicate, was at the core of the Industrial Enlightenment

¹³ Elsewhere he added (Porter, 2000, p. 15) that “British pragmatism was more than mere worldliness: it embodied a philosophy of expediency, a dedication to the art, science and duty of living well in the here and now.”

¹⁴ One may wonder what the roots of this difference were. One possible explanation is that in the late seventeenth century, Britain created a form of government that may not have been democratic, but all the same was widely regarded as functional and legitimate as the central power was Parliament and not the monarchy. Unlike French *philosophes*, most British intellectuals did not strongly object to their regime as a whole even if they often took issue with many of its decisions and policies.

¹⁵ Some of the following is adapted from Mokyr (2005).

¹⁶ In one of his most perceptive essays, Needham (1969, p. 27) pointed out that in Imperial China, real work in engineering was “always done by illiterate or semi-literate artisans and master craftsmen who could never rise across that sharp gap which separated them from the ‘white collar literati.’”

movement. As noted above, the communications ran in both directions: practical people with specific technical problems to solve could access and absorb whatever best-practice knowledge had to offer—which, of course, at most times was rather little. At the same time, the needs of crafts and manufactures to find solutions to specific problems could influence the research agenda of the scientists. The famous close connections between James Watt and his Scottish academic friends John Robison and Joseph Black are a prime example. So was the growing demand for scientifically informed consultants by progressive industrialists (for more details, see Mokyr, 2009a, pp. 57–59).

The Industrial Enlightenment, interestingly enough, was to a considerable extent a provincial affair. London, of course, remained the cultural center of the country where a large number of skilled artisans catered to the local market. But in provincial towns such as Manchester, Liverpool, Newcastle, and Leeds, we find, as J.H. Plumb (1972, p. 23) memorably put it, “knots of enlightened men with a passionate regard for empirical knowledge, secular in their intellectual attitudes, although often muddled, uncertain and tentative, with . . . rational and irrational beliefs combined in the same man.” The networks that made the Industrial Enlightenment so effective can be described as small world networks (a network where the typical number of hops between two randomly chosen nodes is small in some sense, e.g., grows proportionally to the logarithm of the number of nodes in the network). This means that people who mattered were mostly connected in little localized cliques with a few connections between these nodes. Between 1750 and 1850, the number of these internode connections rose as communications improved, but even in its earlier times, they were effective enough.¹⁷ Better roads and coastal ships, better carriages, and improved postal services were at the core of the ever more effective connections between local nodes.

Such a bilateral interchange between formal science and technology seems natural to us in the twenty-first century. In eighteenth-century Europe, however, such interactions developed slowly, impeded by social inertia and snobbery. Yet in the longer run, they proved as irresistible because their role in promoting the continuing progress of technology was recognized by producers. Precisely because the British Enlightenment was more focused on material issues, it was supportive of the growth of communications between natural philosophers and the more progressive and open-minded industrialists of the time (Porter, 1981, 2000).

For information networks to be efficient, trust was essential.¹⁸ Members of associations and scientific societies exchanging knowledge had to trust one another,

¹⁷During the eighteenth century, British mail services improved significantly. In their original form, almost all letters sent within England had to go through London. The idea of cutting out London (which seems obvious enough to us) occurred to Ralph Allen (1693–1764), a Cornwall postal employee and one of the unsung successful entrepreneurs of the first half of the eighteenth century. By the time of Allen’s death in 1764, most of England and Wales received mail daily through his “byway post” (Headrick, 2000, p. 187).

¹⁸For a detailed discussion of the role of trust and social capital in the Industrial Revolution, see Sunderland (2007).

but often knowledge came from strangers. It was therefore essential that agents had the cultural propensity to trust their sources and assume a high probability that the information they accessed was correct. In that regard, it is important to stress that a long tradition of universalistic values, in which people felt that unrelated persons and even strangers were trustworthy, went back to the Middle Ages and the disappearance of extended kinship as the chief form of social organization. A substantial modern literature has found evidence that societies that are less founded on extended kinship tend to be more trusting and more able to cooperate with strangers (Henrich, 2020; Enke, 2019, 2023; Schulz, 2022). As argued in detail in Greif et al. (2024), the prevalence of more universalistic values and trustful social norms was a likely consequence of the European Marriage Pattern that was established in the Middle Ages. The weakening of the extended family made the kind of organizations dedicated to the transfer of knowledge among strangers more prevalent and more effective.¹⁹ In that sense, the Industrial Enlightenment had an unexpected and certainly unintended ancestry.

These networks, then, reduced access costs. The lowering of access costs, to be sure, was driven in part by technological shocks—above all the printing press and better postal services—but it had a critical institutional component as well. The rules of the scientific community and open science increased trust, as investigators shown to be charlatans or plagiarists at least in theory suffered costly damage to their reputations. Access costs did not only matter because of the ability of innovators to access the relevant scientific theories and experimental data that learned people assembled. New knowledge builds upon existing knowledge, tweaking, and recombining it into new forms, but in order to do so, it needed to know what was already there. The intellectuals who subscribed to the ideas of the Industrial Enlightenment recognized this need and created the means to access knowledge through a range of publications that ordered and organized existing knowledge in an accessible way.

Such publications needed two things to be effective. One was an ordering of the contents in a handy way, such as alphabetization. The other was to make it accessible through libraries or cheap editions. The books published through the inspiration of the Industrial Enlightenment were part of one might call the codification project. It consisted of the description and depiction of what had hitherto been mostly tacit knowledge, the kind of knowledge possessed by craftsmen and mechanics, passed on to their apprentices but less easily to others. These objectives were achieved most famously in the *Grande Encyclopédie*, widely regarded as one of the paradigmatic documents of the Enlightenment. Robertson (2021, p. 419) notes that many artisanal

¹⁹ As argued at length in Greif et al. (2024, Chap. 9), many of the social organizations that emerged in the Middle Ages indirectly and through often winding paths, led to the kind of organizations that drove the Industrial Enlightenment. An example is the Republic of Letters, a loose network of intellectuals, who were not related, but cooperated through a common objective to expand useful knowledge. Indirectly, such organizations were made possible through the weakening of extended families and the rise of corporations and associations such as guilds, universities, scientific societies, and the like in medieval times.

skills previously guarded from outsiders were now placed in the public realm. d'Alembert and Diderot's masterpiece was a "landmark in the diffusion and demystification of knowledge." What he fails to add was that in addition to the encyclopedia, assorted compendia and technical dictionaries "of all arts and sciences" were published all over Europe, even if they less ambitious and encompassing than d'Alembert and Diderot's grand design (Headrick, 2000, pp. 142–172). In what is now a classic paper, Squicciarini and Voigtläender (2015) showed how access to the encyclopedia was correlated with technological progress and the rise of modern industries.

Another way of organizing useful knowledge in an accessible way was to publish specialized books describing the state of the art techniques in various industries. The most impressive of these projects was no doubt the over 13,000 pages of the 80 volumes of the *Descriptions des Arts et Métiers* compiled by the French academy before the Revolution—in Gillispie's (1980, p. 344) judgment, the largest body of technological literature ever produced. Many other specialized books on individual industries appeared, especially in France. While Britain had a comparative advantage in tacit knowledge and the "knacks" of the trade, acquired through informal and personal channels such as apprenticeship and personal knowledge, the European Continent seems to have had a comparative advantage in formal and codified knowledge. The most impressive and best-organized textbook of pre-Lavoisier chemistry was P.J. Macquer's encyclopedic *Dictionnaire de Chimie* published in 1766 and translated into English in 1771 by the chemist James Keir. The same was true, surprisingly, for technological knowledge. The most detailed and richest technical book on coal and collieries was written by a French physician, Jean-François C. Morand (2012 [1769–78]). While England had the most extensive copper-smelting industry in eighteenth century Europe, economic historians depend on continental sources for its description (Harris, 1992, p. 22). The English engineered and produced; the French wrote about it. Both nations had been part and parcel of the Industrial Enlightenment, and both realized that knowledge was the key to economic progress; but there were nuances in how they were going to carry out this program.

Finally, access to useful knowledge was furthered by the British patent system, discussed in more detail below. As Dutton (1984, p. 22) remarked, full and detailed disclosure of the technical details of an invention was the price that society charged for patent protection. Once these details were deposited in the office of chancery, they were accessible, albeit at a cost. Bottomley (2014, pp. 177–201) has shown that arguments that inventors submitted deliberately misleading specifications and that the documents were hard and costly to access are overblown. Many details were published in cheap magazines such as *The Mechanics Magazine* published between 1823 and 1858. He provides numerous examples of various individuals consulting the documents that otherwise would have been almost impossible to access. In short, intellectual property rights in new technological knowledge meant that there was a sharp trade-off between patents and secrecy, and the many inventors who made use of the patent office contributed to the lowering of access costs. That, too, was part and parcel of the Industrial Enlightenment.

34.3 High-Powered and Low-Powered Incentives²⁰

As has been long understood, the creation of new knowledge is underincentivized. Knowledge is non-rivalrous, often non-excludable (or costly to exclude), and creates major externalities (since knowledge builds on previous knowledge). As a result, it is difficult to design any incentives—either high-powered or low-powered—that adequately encourage and incentivize inventors and scientists. While a few inventors gained fame even before 1600 (one thinks of Hero of Alexandria, Gutenberg, and William Lee), it is hard to think of many examples of pre-1750 inventors who became rich thanks to their ingenuity even if it contributed materially to economic performance.²¹ On the eve of the Industrial Revolution, after centuries of gradual institutional change, something of a bifurcation in the way contributions to useful knowledge had emerged. Propositional knowledge was rewarded largely through reputation effects and other low-powered incentives; prescriptive knowledge could be rewarded directly through patents or other financial rewards.

The European Enlightenment—and especially British Enlightenment—believed in the sanctity of property rights. Of course, almost all societies had some concept of property rights. However, the idea of property rights in ideas and knowledge—IPR's—was a uniquely Western idea; no such notion can be found for instance in Chinese culture. By the late eighteenth century, European writers realized not only that patents were a moral institution—rightly rewarding a successful inventor and recognizing a moral claim on what was deemed to be property by priority—but that it actually stimulated and encouraged innovation and thus economic development.²² Modern economics has found the idea attractive, as it provided an easy and intuitively powerful connection between institutions and the technological progress that drove modern economic growth (North, 1981, pp. 164–165; Aghion et al., 2021, p. 37). England, after all, had a patent system that predated the Industrial Revolution by over a 100 years. And as this institutional foundation safely predated the Industrial Revolution, it effectively rules out reverse causation. Moreover, as Adam Smith also believed in a famous passage in his *Lectures on Jurisprudence* (1978 [1757], p. 83), it rewarded inventors more or less in proportion to the contribution of their invention and thus created a high-powered incentive to engage in innovative activity. Or so it seemed.

²⁰ Some of the following is adapted from Mokyr (2009b).

²¹ It is telling that nothing is known about the identity of some of the other striking inventions of the Middle Ages such as the mechanical weight-driven clock, eyeglasses, fire arms, and the introduction of windmills.

²² Goethe famously wrote that the “clever Englishman” was able to turn ideas into property and sighed that “no wonder they are in every way ahead of us” (cited by Klemm, 1964, p. 173). In 1795, Chief Justice James Eyre of the court of common pleas observed that many patents existed for new methods of manufactures that were “beyond all calculation important to commercial activity . . . and in my apprehension it is strictly agreeable to the spirit and meaning of the Statute” (cited by Bottomley, 2014, p. 152).

The problem is that a closer look at the British patent system leads to serious reservations regarding its effect on the Industrial Revolution. The net effect of patents on technological change remains in serious dispute till the present day (Boldrin & Levine, 2008), all the more so for the eighteenth century. In a few cases, patents have been argued to have actually blocked innovation by the threat of litigation. Perhaps more damaging, the number of actual patents filed before 1750 was small and took off with the Industrial Revolution.²³ It is hard to assign it a major causal role. Moreover, patents constituted a small fraction (about one in six) of all important inventions. This was demonstrated in devastating detail by Moser (2005, 2021), and while one can question the exact percentages involved, anecdotal evidence strongly supports her findings. For a few inventors, the system worked and they were able to secure a temporary monopoly on a lucrative technological breakthrough. But for many of the most important inventions of the era, the patent system was either not utilized at all or failed to protect the inventor.²⁴ Until 1851, patenting in Britain was expensive beyond the reach of most middle-class inventors coming from the artisanal class, and legal protection from infringements through the courts was weak and unreliable.²⁵ Additional incentives stimulating technological progress had to come from elsewhere.

In some ways, the market system found other ways to reward technological pioneers, above all through the “first-mover advantages” as was famously the case with Richard Arkwright. In other cases, they tried to secure a monopoly by keeping a crucial ingredient secret—if this was feasible. In other cases, the political system stepped in directly, with Parliament voting substantial pensions or awards for people whose innovations had been deemed especially useful (most famously the large awards, totaling £ 30,000, awarded to Edward Jenner, the inventor of the smallpox vaccination). Prestigious prizes to successful scientists whose work was likely to have practical consequences were established. The Copley Medal, established by the Royal Society in 1736, was a powerful reputation-signaling device, and many winners were artisans and engineers. Among them were some of the best inventors of the Industrial Revolution, such as the leading engineer John Smeaton and the

²³ Around 1768, inventors began to use specialized patent agents, which streamlined the clumsy and slow application process and clearly was a factor in the sudden take-off of the number of patents after 1760 or so (Bottomley, 2014, pp. 65–73).

²⁴ Moreover, if it was the patent system that accounted for Britain’s leadership, one might wonder why other economies were not more successful earlier. Britain’s patent system and its policies encouraging invention were not unique: the Netherlands did have a system of patents awarded by both the Estates General and the provincial estates from the late 1580s on, awarding a peak number of 119 patents in the 1620s. Yet in the eighteenth century, the number of patents dwindled into insignificance (Davids, 2008, pp. 400–416). Ancien régime France, too, had a system of rewarding inventors even if it worked quite differently from the English patent system (Hilaire-Pérez, 2000).

²⁵ Because of their dogmatic opposition to monopolies of any kind, judges often decided against patentees. Consequently, patentees who felt their patents were infringed rarely sued: between 1770 and 1850, only 257 patent cases came before the courts out of 11,962 patents granted (Dutton, 1984, p. 71). Bottomley (2014, pp. 79–85) has revised Dutton’s numbers, but even so, less than 2.5% of cases were litigated before 1830.

instrument maker Jesse Ramsden (Bektas & Crosland, 1992). The Rumford medal for Majors discoveries in heat and light (the first of which was awarded to Rumford himself in 1800) was a similar token of respect to innovators. Membership in the Royal Society remained a coveted distinction even if the Society showed some signs of intellectual atrophy in the eighteenth century.²⁶ In a few cases, inventors could shield their monopoly through secrecy, though it seems unlikely that many techniques and formulas could be protected for long from imitation and reverse engineering.

Inventors were driven by other motives as well. Many of the great engineers wanted to become known as ingenious and creative individuals so as to secure commissions from manufacturers and local authorities based on reputations. Some inventors who had been trained in formal schools had absorbed the culture of open science and felt that it was immoral to exclude others from knowledge, which was after all non-rivalrous.²⁷ Clearly, of course, not all of them were pure altruists: many if not most cared deeply about their reputations among their peers, about membership in the Royal Society, medals, knighthoods, and other reputation-signaling devices.²⁸ Institutions, then, encouraged the growth of useful knowledge, but the incentive structure driving innovation was more complex than the simple-minded greed driving efforts that noneconomists attribute to the beliefs of economists (for more details, see Mokyr, 2018).

The same holds for the patent system. Human behavior and psychology played just as much a role as income-maximization. Entrepreneurship and invention—and a fortiori patenting in England during the Industrial Revolution—are like unfair gambles or lotteries. What makes such lotteries work is that a few spectacular and

²⁶In a few cases, specific prizes were promised for inventors who would solve concrete problems that were regarded as especially important, such as the problem of longitude at sea. While the standard tale of the British Board of Longitude, which was formally in charge of administering the funds, has been radically revised in modern times (Howes, 2022), the basic fact remains that two well-known scientists, William Whiston and Humphrey Ditton, petitioned Parliament to establish the Board, realizing that exerting efforts to solve required both a reward and financial support during the (rather tedious and lengthy) research effort.

²⁷Michael Faraday only took out one patent in his life and one that had nothing to do with his main line of research, namely a better chimney for the use of lighthouses (which he made over to his brother). His mentor, Humphry Davy, refused to take out a patent for the famous “miner’s friend” lamp he invented and claimed repeatedly to be “philosophically” opposed to patents (Ruston, 2019).

²⁸Economists have long argued that especially in the upper tail of the human capital distribution—the one that matters most the economic development—agents are driven by more than material incentives. As pointed out by Kreps (1997), the distinctions between intrinsic motivation and status incentives can be fairly tricky. Kreps notes that in most employment situations (and that covers the bulk of the citizens of the Republic of Letters), it is hard to detect intrinsic motivation. He notes that “what is called intrinsic motivation may be (at least in part) the worker’s response to fuzzy extrinsic motivators, such as fear of discharge, censure by fellow employees, or even the desire for coworkers’ esteem.” Here “coworkers’ esteem” should be interpreted as “the opinion of other scientists and intellectuals.” Reputations, then, led to memberships, appointments, knighthoods, and were of course also desirable for their own sake.

well-publicized successes led others to believe that their odds were better than they really were. Its very existence drove many would-be inventors toward exerting efforts in the hope of securing the kind of riches that James Watt and Charles Tennant had gained from the patent system. This is all the more powerful precisely, because by definition, all inventions are in some qualitative dimension different from one another (unlike lottery tickets) and a function of the ability of the inventor. It is therefore easier for “gamblers” on the patent system to persuade themselves that the conditional odds that underlie decisions in this activity may be systematically higher than the unconditional ones. In other words, would-be inventors overestimated their abilities and thus their chances for a successful patent as Adam Smith famously pointed out (Smith, 1976 [1776], p. 120). The expected rate of return on inventive activity for the *entire* population of inventors was in all likelihood negative, but because of a salient upper tail, the *ex ante* incentives were powerful (Nye, 1991). And they were what mattered. All the same—and this is an important point if we are to understand the Industrial Revolution—for society as a whole, the benefits were substantial, since the social returns of inventive activity exceeded the private return by a large factor. Hence anything that can inflate the *ex ante* perception of private returns to get it a bit closer to the social returns will encourage invention and is thus socially beneficial. It is this mechanism that best supports Bottomley’s belief that the patent system in Britain “did indeed encourage the development and diffusion of technology during the Industrial Revolution” (Bottomley, 2014, p. 174).

All in all, the Industrial Enlightenment became permeated with the notion that new prescriptive knowledge should be seen as “property” and that inventors “owned” the rights to it even though they were required to divulge the details of their invention. Over time, thus, the eighteenth century turned the concept of a “patent” from a monopoly and a privilege (and thus a form of rent-seeking) into a property right that increased efficiency, reflecting the ideas of the Industrial Enlightenment.²⁹ The idea of a patent was perfectly consonant with the Industrial Enlightenment—a patent had to be potentially useful for practical purposes to be granted at all but would only become financially valuable if it met the market test.

Following their respective revolutions, both the United States and Republican France established patent offices within a year from one another in 1790–1791. By securing this property, it was believed that technological creativity was incentivized and hence was encouraged.³⁰ In short, it was a mechanism supporting material progress. As John Stuart Mill (1929 [1848], p. 933) summed it up, the patent system clearly dominated any reward system run by the state. All the same, it was widely understood to be a second-best solution. Enlightenment writers overwhelmingly opposed monopolies, and a patent was a monopoly. There was an intuitive sense

²⁹ Biagioli (2006, p. 1136) has argued that it was that “it was the introduction of patent specifications that made that contract [between patentee and the public] politically defensible by distancing it from the ‘odious monopolies’ of the ancien régime.”

³⁰ James Madison wrote that “The utility of the clause will scarcely be questioned. . . The right to useful inventions seems with equal reason to belong to the inventors. The public good fully coincides in both cases with the claims of the individuals.” (Madison, 1788).

that knowledge should be free-access because anything that limited access to useful knowledge was obviously in contradiction with the other objective of easy access. It was also realized that patents were used strategically, for example, to block research by non-patentees in some directions and thus actually slowed down innovation (as the classic examples of Thomas Savery's patent blocking Thomas Newcomen from patenting his steam engine and James Watt blocking high pressure engines attest).³¹

The Society of Arts, one of the most paradigmatic organizations reflecting the values of the Industrial Enlightenment founded in 1754, awarded medals and small prizes to inventors—provided the invention had not been patented (Howes, 2020).³² Over time, the Society gave a preference to projects that for one reason or another could not be patented or made profitable. It emphasized inventions that saved lives, made workplaces safer, prevented traffic accidents, replace child-labor in chimney sweeping and the like (Howes, 2020, pp. 73–81). Moreover, through its meetings and its library and other facilities, it became a clearinghouse for useful knowledge, a repository for information on technology (*ibid.*, p. 84). Much like the patent system discussed above, it managed to combine in one organization both incentives and diffusion of useful knowledge. It embodied the essence of the Industrial Enlightenment.

Incentivizing and rewarding propositional knowledge was more difficult as it could not be patented. This kind of knowledge shares many of the features of the commons (Ostrom & Hess, 2007). Much like a commons, knowledge is a shared resource but one that is produced in the system and has a positive opportunity cost. Unless people somehow can be charged for its use, at first blush no one has any financial incentive to produce it. The putative “tragedy of the commons” suggests that any system of knowledge production would flounder on the non-excludability of knowledge. And yet, just as the commons do not always become a tragedy but can be managed by a community, knowledge production can be and was properly incentivized if the correct institutions emerged. The cultural background to the emergence of these institutions was the Industrial Enlightenment.

³¹ Charles Babbage, never one to mince words, denounced the patent law as a system that deprived the inventor of the fruits of his genius and put the most productive citizens of society in a position of “legalized banditti.” It was a “fraudulent lottery which gives its blanks to genius and its prizes to knaves” (Babbage, 1830, pp. 333, 321). The objections were not so much against the system in general as much as against the way the law was written and carried out in Britain, especially the high cost “system of vicious and fraudulent legislation” which deprived the inventor of the fruits of his genius and that even the granting of a patent was “almost wholly illusory” till the patent had been sustained by a court of law, at an even higher cost (*ibid.*, p. 334).

³² William Shipley, its founder, viewed its purpose as follows: “Whereas the Riches, Honour, Strength and Prosperity of a Nation depend in a great Measure on Knowledge and Improvement of useful Arts, Manufactures, Etc. . . . several [persons], being fully sensible that due Encouragements and Rewards are greatly conducive to excite a Spirit of Emulation and Industry have resolved to form [the Society of Arts] for such Productions, Inventions or Improvements as shall tend to the employing of the Poor and the Increase of Trade” (cited by Mokyr, 2009, p. 52).

To see how this came about, note that the Industrial Enlightenment was in large part the product and logical continuation of the Republic of Letters, the earliest signals of which can be traced to the late fifteenth century (Mokyr, 2016). The Republic of Letters was an information network that served to diffuse knowledge among intellectuals, and in the process, it created reputations of the best and the brightest among learned men and women. These reputations were of course valuable in and of themselves for intellectuals craving peer recognition and respect, but they were also strongly correlated with coveted royal and aristocratic patronage, which meant both enhanced social status and economic security (and in a few cases political influence).

Patronage and fame were of course not the sole driver of the growth of propositional knowledge in the era of the Industrial Revolution. Science had become respectable, virtuous, something that rich burghers or gentlemen such as Anthonie van Leeuwenhoek, Henry Cavendish, and Antoine Lavoisier kept themselves busy with.³³ Others, however, still depended on reputations to win gainful employment, especially when universities declined them positions.³⁴ In short, the concept of “open science” proposed by Dasgupta and David (1994; see also David, 2008), in which priority rights (that is, the credit given to the person who claimed to be the first to have discovered a natural phenomenon or regularity) operated much like property rights in physical assets, flourished in the eighteenth century and operated in harmony with the Industrial Enlightenment. Superstar intellectuals, inventors, and scientists achieved what we would call today celebrity status, sometimes deserved, sometimes not. But the enormous reputation enjoyed by Isaac Newton could not but stimulate countless ambitious and talented mathematicians and physicists “to be like Ike.” A century after Newton, Edward Jenner’s fame (and subsequent financial reward) must have done the same for medical researchers.

Incentives to scientific research came in many flavors. One of the great insights was that it turned out to be fairly cheap to incentivize innovative intellectuals to add to propositional knowledge. International fame and ambitious and peer recognition were and still are major drivers of intellectual excellence and all that was needed was a few inexpensive signals that broadcast eminence. The idea was an old one: Marcus Aurelius endowed some chairs in philosophy in the second century AD (Frede, 2013). The practice was revived in early modern Europe, with the emergence of the Republic of Letters. A key figure in this regard was Lady Margaret Beaufort, the mother of the first Tudor king Henry VII, who established the Lady Margaret’s

³³ Cavendish seems to have been remarkably agnostic about his reputation as a scientist. The discoverer of hydrogen, and in some views “without peer in eighteenth century English natural philosophy. . . the first after Newton to possess mathematical and experimental talents at all comparable to Newton’s” (McCormach, 2008, p. 159), he published a number of things, but many of his most important findings remained unpublished and were discovered only many years after his death.

³⁴ An example is the career of Leonard Euler, refused a professorship by the University of Basel, who found a patronage position at the Russian Imperial Academy and later at the Berlin Academy at the invitation of King Frederick II.

professorship of divinity at the University of Cambridge in 1502 and a lectureship in divinity at Oxford. A few years earlier, the first Regius Professorship had been endowed in Aberdeen, and Henry VIII established the first ones in England among others in Law, Greek, and Medicine. The Lucasian Chair in Mathematics at Cambridge was the first privately endowed chair, per Henry Lucas's will, in 1663. Isaac Newton was the second holder of the chair and held it for 33 years. It is still considered to be one of the most prestigious academic chairs in the world. Membership in the Royal Society remained a desirable honor, with "FRS" becoming a badge of accomplishment. Such distinctions became more prevalent in the eighteenth century as part of the program to encourage the progress of useful knowledge by rewarding persons who were regarded as having added substantially to useful knowledge such as the abovementioned Copley Medal. Other European nations followed similar practices (Hilaire-Pérez, 2000).

Furthermore, the interest in promoting useful knowledge supported the idea of natural philosophers and great engineers as role models and high social status people, whom young and ambitious persons would want to imitate. The full acceptance of this value took a long time to reach maturity, but at least two giants in British history, Newton and Watt, managed to achieve exactly that (Fara, 2002; MacLeod, 2007). On the Continent, things were moving along the same lines. In early eighteenth-century France, the new science was especially valued and became part of high society and a new political culture in which a powerful alliance was created between the *savants* of the Republic of Letters and the royal administration (Shank, 2008, p. 88).³⁵ The Dutch physician Herman Boerhaave was so famous in his time that Peter the Great sat in on his classes and Voltaire and Linnaeus traveled to Leyden to meet him.³⁶ The great Pierre-Simon Laplace himself was elected to the French *Académie* at age 24 and eventually was made a count of the Empire in 1806 and a marquis in 1817, as well as appointed to a rather unsuccessful stint as Minister of Interior affairs in 1799.³⁷ The effective allocation of talent and human capital in the extreme upper tail of the distribution of talent was sensitive to such signals. It was those kinds of institutions that provided the incentives for modern science to evolve eventually into the fuel that powered the engines of progress.

One driver of scientific progress that is underestimated is sheer curiosity. Whether curiosity is more or less hard-wired in humans is hard to know, but clearly culture has

³⁵ Shank explains that royal administrators saw academics as a model and a catalyst for bringing about a wider transformation of French society, but the social boundary between savants and royal officeholders was often quite blurry.

³⁶ A possibly apocryphal story has it that a letter sent from China by a Mandarin, addressed simply to "the illustrious doctor Boerhaave, physician in Europe," reached him without delay.

³⁷ One way of raising the social status of science and scientists was to appoint them to high political positions—something that Napoleon, in many ways an adherent of the Industrial Enlightenment, excelled in. In addition to Laplace, he appointed the chemists Antoine-François de Fourcroy as François Minister of Public Instruction and Jean-Antoine Chaptal as Minister of the Interior, as well as making the mathematician Gaspard Monge a count in 1808 and appointing him to a key position in the Senate.

the capability to encourage or discourage it.³⁸ In early medieval Christianity, curiosity was broadly condemned as unbecoming to a good Christian, and regarded as an attempt to peek into God's domain. St Bernard of Clairvaux (1090–1153), an influential monastic reformer, promoted curiosity to one of the seven deadly sins as a subspecies of “lust” (Daston, 1995, p. 393). Two centuries later, however, Thomas Aquinas made some fine distinctions between the kinds of knowledge that were virtuous to pursue and those that were not.³⁹ The aversion to new knowledge for its own sake was thus already weakening in the later Middle Ages, but it was still powerful and can still be found in Jean Gerson and Erasmus (Robertson, 2021, p. 44; Rummel & MacPhail, 2021).

In any case, early modern Europe experienced a cultural transformation that turned curiosity into a virtue rather than a sin, especially for the elites. Renaissance courts and academies proudly paraded their curiosity and saw the display of new knowledge as a symbol of the superiority and power of the ruling classes. This new attitude was particularly manifest in the proto-museums known as “curiosity cabinets,” which displayed exotic animal and plant specimens and antiques (Eamon, 1994, pp. 223–224).⁴⁰ The intellectual father of the Industrial Enlightenment, Francis Bacon, insolently warned his readers in *The Great Instauration* not to fall into the error of thinking “that the inquisition of nature is in any part interdicted or forbidden” and cited with approval Proverbs 25:2 that stated that “it is the honor of God to conceal a thing and the honor of kings to investigate them” (my translation from the Hebrew) (Bacon, 1999 [1620], pp. 74–75). While Bacon still warned that the “true end of knowledge” should not be “for the pleasure of the mind,” research of any kind was turned into a virtuous activity.⁴¹ Clearly he realized that research was fun—an intrinsic motivation indeed.) Renaissance Europe experienced the emergence of a class of “virtuosi”—upper-class gentleman devoted to study of arts and sciences. While much of their work was amateurish and dilettantish, their activities

³⁸ Some of the following is adapted from Mokyr (2018).

³⁹ Aquinas (1947, questions 166–167) viewed the intrinsically motivated search for knowledge favorably within limits. As long as the philosophers do not seek truth “to assail the faith” and not study “to know the truth above the capacity of his own intelligence, since by so doing men easily fall into error,” he was willing to allow a certain level of curiosity. He distinguished between “studiositas” and “curiositas,” the former being a virtue, the latter a vice under certain circumstances.

⁴⁰ One of the most striking of these collection was the one established in Rome by the polymath Athanasius Kircher, a German-born Jesuit (1602–1680) whose museum of curiosities was the most famous in Europe and contained a range of archaeological, technological, and scientific objects. Kircher was a person of almost infinite curiosity, “a one-man intellectual clearing house for cultural and scientific information” as the Encyclopedia Britannica described him. His 44 books covered a bewildering range: geography, astronomy, mathematics, language, medicine, ancient history, and music.

⁴¹ Thomas Hobbes defined curiosity as the desire to know the causes of things and noted that “such as is in no living creature but Man; so that Man is distinguished, not onely by his Reason; but also by this singular Passion from other Animals; in whom the appetite of food, and other pleasures of Sense, by praedominance, take away the care of knowing causes. . . is a Lust of the mind, that by a perseverance of delight in the continuall and indefatigable generation of Knowledge, exceedeth the short vehemence of any carnall Pleasure” (Hobbes (1929 [1651], p. 44).

raised the prestige of intellectual discourse and they provided a constituency and readership for the people who did the heavy lifting in research. The virtuosi provided much-needed social respectability to those who contemplated engaging in intellectual endeavors and they helped turned curiosity from vice to virtue.

By the early years of the Industrial Enlightenment, the moral acceptability of research driven by curiosity had become relatively uncontroversial and the social status of great scientists kept rising.⁴² Many noted scientists and inventors came from the very top of a highly hierarchical society. Robert Boyle was one of the wealthiest men in England, and Antoine Lavoisier a wealthy tax official in France. The abovementioned reclusive Henry Cavendish (the grandson of the Duke of Devonshire) and the Scottish experimentalist Archibald Cochrane, the Earl of Dundonald (who worked on a chemical process to make coal tar and ruined himself financially in the process), were equally driven by curiosity. Ferdinando de Medici II, grand duke of Florence and Galileo's patron (1610–1670), is credited with the invention of a substantially improved thermometer in the mid-seventeenth century. Whereas financial gain, social status, and economic security were important incentives in the creation of scientific knowledge, intrinsic motivation and curiosity cannot be discounted. Some of the great scientists in the Scientific Revolution and research in the age of Enlightenment were driven by it.

Finally, it is worth reiterating that incentives for expanding knowledge can be both positive (a reward) or negative (penalizing people for heretical ideas, disrespectful of sacrosanct ancient wisdom). The well-known tales of Galileo and Copernicus illustrate the risks that scholars proposing radical new ideas were taking. Less well-known but perhaps more damaging was the persecution of Jan Baptist van Helmont, a Brussels-born chemist and the founder of pneumatic chemistry, who was condemned for “perverting nature by ascribing to it all magic and diabolic art” and interrogated by the inquisition, as was Giambattista della Porta, an Italian polymath and inventor, whose Neapolitan scientific society known as the *Accademia dei Segreti* founded in 1560 was disbanded on orders of the Inquisition. We will never know with precision how many aspiring young scientists were discouraged or frightened into abandoning their careers on account of these persecutions. But it has been shown that the counter-reformation seriously damaged scientific work, especially south of the Alps as well as the number of scientists in countries strongly affected by the actions of the Inquisition (Cabello, 2023).

The Enlightenment movement as a whole was profoundly committed to tolerance, freedom of speech and association, and had nothing but scorn for censorship. Led by such luminaries as Locke and Voltaire, Enlightenment intellectuals struggled to be rid of any kind of ideological persecution or censorship of new ideas, and this included useful knowledge. Controversial writers such as La Mettrie and Helvétius could still get in trouble for provocative ideas (though rarely scientific discoveries), and in Britain, where censorship had been abolished, political rebels and excessively

⁴² As Lorraine Daston phrases it, “moralists continued to thunder away against such frivolous and potentially dangerous interests. . . but the decibel level of their complaints suggests that by the late seventeenth century they were on the defensive” (Daston, 2005, p. 37).

disrespectful authors could be sued for obscene or seditious libel. Such legal suppression of thought that strayed too far from what the authorities deemed acceptable was still fairly common in Enlightenment Europe, but legal sanctions against scientists and physicians—no matter how outrageous their arguments were thought to be—are hard to find. The Industrial Enlightenment was part of a greater movement toward toleration and freedom of expression that rid Europe of a serious obstacle to progress and by and large eliminated negative incentives. The first amendment to the US constitution was a concrete example of the influence that these intellectuals had on eighteenth century institutions. Again we can see how culture affected a set of specific institutions that turned out eventually to be crucial to the economic development of Western Europe.

34.4 Conclusions

Institutions, in the sense of setting incentives and creating both opportunities and the rewards for taking advantage of them, have become a major theme in economic history. Yet oddly enough, with some exceptions, they have not been systematically explored in trying to understand why the Industrial Revolution occurred at all and why it took place first in Britain.⁴³ What is needed above all is an explanation of the impact of institutions on innovation and technological creativity in an eighteenth-century context. In other words, to understand the Industrial Revolution, we need to lay bare why innovation occurred at all, as most societies that existed before 1750 were far less innovative than eighteenth-century Britain. An examination of the various institutional aspects shows, first and foremost, that political institutions and the other institutions that determined the distribution of power and authority in Britain were important, but so were private-order institutions. What is now needed above all are measurement and testing of the impact of institutions as much as possible; in this debate, an ounce of data is worth more than a pound of theory.⁴⁴

The deeper question, and one that the neo-institutional literature has struggled to answer, is why some economies have institutions that are more conducive to technological progress than others and therefore become technological leaders, if only for limited periods. Are institutions themselves endogenous to economic and material factors as Marxian orthodoxy maintained? Or are they wholly determined by what people value and believe, as recently argued by McCloskey (2021b)? I have made a similar argument at considerable length in Mokyr (2016).⁴⁵ It seems

⁴³For earlier surveys, see Kapás (2012) and Mokyr (2008, 2021).

⁴⁴A pioneering attempt to test some of the alternative hypotheses is Dowey (2016) and Galofré-Vilà (2023); see also Kelly et al. (2023).

⁴⁵There is a somewhat pedantic debate on whether the word “culture” is an appropriate term. McCloskey (2021b) feels that “‘culture’ is merely the vague way in which economists talk when they have not actually taken on board the exact and gigantic literature about ideas, myths, stories, rhetoric, ideology, metaphors, ceremonies, and the like since the Greeks, the Talmudists, or the Sanskrit grammarians.” Basically, however, the standard definitions of culture as the “unions of the sets of shared beliefs and values in a society” as proposed by Boyd and Richerson (1985, p. 2) in their pioneering work on cultural evolution seems perfectly adequate for the purpose at hand.

reasonable to argue that rather than one “driving” the other, they coevolve, affecting one another but also both affected by exogenous shocks and events. History, however, is rich with examples in which institutions were not in any way chosen by societies but imposed through raw political and military power.

The exact ways and channels in which culture affects institutions is a topic that still awaits full incorporation in economic history and will be a project large enough to keep an entire generation of scholars busy (Alesina & Giuliano, 2015). But proof of the notion that any *single* cultural factor is either necessary or sufficient to create the kinds of institutions needed for technological progress seems so far elusive. Moreover, the dynamics of institutional change are still poorly understood. In many cases, institutions were created for one reason and led to unintended consequences. Neither the Poor Law of 1601 nor the Patent Office of 1624 was created by people with an Industrial Revolution in mind. In other cases, fairly small events led to contingent institutional outcomes with far-reaching implications for innovation and economic progress.

Moreover, discrepancies can persist between what people want and like and the kind of institutions they actually experience. After all, political revolutions can occur quite abruptly, and it seems implausible that culture, which is quite sticky and resistant to rapid change, would change *pari passu*. For instance, was Russian culture completely turned upside down between 1985 and 1991 with the demise of Marxism-Leninism? Or did Iranian culture and devoutness change dramatically between 1978 and 1980 when a theocratic Islamic state took over? Closer to the topic at hand here, McCloskey (2010, 2021a) argued that an ideology of liberalism and individual freedom led to “a liberal releasing of opportunity” and was the key to growth. In fact, however, eighteenth-century Britain was hardly the liberal (in the classical sense) *laissez faire* society it was to become in the second quarter of the nineteenth century. It was a highly regulated and protectionist state in which rent-seeking and mercantilist ideas constrained and distorted many markets. Moreover, non-liberal institutions can, under the right kind of government, generate innovation and technological progress, even if in the very long run they are perhaps less sustainable and welfare-enhancing as free-market pluralist societies.

Furthermore, anyone arguing for the primacy of culture over institutions owes us—no mean requirement—some kind of theory of where ideas come from and why they become prevalent, that is, why they triumph in the market place for ideas. To complicate matters further: institutions in turn can successfully bring about cultural change by successful propaganda and steering education and social conditioning in one direction or another.⁴⁶ In the absence of a more general theory explaining the coevolution of culture and institutions, we have to be satisfied with the more modest project of finding how cultural phenomena such as the Industrial Enlightenment led to the institutions that helped bring about the Great Enrichment.

⁴⁶Of the many examples that illustrate this point, a striking one is the success of Nazi educational policies to install antisemitic ideas in German youngsters, a culture that survived into the postwar period. See Voigtländer and Voth (2015).

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., & Robinson, J. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Crown.
- Acemoglu, D., Johnson, S., & James, R. (2005). Institutions as a fundamental cause of economic growth. In P. Aghion & S. Durlauf (Eds.), *The handbook of economic growth* (pp. 385–465). Elsevier.
- Aghion, P., Antonin, C., & Bunel, S. (2021). *The power of creative destruction: Economic upheaval and the wealth of nations*. Harvard University Press.
- Alesina, A., & Giuliano, P. (2015). Culture and institutions. *Journal of Economic Literature*, 53(4), 898–944.
- Aquinas, T. (1947). *The summa theologica* (Benziger Bros. edition). Translated by Fathers of the English Dominican Province. <http://dhspriority.org/thomas/summa/SS.html>
- Babbage, C. (1830). Reflections on the decline of science in Britain and on some of its causes. *Quarterly Review (London)*, 43, 307–342, repr. Augustus Kelley, 1970.
- Bacon, F.. (1999 [1620]). *The great instauration* (Reprinted in *Selected Philosophical Works*, R.-M. Sargent, Ed.). Hackett Publishing.
- Bektas, M. Y., & Crosland, M. (1992). The Copley Medal: The establishment of a reward system in the royal society, 1731–1839. *Notes and Records of the Royal Society of London*, 46(1), 43–76.
- Bertucci, P. (2017). *Artisanal enlightenment: Science and the mechanical arts in old regime France*. Yale University Press.
- Biagioli, M. (2006). Patent republic: Representing inventions, constructing rights and authors. *Social Research*, 73(4), 1129–1172.
- Blackstone, W. (1765–69). *Commentaries on the laws of England* (1st ed.). Printed at the Clarendon Press. <http://www.yale.edu/lawweb/avalon/blackstone/blacksto.htm>
- Bogart, D. (2011). Did the glorious revolution contribute to the transport revolution? Evidence from investment in roads and rivers. *The Economic History Review*, 64(4), 1073–1112.
- Bogart, D. (2019). Investing in early public works: Financial risks and returns in English and Welsh turnpikes, 1820–82. *Economic History Review*, 72(3), 848–868.
- Boldrin, M., & Levine, D. K. (2008). *Against intellectual monopoly*. Cambridge University Press.
- Bottomley, S. (2014). *The British patent system during the industrial revolution, 1700–1852*. Cambridge University Press.
- Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process*. University of Chicago Press.
- Bury, J. B. (1955 [1920]). *The idea of progress: An inquiry into its growth and origin*. Dover.
- Cabello, M. (2023). *The counter-reformation, science, and long-term growth: A black legend?* (Unpublished Working Paper). Martin Luther University.
- Christie, J. R. R. (2018). Chemical Glasgow and its entrepreneurs. In L. L. Roberts & S. Verrett (Eds.), *Compound histories: Materials, governance and production, 1760–1840* (pp. 311–332). Brill.
- Clark, P. (2000). *British clubs and societies, 1580–1800: The origins of an associational world*. Clarendon Press.
- Clark, G. (2007). *A farewell to alms*. Princeton University Press.
- Clow, A., & Clow, N. L. (1952). *The chemical revolution: A contribution to social technology*. Batchworth, repr. Gordon and Breach, 1992.
- Cowan, B. (2005). *The social life of coffee: The emergence of the British coffeehouse*. Yale University Press.

- Cunningham, A., & French, R. (Eds.). (2006). *The medical enlightenment of the eighteenth century* (Revised edition). Cambridge University Press.
- Dasgupta, P., & David, P. A. (1994). Toward a new economics of science. *Research Policy*, 23(5), 487–521.
- Daston, L. (2005). All Curls and Pearls: a review of Neil Kenny. The uses of curiosity in early Modern France and Germany.” *London Review of Books*, 23, 37–8.
- Daston, L. (1995). Curiosity in early modern science. *Word and Image*, 11(4), 391–404.
- David, P. A. (2008). The historical origins of ‘open science’: An essay on patronage, reputation and common agency contracting in the scientific revolution. *Capitalism and Society*, 3(2), 1–103.
- Davids, K. (2008). *The rise and decline of Dutch technological leadership*. Brill, 2 Vols.
- Dowey, J. (2016). *Mind of matter: Access to knowledge in the industrial revolution* [Unpublished Ph.D. dissertation, London School of Economics].
- Dutton, H. I. (1984). *The patent system and inventive activity during the industrial revolution 1750–1852*. Manchester University Press.
- Eamon, W. (1994). *Science and the secrets of nature*. Princeton University Press.
- Enke, B. (2019). Kinship, cooperation and the evolution of moral systems. *Quarterly Journal of Economics*, 134(2), 953–1019.
- Enke, B. (2023). *Moral boundaries* (NBER Working Paper No. 31701). NBER.
- Fara, P. (2002). *Newton: The making of a genius*. Columbia University Press.
- Foucault, M. (1984). What is enlightenment? In P. Rabinow (Ed.), *The Foucault Reader*. Penguin.
- Frede, D. (2013). Alexander of Aphrodisias. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. <http://plato.stanford.edu/archives/sum2013/entries/alexander-aphrodisias/>
- Galofré-Vilà, G. (2023). The diffusion of knowledge during the British industrial revolution. *Social Science History*, 47, 167–188.
- Garrard, G. (2006). *Counter-enlightenments: From the eighteenth century to the present*. Routledge.
- Gillispie, C. C. (1980). *Science and polity in France at the end of the old regime*. Princeton University Press.
- Greif, A., & Iyigun, M. (2013). Social organizations, violence, and modern growth. *American Economic Review (Papers and Proceedings)*, 103(3), 534–538.
- Greif, A., Iyigun, M., & Sasson, D. (2013). *Social institutions and economic growth: Why England rather than China became the first modern economy* (Stanford Center for International Development, Working Paper 465). Stanford Center for International Development.
- Greif, A., Mokyr, J., & Tabellini, G. (2024). *Two paths to the twentieth century: Culture and institutions in Europe and China, 1000–2000*. Princeton University Press.
- Hall, A. R. (1974). What did the industrial revolution in Britain owe to science? In N. McKendrick (Ed.), *Historical perspectives: Studies in English thought and society* (pp. 129–151). Europa Publications.
- Hanlon, W. W. (2022). *The rise of the engineer: Inventing the professional inventor during the industrial revolution* (NBER Working Paper No. 29751). NBER.
- Harris, J. R. (1992). Skills, coal and British industry in the eighteenth century. In J. R. Harris (Ed.), *Essays in industry and technology in the eighteenth century*. Ashgate/Variorum.
- Headrick, D. (2000). *When information came of age: Technologies of knowledge in the age of reason and revolution, 1700–1850*. Oxford University Press.
- Henrich, J. (2020). *The weirdest people in the world: How the west became psychologically peculiar and particularly prosperous*. Farrar Strauss and Giroux.
- Hilaire-Pérez, L. (2000). *L’Invention technique au siècle des lumières*. Albin Michel.
- Hobbes, T. (1929 [1651]). *Leviathan*. Clarendon Press.
- Howes, A. (2020). *Arts and minds: How the royal society of arts changed a nation*. Princeton University Press.
- Howes, A. (2022). Why innovation prizes fail. <https://worksinprogress.co/issue/why-innovation-prizes-fail/>
- Jacob, M. C. (1988). *The cultural meaning of the scientific revolution*. Alfred A. Knopf.

- Jacob, M. C. (1997). *Scientific culture and the making of the industrial west* (2nd ed.). Oxford University Press.
- Jacob, M. C. (1998). The cultural foundations of early industrialization. In M. Berg & K. Bruland (Eds.), *Technological revolutions in Europe* (pp. 67–85). Edward Elgar.
- Jacob, M. C. (2000). Commerce, industry, and the laws of Newtonian science: Weber revisited and revised. *Canadian Journal of History*, 35(2), 275–292.
- Johnson, Samuel. 1759. The prince of Abisinia. : Printed for R. and J. Dodsley; and W. Johnston.
- Jones, P. M. (2016). *Agricultural enlightenment: Knowledge, technology, and nature*. Oxford University Press.
- Kapás, J. (2012). Which institutions caused the British industrial revolution? In J. Kapás & P. Czeglédi (Eds.), *Institutions and the industrial revolution* (Competitio Books, Vol. 12, pp. 35–59). University of Debrecen.
- Kelly, M., & Ó Gráda, C. (2021). Connecting the scientific and industrial revolutions: The role of practical mathematics. *Journal of Economic History*, 81, 239–275.
- Kelly, M., Mokyr, J., & Ó Gráda, C. (2014). Precocious Albion: A new interpretation of the British industrial revolution. *Annual Review of Economics*, 6, 363–391.
- Kelly, M., Mokyr, J., & Ó Gráda, C. (2023). The mechanics of the industrial revolution. *Journal of Political Economy*, 131(1), 59–94.
- Kelly, M., Mokyr, J., & Ó Gráda, C.. (2024). *Why Britain? A new view of the industrial revolution*. Unpublished manuscript.
- Klemm, F. (1964). *A history of western technology*. MIT Press.
- Koschnick, J. (2023). *On the shoulders of science – Early science as a driver of innovation during the early industrial revolution* [Unpublished Ph.D. dissertation, London School of Economics].
- Kreps, D. M. (1997). Intrinsic motivation and extrinsic incentives. *American Economic Review*, 87(2), 359–364.
- MacLeod, C. (2007). *Heroes of invention: Technology, liberalism and British identity*. Cambridge University Press.
- Madison, J. (1788). The Federalist Papers No. 43. <https://guides.loc.gov/federalist-papers/text-41-50#s-lg-box-wrapper-25493407>
- Mathias, P. (1979). *The transformation of England*. Columbia University Press.
- McCloskey, D. N. (2010). *Bourgeois dignity: Why economics can't explain the modern world*. University of Chicago Press.
- McCloskey, D. N. (2016). Max U vs humanomics: A critique of neo-institutionalism. *Journal of Institutional Economics*, 12(1), 1–27.
- McCloskey, D. N. (2021a). *How growth happens*. Presented to the Cato Institute Conference, Unpublished.
- McCloskey, D. N. (2021b). *Institutions matter; but not as much as neo-institutionalists believe*. Presented to the Economic History Association in Tucson, Arizona, October.
- McCormach, R. (2008). Cavendish, Henry. In C. C. Gillispie (Ed.), *Complete dictionary of scientific biography* (Vol. 3, pp. 155–159). Charles Scribner's Sons.
- Mill, J. S.. (1929 [1848]). *Principles of political economy* (W. J. Ashley, Ed.). Longmans, Green and Co.
- Mokyr, J. (2002). *The gifts of Athena: Historical origins of the knowledge economy*. Princeton University Press.
- Mokyr, J. (2005). The intellectual origins of modern economic growth. [Presidential address]. *Journal of Economic History*, 65(2), 285–351.
- Mokyr, J. (2008). The institutional origins of the industrial revolution. In E. Helpman (Ed.), *Institutions and economic performance* (pp. 64–119). Harvard University Press.
- Mokyr, J. (2009a). *The enlightened economy*. Yale University Press.
- Mokyr, J. (2009b). Intellectual property rights, the industrial revolution, and the beginnings of modern economic growth. *American Economic Review*, 99(2), Papers and Proceedings, May 2009, 349–355. Repr. in Correa, C. M. (Ed.), *Intellectual property and economic development*. Edward Elgar.
- Mokyr, J. (2016). *A culture of growth: Origins of the modern economy*. Princeton University Press.

- Mokyr, J. (2018). Bottom up or top-down? The origins of the industrial revolution. *Journal of Institutional Economics*, 14(6), 1003–1024.
- Mokyr, J. (2019). The economics of apprenticeship. In M. Prak & P. Wallis (Eds.), *Apprenticeship in early modern Europe* (pp. 20–43). Cambridge University Press.
- Mokyr, J. (2021). ‘The holy land of industrialism’ – Rethinking the industrial revolution. *Journal of the British Academy*, 9, 223–247. <https://www.thebritishacademy.ac.uk/documents/3393/JBA-9-p223-Mokyr.pdf>
- Mokyr, J., Sarid, A., & van der Beek, K. (2021). The wheels of change: Technology adoption, millwrights, and the persistence in Britain’s industrialization. *Economic Journal*, 132, 1894.
- Morand, J.-F.-C.. (2012 [1769–78]). *L’art d’exploiter les Mines de Charbon de Terre. Du Charbon de Terre et de ses Mines* (Repr. ed.). Hachette Livres.
- Moser, P. (2005). How do patent laws influence innovation? Evidence from nineteenth-century world fairs. *AER*, 95, 1214–1236.
- Moser, P. (2021). *Pirates and patents: An economic history of innovation in Europe and the United States*. Unpublished Book Manuscript.
- Musson, A. E., & Robinson, E. (1969). *Science and technology in the industrial revolution*. Manchester University Press.
- Needham, J. (1969). *The grand titration*. University of Toronto Press.
- North, D. C. (1981). *Structure and change in economic history*. Norton.
- North, D. C., & Weingast, B. (1989). Constitutions and commitment: Evolution of institutions governing public choice in seventeenth century England. *Journal of Economic History*, 49(4), 803–832.
- Nye, J. V. (1991). Lucky fools and cautious businessmen: On entrepreneurship and the measurement of entrepreneurial failure. In J. Mokyr (Ed.), *The vital one: Essays in Honor of Jonathan R.T. Hughes* (Research in Economic History, Vol. 6, pp. 131–52). JAI Press
- Ó Gráda, C. (2016). Did science cause the industrial revolution? *Journal of Economic Literature*, 5(1), 224–239.
- Ostrom, E., & Hess, C. (2007). A framework for analyzing the knowledge commons. In C. Hess & E. Ostrom (Eds.), *Understanding knowledge as a commons* (pp. 41–81). MIT Press.
- Outram, B. (1995). *The enlightenment*. Cambridge University Press.
- Pagden, A. (2013). *The enlightenment and why it still matters*. Oxford University Press.
- Plumb, J. H. (1972). Reason and unreason in the eighteenth century. In *In the light of history*. Delta Books.
- Porter, R. (1981). The enlightenment in England. In R. Porter & M. Teich (Eds.), *The enlightenment in national context* (pp. 1–18). Cambridge University Press.
- Porter, R. (1982). Was there a medical enlightenment? *British Journal for Eighteenth Century Studies*, 5, 49–63.
- Porter, R. (2000). *The creation of the modern world: The untold story of the British enlightenment*. W.W. Norton.
- Priestley, J. (1771). *An essay on the first principles of government* (2nd ed.). J. Johnson.
- Robertson, R. (2021). *The enlightenment: The Pursuit of happiness, 1680–1790*. Harper and Row.
- Rodrik, D., Subramanian, A., & Trebbi, F. (2004). Institutions rule: The primacy of institutions over geography and integration in economic development. *Journal of Economic Growth*, 9, 131–165.
- Rummel, E., & MacPhail, E. (2021). Desiderius Erasmus. In *The Stanford encyclopedia of philosophy*. <https://plato.stanford.edu/archives/win2021/entries/erasmus/>
- Ruston, S. (2019). Humphry Davy: Analogy, priority, and the ‘true philosopher’. *Ambix*, 66(2–3), 121–139.
- Schulz, J. F. (2022). Kin networks and institutional development. *The Economic Journal*, 132(647), 2578–2613.
- Shank, J. B. (2008). *The newton wars and the beginnings of the French enlightenment*. University of Chicago Press.
- Slack, P. (2015). *The invention of improvement: Information and material progress in seventeenth-century England*. Oxford University Press.

- Smith, A. (1976 [1776]). *The wealth of nations* (E. Cannan, Ed.). University of Chicago Press.
- Smith, A. (1978 [1757]). *Lectures on jurisprudence* (R. Meek et al., Eds.). Clarendon Press.
- Solar, P. M. (1995). Poor relief and English Economic development before the industrial revolution. *The Economic History Review, New Series*, 48(1), 1–22.
- Squicciarini, M. P., & Voigtländer, N. (2015). Human capital and industrialization: Evidence from the age of enlightenment. *Quarterly Journal of Economics*, 130(4), 1825–1883.
- Stewart, L. (1992). *The rise of public science*. Cambridge University Press.
- Sunderland, D. (2007). *Social capital, trust and the industrial revolution, 1780–1880*. Routledge.
- Szostak, R. (1991). *The role of transportation in the industrial revolution*. McGill's-Queen's University Press.
- Terpstra, T. (2024). *Rome and the rise of the west*. Unpublished manuscript.
- Voigtländer, N., & Voth, H.-J. (2015). Nazi indoctrination and anti-Semitic beliefs in Germany. *PNAS*, 112(26), 7931–7936. <https://doi.org/10.1073/pnas.1414822112>
- Wallis, P. (2025). *The market for skill: Apprenticeship in England, 1500–1800*. Princeton University Press.
- Wootton, D. (2015). *The invention of science: A new history of the scientific revolution*. Allen Lane.
- Zilsel, E. (1942). The sociological roots of science. *American Journal of Sociology*, 47(4), 544–560.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Formal and Informal Institutions Explaining China's Modern Economic Development: Their Historical Roots and Recent Evolution 35

Mary M. Shirley  and Lixin Colin Xu

Contents

35.1	Introduction	928
35.2	What Type of Institutions Are Needed to Encourage Private Investment and Growth?	928
35.3	The Evolving Roles of Government and Business in China's Long History	929
35.4	State and Private Sectors in Post Mao China	933
35.4.1	Early Reforms (1978–1992)	933
35.4.2	Increasing Support to Private Activity (1992–2007)	934
35.4.3	Laws and Private Growth (2007–2012)	935
35.4.4	Reversals and Uncertainty (2012–Present)	936
35.5	Institutions Fostering Private Sector Development in China	938
35.5.1	Enforcement of Contracts	939
35.5.2	Protection Against Expropriation	942
35.6	What Does the Future Hold?	945
	References	948

Abstract

How has China grown so rapidly through private initiatives despite weak institutions enforcing contracts and protecting against expropriation? Why has private investment, after overcoming so many obstacles, stagnated over the last decade? This chapter explores the changing roles of the state and private sectors during China's history and since the death of Mao in 1979. It analyzes the formal and informal institutions that have enforced contracts and protected against expropriation in China. It concludes with a consideration of the adverse effects of recent changes on China's private sector and the future trajectory of China's growth.

M. M. Shirley (✉)
The Ronald Coase Institute, Philomont, VA, USA
e-mail: mshirley@coase.org

L. C. Xu
Cheung Kong Graduate School of Business, Beijing, China

Keywords

Property rights · Society capacity · Contract enforcement · Government expropriation · Confucianism · China's state-owned enterprises

35.1 Introduction

China's extraordinary growth, averaging 10% yearly in real terms since 1976, has been driven largely by the private sector (Brandt et al., 2020). Private firms surged after they were first grudgingly accepted in the 1980s, and subsequently mushroomed to account for 70% of China's GDP by 2012 (Lardy, 2019, p. 19). It is puzzling, as Qian (2002) points out, that China managed to grow so rapidly through private initiatives despite the weakness of institutions traditionally viewed as necessary, such as rule of law and secure property rights. Also puzzling is why private investment, after overcoming so many obstacles, has stagnated over the last decade. The role of government and the role of informal institutions (including culture) in supporting private initiatives are crucial to understanding the evolution of China's private enterprises, both during its rapid growth and its recent stagnation.

This chapter first describes the two main functions of institutions fundamental for private development: contractual enforcement and protection against expropriation by government and elites. Next it highlights the changing roles of the state and private sectors in China's history and now, and then analyzes the alternative institutions that have fostered private investment in China over the last five plus decades. It concludes with a consideration of the adverse effects of recent changes on China's private sector and the future trajectory of China's growth.

35.2 What Type of Institutions Are Needed to Encourage Private Investment and Growth?

Acemoglu and Johnson (2005), following North (1981), identified two types of institutions critical for prosperity: (1) those supporting private contracts ("contracting institutions") and (2) those protecting citizens against expropriation by government and elites ("anti-expropriation institutions"). The contracting institutions that help firms and individuals formulate and enforce contracts consist of formal laws and informal constraints, including business norms, reputational concerns, future business relations, etc. (Greif, 2006; Klein et al., 1978; Williamson, 1985, 2002).

Anti-expropriation institutions include the formal laws and informal norms that constrain governments and powerful people to behave within the law. To protect property rights and maintain order, the state must be sufficiently powerful to enforce the law and prevent anarchy. But how to assure that a competent and powerful state does not arbitrarily infringe on others' rights in favor of allied elites (Shirley, 2008; Weingast, 1997)? NIE's answer is rule of law—no one is above the law and the law

is impersonal, treating all the same—combined with mechanisms to hold government and elites accountable for conforming to the law, such as an independent judiciary, competitive elections, and unconstrained media. (Acemoglu & Robinson, 2012, 2019; North, 1990; North et al., 2009; Shirley, 2008; Weingast, 1997). A mobilized and informed civil society able to form organizations such as political parties, unions, business associations, or lobbying groups to advance their interests, can curb elite confiscation and hold state actors accountable. Among the institutions that empower civic activism and state accountability are legal guarantees of freedom of the press, expression and right to assemble; elections (Acemoglu & Robinson, 2012, 2019; Keefer, 2011; North et al., 2009), federalism (Weingast, 1997); and separation of powers (Carey, 2005; Haggard & McCubbins, 2001). Historical traditions can also empower civic activism. For instance, evidence from China suggests that shared experience in major wars under strong local leadership can promote a tradition of civic engagement (Lai et al., 2023).¹

While there are overlaps between the two types of institutions, there are also important differences. Acemoglu and Johnson (2005) argue that when contracting institutions are weak, citizens can find ways to protect themselves against opportunistic behavior by changing the terms of their contracts or the nature of their activities, but it is not possible for citizens to protect themselves from a predatory government or powerful elites by writing alternative contracts (p. 951). This may explain why they also find that contracting institutions have less effect on long-run growth, investment, and financial development, compared to anti-expropriation institutions. Yet China has managed to grow rapidly even though improvement in anti-expropriation institutions has “remained elusive” (Long et al., 2022, p. 3). One reason for this is improvement in contracting institutions (Ibid.) and another is the development of successful alternatives to anti-expropriation institutions. To understand how, we next consider the evolving roles of the government and private sectors in China's long history and the more recent past, and then describe China's alternative institutions fostering private development.

35.3 The Evolving Roles of Government and Business in China's Long History

China has a long and strong tradition of state dominance of a weak society (one whose members cannot credibly constrain the state) for reasons outlined by Acemoglu and Robinson (2019) and Zhao (2022). First is the weakness of the aristocracy since the centralization of China. During the Transitional Era before the first unification of China (BC 545 to BC 420), a strong hereditary aristocracy challenged and divided the powerful Jin (晉) State. To avoid future such challenges,

¹Similarly, when county-level governments under China's top-down appointment system had members with local guerrilla experience, they facilitated local development and reduced the damage caused by central planning disasters such as the 1958–60 Great Famine (Fang et al., 2023).

the newly founded states abolished the hereditary aristocracy and began to rely on professional bureaucrats. Since then, the aristocracy had rarely been strong enough to challenge the emperors. The professional bureaucrats who replaced them were loyal to the emperors who had selected them and seldom constituted a serious challenge to the state.

Second, since centralization during the Qin Dynasty more than two millennia ago, wealthy private merchants and entrepreneurs never had sufficient power or cohesion to challenge the state (Zhao, 2022; von Glahn, 2018). Before Qin was founded in BC 221, merchants had relatively higher status. During the Qin unification, however, state policies geared toward winning the war against other states promoted farming and soldiering exclusively; merchants were heavily discriminated against. Subsequent Chinese states continued to discriminate against merchants. In contrast, in many European states, merchants gradually amassed sufficient financial and military power by the seventeenth and eighteenth centuries to challenge the royal families. Consider the Glorious Revolution of 1688 in England when manufacturers, traders, and the “newly dynamic cities and towns” were able to form the Whig political party that successfully imposed their vision on the state (Pincus & Robinson, 2014, p. 15). Chinese merchants, in contrast, seldom held sufficient power to counter the state. Moreover, key manufacturing industries such as salt, metallurgy, and weapon-making were largely state-owned and controlled (von Glahn, 2018, pp. 82, 93–94, 185). Reflecting their lower status, Chinese merchants and entrepreneurs long relied on government connections, including family members serving as government officials, to protect their property rights (Zhao, 2022, p. 387).

Relatedly, China never had influential philosophies legitimizing the power of merchants and entrepreneurs. In contrast to increasingly influential Western portrayals of merchants as reflecting the spirit of capitalism of Max Weber or wielding Adam Smith’s “invisible hand” to benefit society, ruling Confucian thought legitimized the rule of emperors and their bureaucrats as holding “mandates of heaven.” The selection of bureaucrats via merit-based exams further legitimized their prominent position in Chinese society, while merchants and entrepreneurs were often denigrated.

Third, the shared beliefs and cultural traditions in China further strengthened China’s authoritarian tendencies (Zhao, 2022).² Of the three major spiritual/belief traditions, Taoism, Buddhism, and Confucianism, two were hands-off towards government, and the third actively bolstered government authority.³ The founders

²The source for the rest of this section is from Zhao (2022) unless other sources are explicitly cited.

³Another fundamental governance system was termed “legalism,” which was a distinct school of thought from Confucianism. Legalism initiated the strong central state tradition during the Warring States period (BC 419 to BC 221) (see Chap. 7, Zhao, 2022). The legalists emphasized the absolute authority of the state (and emperor) built on a strong bureaucracy and severe laws and punishments to maintain order. While legalism quickly strengthened the state and unified the warring states, its draconian rules also resulted in the shortest reign of any dynasty in Chinese history. Confucian reforms in subsequent centuries adopted some elements of legalism, and Confucianism can be thought broadly to include legalism elements.

of Taoism, such as Lao Zi (老子) and Zhuang Zi (庄子), argued that the best way to organize a society was letting people govern themselves without government interference. Emperors and their advisors rarely supported Taoism, since it did not grant government authorities legitimacy and significance. Although Taoism's influence on government policies was muted over the millennium, it did engender a norm among the population that "ordinary folks" should not be involved in government affairs, but should instead focus on self-development. Furthermore, the strong starting position of the central state dating from the Qin Dynasty, as well as the continuous dominance of the emperor-Confucian-scholars coalition prevented Taoism from being pushed to the center stage of Chinese ideology. Consequently, Taoism has not contributed to constraining the government.

The second belief tradition, Buddhism, has also rarely constrained government authority in China. Buddhism was introduced to China in the fourth century and spread quickly. Buddhism did not give special respect to government authority and its goal of achieving personal enlightenment by reducing material desires did not add to the wealth of families or the government, which alienated most Chinese rulers. The majority of the rulers spurned this tradition and Buddhism never became significant in constraining the government.

The third set of beliefs, Confucianism, was largely co-opted to work for most rulers throughout Chinese history.

After the collapse of the Qin Dynasty, an influential philosopher, Dong Zhongshu (董仲舒, 179-104 BC), mixed elements of the ruling craft of legalism, the world view of Taoism, and most critically, the moral and behavior codes of Confucianism. The Confucian elements included the premises that: (i) people should follow the orders of heaven, with emperors being the conduit ("mandate of heaven"); (ii) maintaining social order requires accepting the hierarchical relationship giving authority to an emperor over his ministers (subordinates), a father over his son(s), and a husband over his wife. Emperors and their bureaucrats embodied overwhelming legitimacy from their positions. The main constraint on emperors was the "mandate of heaven"—bad behavior of emperors would be punished by heaven—and the idea that emperors were supposed to be the model for all men and women. The highest responsibility for emperors was to encourage their subjects to follow Confucian moral codes.

The Confucian moral codes of Dong became an enduring ruling philosophy for future emperors. Emperors embraced the codes since they gave them the highest authority with few constraints (except "heaven"). In subsequent millennia, the selection of bureaucrats was institutionalized through the imperial examination, or Keju system, based on their familiarity with Confucian classics and loyalty to emperors and the Confucian moral codes. Keju reached its apex during the Song dynasties (960-1279), which happened to be the most prosperous in Chinese history especially in comparison to the rest of the world (Broadberry et al., 2018). Another key development during Song was the introduction of private academies run by well-educated Confucian scholars that trained elites who became involved in local affairs, such as founding local clan organizations or providing disaster relief and local basic education. Thus, local elites originated from an educational system, not private

wealth, as in England, for example. Scholars were encouraged to cultivate the self (to be trustworthy, for example), improve the family, and govern the country; their top ambition was to serve in the government. As a result of the Keju system, and more generally the Confucian culture of educating children to understand the classics and uphold the Confucian moral codes (Zhu, 2021, p. 156), China's well-established educational infrastructure resulted in relatively high levels of basic education throughout most of Chinese history. Until the nineteenth century, China's levels of basic numeracy outperformed all countries for which there are data (including Japan, Korea, Britain, Germany, France, Sweden, India, Egypt, and Russia) (Huang, 2023, pp. 108–109). Another positive legacy of Confucianism stemmed from its emphasis on trust, as we discuss below.

Bureaucrats, benefiting from the Keju system, were co-opted to embrace Confucian norms. The strong emphasis on loyalty and vast differences in wealth and power between emperors and bureaucrats left little chance for collective actions by bureaucrats. The emperors/bureaucrats complex became a highly stable ruling structure as a result: two millennia of Chinese states were largely under the guidance of the Confucian-Legalist philosophy, despite some attempts at innovation.⁴

China's long tradition of super-strong state/weak society casts a shadow over modern China, and was likely an important factor in the way that state and private sectors have evolved over recent decades.⁵ First, the strong state/weak society tradition helps explain why China would have difficulties constraining government expropriation, but find it easier to protect property rights in transactions among non-government actors. Encouraging transactions among non-state actors increases society's wealth without threatening the dominance of the state. Second, the robust presence of state-owned enterprises (SOEs) and the weak status of merchants and entrepreneurs throughout its history supports the premise that China's private sector development would encounter fierce resistance from the state sector. Third, China's lengthy history as a centralized state with a Confucian culture emphasizing serving the country suggests that incentives based on national identity could be a useful tool for the government. Government's historically tight control of its people and overwhelming power advantage also suggest that it would be tougher to change the balance of power between government and civil society in China than in many other countries where the government does not have such an extensive tradition of absolute power. Finally, the Confucian tradition of emphasizing education as well as the well-established local educational infrastructure would put China in a better

⁴For example, during late sixteenth and early seventeenth centuries, scholars such as Wang Yangming (王阳明), Zhu Xi (朱熹), and Gu Yanwu (顾炎武) introduced modern concepts, such as viewing oneself as the source of knowledge, viewing the state as belonging to both rulers and subjects, constraining rulers' authority, increasing local authority, and promoting commerce. Without the support of the bureaucrats-scholars, these innovations lost influence quickly (Zhao, 2022).

⁵Why didn't this constrain countries with a similar heritage, such as Japan? Huang (2023, p. 328) argues that Meiji Japan confronted Westernization with systems inherited from a period of Chinese history (the Tang dynasty) that was "ideologically diverse and flexible" rather than the "sclerotic autocratic systems of the Song, Yuan, Ming and Qing."

position to take off economically more supportive institutions and policies were introduced.,.

35.4 State and Private Sectors in Post Mao China

35.4.1 Early Reforms (1978–1992)

Under Mao, China was dominated by state and communal property rights. After Mao's death in 1976 and especially after the reforms launched in 1978, the scope for private initiatives began to expand. Private farming, which had been banned until 1980, was officially recognized in 1982 under the Household Responsibility System (HRS) and spread from about 1% of rural households in 1979 to 99% by 1984 (Huang, 2008, pp. 86–7).⁶ Private farming, however, remained a use right, not an ownership right; the land itself continued to be owned by the state and could be transferred to other uses by local government fiat.⁷ Nevertheless, rights under the HRS were sufficiently secure to facilitate labor market integration and the movement of labor out of agriculture (Giles & Mu, 2018), which substantially improved rural productivity and standard of living (Lin, 1992).

As for non-farm private businesses, initially private firms with more than 7 employees were illegal and by the end of 1982 some 30,000 people had been arrested for “profiteering” (Coase & Wang, 2012, p. 78). During this period, the government viewed private firms as mostly self-employment for marginal economic actors, such as unemployed, laid-off, or retired workers (Nee & Oppen, 2012, p. 5). Government leaders did not foresee the surge of private entrepreneurship unleashed when the Mao era restrictions were reduced. Although the business environment was unfavorable, it was a major improvement from the repression under Mao, and many would-be entrepreneurs found private business more appealing than the drudgery of employment in a state-owned enterprise (SOE). Even though the laws then did not offer protection of property rights, would-be entrepreneurs judged the security of private investment by signals, such as *People's Daily* writing about raising cows rather than class struggle, Deng Xiaoping's reemergence out of multiple purges, and the government return of confiscated assets to capitalists who were released from prison (Huang, 2023, p. 127). Much of this dynamism was initially channeled into township and village enterprises (TVEs), which grew rapidly to almost 25 million by the mid-1990s (of which 22.5 million were household businesses), employing over 135 million people in 1996 (Huang, 2008, Table 2.1; Chang & Wang, 1994). TVEs were nominally owned by township residents, but local governments often retained

⁶In parallel, the Chinese government also allowed rural residents use rights for forest land, which has been shown to induce an increase in investment in the forest land (Yi et al., 2014), especially where local elections improved local accountability (Deininger & Jin, 2009).

⁷Indeed, after the 1994 fiscal reform recentralized tax revenue, local governments scrambling for resources increasingly confiscated suitable land and leased it to developers for substantial fees which they used as collateral for loans to finance spending projects (Ang, 2016).

control over their profits since they controlled access to necessary resources and obtained higher share of revenue (relative to the standard private ownership form) (Chang & Wang, 1994; Qian, 2002). Nevertheless, from the outset many were run as private firms, and economically efficient (see Huang, 2008, pp. 75–85). Another key reason behind the TVEs' fast growth and success was the trusting cooperative culture that facilitated the formation of new business organizations and inter-firm relationships, including relationship-specific investment (Weitzman & Xu, 1994). According to data from the World Values Survey (averages between 1981 and 2014), China ranked the highest in the world in having 60% answering in the affirmative to the question: "Do you think most people are trustworthy?", which is consistent with the Confucian moral code's emphasis on trust (诚信) (Zhu, 2021, Chap. 7).

During the 1980's reformist era, the government not only allowed more private initiatives, but also began to shift SOEs from centralized bureaucratic control to more market-oriented control by management (Xu, 1998). SOEs were allowed to retain an increasing share of their profits and more managers were given autonomy over production decisions and employee wages.⁸ SOEs' productivity rose significantly in the 1980s due to stronger incentives, intensified market competition (mostly from TVEs), and improved factor allocation (Li, 1997).

The reforms from 1978 to 1989 helped facilitate private investors' perception that *de facto* property rights protection was improving. For instance, in the late 1980s, Zhao Zhiyang initiated the separation of the CCP and the state, which led to the perception of more rule-based government less subject to party intervention (Huang, 2023, p. 134). (This reform was reversed after the Tiananmen incident in 1989 and the power of the general secretary of the CCP was boosted at the expense of other power centers with important future consequences, according to Huang, 2023, p. 121.) Private investors were also reassured by political polarization in the 1980s, with both powerful reformers (led by Deng Xiaoping) and powerful conservatives (led by Chen Yun) exerting important influence, and reformers in the driver seat (Huang, 2023, pp. 122–124).

35.4.2 Increasing Support to Private Activity (1992–2007)

After Deng Xiaoping's Southern Tour in 1992, where he endorsed economic experimentation ("crossing the river by feeling for the stones"), there was an explosion in the number of private firms that were not TVEs, collectives, or individually owned (i.e., very small). Such enterprises got another boost in October 1992 when the Fourteenth Congress of the Chinese Communist Party described the goal of reform as establishing a "socialist market economy" (Clarke et al., 2007, Table 11.1).

⁸ Government also signed performance contracts with SOE managers (88% by 1989), ceding them greater discretion while setting specific targets (Xu, 1998). However, Shirley and Xu (2001) found that on average it was not performance contracts that improved SOE performance, but rather conducive circumstances such as greater competition that provided stronger incentives.

Subsequent laws gradually strengthened the legal basis for private firms (see *Ibid.* for details), and in 2004 the Constitution was amended to state that “The state encourages, supports and guides the development of the private sector” (Ang, 2016, p. 162). Under the “Three Represents” of President Jiang Zemin, the recruitment of private entrepreneurs into the party was encouraged, sending another signal of implicit property rights protection in China (Huang, 2023, p. 134).

Once private firms were allowed to expand beyond 7 employees, they ballooned. Simultaneously and partly because of the central government’s policy of shrinking rural bank branches (Han & Hare, 2013), credit-hungry TVEs declined as a share of the economy from the mid-1990s and small household businesses fell as well from 2000 onward (Ang, 2016, pp. 162–3). The expansion of the rest of the private sector (large and mid-sized firms, the remaining TVEs, and small household businesses) was combined with shrinkage of SOEs. In 1998 the government encouraged local governments to divest smaller SOEs and roughly 80% of local SOEs were privatized by 2009 (Huang et al., [forthcoming](#)). Most of the privatized firms were poorly performing small- and mid-sized firms, and their divestiture substantially increased the average productivity of all firms (Huang et al., [forthcoming](#)). China’s privatization was the largest in the world, raising about US\$700 billion in sales revenues (Huang et al., [forthcoming](#)). Nevertheless, as common throughout China’s dynasties, the government retained “commanding-heights” SOEs and by 2012 SOEs still accounted for 30% of industrial output (Long et al., 2022, p. 6). SOEs also still dominated the key service sectors (such as finance) and continued to receive most formal bank credit despite being less productive than private firms (Allen et al., 2018).

Another advance during this period was the high rate of foreign direct investment (FDI) into China, much of it from Hong Kong, Taiwan, and Macau (Wang et al., 2012). Since most foreign investors were overseas ethnic Chinese, they based their confidence in the security of property rights on informal rules of thumb, such as their perception that local leaders would make favorable investment decisions since they needed FDI to achieve their growth targets. These ethnic investors could also count on “voting with feet” and transferring gains abroad for added security.

35.4.3 Laws and Private Growth (2007–2012)

An important change during this period was the introduction of *de juri* private property rights with the first Property Law in 2007. “The law included measures that checked the power of local government to expropriate assets from private firms. The law gave creditors the right to seize their collateral if a private firm defaulted on a loan...” (Berkowitz et al., 2015, p. 3). This law had real consequences; it was associated with higher firm values on the stock market and increases in investment and productivity of enterprises (Berkowitz et al., 2015; Long et al., 2022). Berkowitz et al. (2015) also found that the law gave firms more access to external finance, especially those without political connections and good growth opportunities, making them less dependent on reinvesting their internal cash flows.

Despite the introduction of the Property Law, both private firms and SOEs continued to be adversely affected by regulatory barriers to entry and other barriers to competition, corruption, and other forms of confiscation that raised uncertainty and the cost of doing business (Long et al., 2022). Moreover, although the Property Law reduced the benefit of political connections somewhat, connections remained highly salient and unconnected firms continued to be more vulnerable, resulting in wide variance in firm circumstances. For example, a 2012 World Bank survey showed that firms in cities with worse institutional environments (i.e., deficiencies in rule of law, lower regulatory quality, or more corruption) were less likely to innovate than firms in cities with better institutional environments (Rodríguez-Pose & Zhang, 2020). Firms in better institutional environments, and those with lower state ownership, not only innovated more, their managers spent less time dealing with regulations.

Nevertheless, private firms continued to grow in size. Surveys suggest that employment in private firms grew on average by 167% from 1993 to 2012—from an average of 61 to 163 employees (Long et al., 2022, p. 14). Surveys also suggest that Chinese private firms became more innovative, investing more in R&D and training over time; in 2012 the ratio of R&D expenditures to revenues of private firms was seven-fold that of SOEs (Long et al., 2022, p. 14).

As for SOEs, restructuring and privatization contributed to a gradual convergence of productivity in private firms and SOEs between 1998 and 2007, although SOE returns to capital remained 40% below that of private firms (Hsieh & Song, 2015). This convergence proved short-lived. By 2016 the return on assets (ROA) of private industrial firms was 3.5 times that of SOEs, and 25% of industrial SOEs were losing money compared to less than 10% of private firms (Lardy, 2019). Nevertheless, SOEs continued to capture most banking credit and government subsidies. Moreover, although the reinvestment rate fell from 47% in 2010 to 10–22% in 2012,⁹ private firms continued to rely heavily on reinvestment of profits to finance their investments (Long et al., 2022, p. 14).

35.4.4 Reversals and Uncertainty (2012–Present)

The business environment for privately owned firms, which became less favorable after the 2008 global financial crisis, has continued to deteriorate. This deterioration stemmed partly from the rising role of local governments and local SOEs (in the form of local investing vehicles) implementing fiscal stimulus policies. Nevertheless, the number of private firms continued to increase, doubling between 2012 and 2022 according to China's National Market Supervision Bureau (Yao et al., 2022, p. 2), and increasing four-fold according to statistics from the National Bureau of

⁹Slowing GDP growth (from 10.6% in 2010 to 7.9% in 2012) might have affected this drastic drop in the reinvestment rate.

Statistics (NBS) website.¹⁰ Furthermore, the population of very large private firms has also grown, even though the largest firms remain overwhelmingly state-owned. In 2011 only four private firms were large enough to be listed in the Fortune Global 500, and their revenues represented about 4% of the total revenues for all 58 Chinese firms on the list. By 2021, 32 private firms made the list with revenues representing 19% of the revenue of the 130 Chinese companies listed (Huang & Véron, 2022, p. 3). Moreover, private firms increasingly dominated the Chinese economy even though property rights protection, while improving, remained weak. According to the then head of China's industry and commerce federation (cited in Ang, 2020, p. 67), in 2017 the private sector was responsible for 70% of innovation, 80% of employment, and 90% of new jobs and businesses. Despite these positive signs, policies characterized as "the state advances and the private sector retreats" have altered the regulatory and financial framework in favor of SOE dominance and government and Chinese Communist Party (CCP) control. These include the anti-corruption campaign, concerted efforts to strengthen and expand SOEs, nationalization of private and privatized firms (Huang et al., 2021), changes to bank lending standards reducing private access to loans and regulatory tightening that diminished the market value of some of China's most successful private enterprises (see also Zhang, 2018).¹¹ This worsening business climate has had important consequences. Du et al. (2014) find that the phenomenon of "state advances and private sector retreats" has reduced aggregate total factor productivity in manufacturing due to systemic misallocation of resources in the manufacturing sector, where less efficient state-owned firms are more likely to enter markets, more likely to expand their size, and less like to exit, than their private counterparts.

Adverse policy changes also contributed to slower growth in private investment which fell well below growth in SOE investments in 2016 (Lardy, 2019, p. 39). Lardy (2019, p. 40) attributes this in part to banks directing a larger share of credit to SOEs since 2012. Also, the emphasis on the increased role of state and party in the economy and the seizure of private enterprises by the state "chilled the investment enthusiasm of private entrepreneurs" (Ibid.). In 2011, 54% of domestic currency bank loans went to private firms and 28% to SOEs, but by 2016 private firms received 11% and SOEs, 83% (Lardy, 2019, p. 141). A crackdown on shadow banking (explained below) also reduced funding for firms unable to access commercial banks. The government, moreover, has directed a much larger share of state

¹⁰The NBS numbers include private firms only, and exclude firms with foreign and overseas Chinese ownership, which did not change much during this period.

¹¹The avowed policies of the government toward private enterprises have been inconsistent. In speeches in 2017 and 2018 President Xi called for an expanded role for the CCP, including creating party committees in private enterprises. A different tone was struck by the CCP Central Committee and State Council in 2016 when they jointly called for improved protection of property rights and acknowledged that the state had illegally seized the property of private entrepreneurs (Lardy, 2019, p. 41). And at a March 2018 National People's Congress, Premier Li Keqiang stated that weak private investment was due to "weak protection of property rights and some other factors" (*China Daily* March 21, 2018, p. 6 cited in Ibid.).

subsidies to SOEs compared to private firms even though private firms are more effective in converting research and development spending into patents (Wei et al., 2017). Long et al. (2022) further report a decline in private firms' reinvestment rates since 2010, consistent with the slowdown in private investment. In 2022 investment by private firms grew by only 0.9% compared to 10% for SOEs (Huang & Lardy, 2023). Tellingly, when an urban land titling reform in Shenzhen gave firms with ambiguous property rights formal titles to their land it raised the investment rate only in SOEs and politically connected firms, not in private firms without connections or state affiliation (Miao et al., 2023). This contrasts with the earlier finding that the Property Law had more positive effects in *less* politically connected firms before 2010 (Berkowitz et al., 2015), suggesting that the importance of political connections for securing property rights protection likely increased in the subsequent decade.

The anticorruption campaign launched in 2012 arguably had profound effects on the private sector. Corruption has been widespread in China. For example, Cai et al. (2011) found that entertainment expenditures in their sample of firms were a mix of bribes and relational capital amounting to 20% of wage bills with negative effects on firm performance. The wide-ranging anticorruption campaign has "punished over 2.9 million people" including managers of corporations (Griffin et al., 2021, p. 6). Griffin et al. (2021) found that the anticorruption campaign did indeed capture firms that had more signs of corruption than their matched counterparts (such as lower minority share ownership, higher abnormal CEO compensation, and higher business entertainment expenditures), but managers with stronger connections to the central government were less likely to be investigated. Given how much the campaign undermined local government connections, a key institution protecting private property rights from expropriation (as we explore below) was likely shaken as well.

The next section analyzes how the development of the private sector in China was and remains highly dependent on informal institutions, despite improvements in formal institutions.

35.5 Institutions Fostering Private Sector Development in China

The institutions that gave private firms in China *enough* security to motivate their extraordinary growth were both formal and informal. Formal institutions according to North (1990) are codified, traceable, and transmissible rules enforced by organized third parties such as state-run legal systems, while informal institutions are unwritten codes of conduct that arise to coordinate repeated human interactions, such as community-sanctioned norms of behavior or internally enforced standards of conduct. Table 35.1 lists the formal and informal institutions in China that performed the two functions we described in Sect. 35.2: (1) enforcement of contracts, and (2) protection from expropriation by government and elites. Formal laws and regulations co-existed with informal rules and this section explains how Chinese private firms relied on both depending on the circumstances.

Table 35.1 Institutions protecting property rights in China

Informal institutions	
Personal relationships	Bribes, favors, kinship ties, <i>guanxi</i> , fellow-townsmen
Firm relations	Business networks and associations, clusters, trust, reputation
Political connections	Links/importance to local governments, manager's/owners' political position, equity link to SOEs
Confucian culture	Culture emphasizing trust and moral responsibilities
Formal institutions	
Legal	Laws
Government	Regulations

35.5.1 Enforcement of Contracts

Initially private firms in China necessarily relied heavily on informal institutions that had enforced contracts during China's long history and are still important today. As summarized in Table 35.1 these include: (1) personal relations, such as *guanxi*, family ties and clans; fellow-townsmen; (2) firm-level institutions, such as networks (including affiliation to group companies), business associations and industrial clusters, and trust- and reputation-based institutions, (3) political connections, such as links to local governments,¹² and (4) the Confucian culture of emphasizing trust and responsibilities.

Many scholars argue that kinship ties, including clan membership, were important in the early success of private Chinese firms by partly compensating for their lack of access to bank financing and tenuous legal status (Allen et al., 2005; Chen et al., 2017, 2022; Greif & Tabellini, 2010, 2017; Lu & Tao, 2009; Nee & Oppen, 2012; Peng & Heath, 1996). China has a long tradition of sustaining cooperation through kinship obligations enforced by concerns about reputation. Allen et al. (2005, p. 60) argue that Confucian beliefs in the centrality of family, social order, and trust have been the most important force shaping China's social values and institutions (see also Zhu, 2021). Greif and Tabellini (2017) also point to the Confucian emphasis on family as one reason for the endurance of clans as China's predominant social organization. Indeed, the Chinese characters for the state (国家) literally mean that the state shelters a family with substantial wealth under its roof.

A related informal institution in China is the fellow-townsmen (同乡) relationship, the enhanced trust of people who are now or were previously from the same location. Being a continent-size country with huge regional variations (Pomeranz, 2000, p. 7) and without sophisticated third-party enforcement, Chinese people have historically relied on strong trust in fellow townsmen. Such location-based trust has

¹²Park and Luo (2001, p. 1) define *guanxi* as an "intricate and pervasive relational network" characterized by mutual obligations, assurances, and understandings that creates a "web of connections" that can be drawn on to secure favors.

facilitated regional merchant groups, such as the key groups formed by people from Shanxi (晋商) or Anhui (徽商), which helped the Chinese economy operate smoothly. Recent evidence suggests that the institution of fellow-townsmen is still influential. For example, counties in Fujian province where the local guerrilla force participated in county government right after the founding of the People's Republic of China in 1949 had about 2 percentage points higher annual growth rates, stronger private-sector development, better local education after the reform; they also had fewer deaths during the Great Famine of 1958–1960 (Fang et al., 2023). Similar evidence is found in Zhejiang province where the private sector was most developed in China (Zhang & Liu, 2019). Relatedly, Chinese provinces with provincial leaders who had spent more years living in the province (such as being born there or having worked there for some time) had higher expenditures on education and health and better local welfare (Persson & Zhuravskaya, 2016).

Even while private firms depend on kinship and fellow-townsmen relations for seed money, they also make use of non-family, non-bank sources of credit supplied under contracts enforced through firm-level informal institutions. This so-called “shadow banking system” consists of “informal financial intermediaries, internal financing and trade credits, and coalitions among firms, investors, and local governments” (Allen & Qian, 2014, pp. 501–2). Shadow bank loans are usually enforced not through the law, but through “trust, reputation, and relationships” (p. 502). Indeed, SOEs with better access to bank loans often grant trade credit to their trading partners (Cull et al., 2009). The importance of these informal institutions goes beyond start-up capital. Nee and Oppen (2012) argue that the informal arrangements within networks of like-minded individuals enabled private firms to start up, survive and grow outside the state-owned manufacturing sector despite high barriers to entry and in the process created China's “capitalism from below.” Typically, such networks are self-enforcing: firms comply with informal agreements and shadow loan obligations to secure their reputations as trustworthy members of their vital networks.

Firms in all economies rely on relations, networks, and informal institutions to secure their property rights and enforce contracts (Williamson, 2002). Williamson argues that, given the costs of litigation, businesses logically prefer such “private ordering” over courts, police, or other forms of “public ordering.” In China, however, reliance on informal institutions is not a “strategic choice by design” but a “strategy by default. . . the most viable way of doing business” according to Peng and Heath (1996, p. 518). Smaller private firms have had to rely on networks for seed capital, since the financial system has long been dominated by state-directed funding to SOEs. They have also relied on *guanxi*, kinship, and clan connections to settle disputes, since the legal system has been uncertain and biased. For example, firms in the Yangzi Delta reported in a 2006 survey that they relied on networks of family and close friends for seed capital, advice, mutual cooperation, and contacts with first customers (Nee & Oppen, 2012).

Industrial clusters were another tool helping smaller firms overcome their credit constraints and enforce contracts. Industrial clusters grew rapidly between 1995 and 2004 when thousands of manufacturing firms, each specializing in a finely defined

production step, located together in a densely populated area, typically with local governments supporting the necessary infrastructure (Long & Zhang, 2011). By lowering minimum capital requirements, such clustering allowed wealth-constrained entrepreneurs to start a business. It also changed firm incentives. "Close proximity and intense competition of firms within a cluster may also reduce the incentive to act dishonestly, making frequent trade credit among firms within a cluster possible" (Long & Zhang, 2011, p. 122). The strong trust stemming from the long Confucian tradition also facilitated inter-firm networks.

Increasingly private firms have also turned to formal institutions. Long, 2010 provided evidence that the Chinese court system started to provide effective property and contracting protection for businesses, especially those involved in long-term investment and long-distance trade. Long et al. (2022) find firms are more willing to rely on formal contracts and to use courts to settle economic disputes thanks to reforms in the legal system and the stronger legal basis for private firms after enactment of the 2007 Property Law. The share of firms using the court system to resolve disputes more than tripled, from a modest 9% in 1995 to a respectable 32% in 2006, and satisfaction with the results increased from 54% to 83% over that same period (Long et al., 2022, p. 14). Similarly, interviews by Chen et al. (2017, p. 3) between 2013 and 2015 found evidence of "a transition from *guanxi*-based transacting to a more formal, legally driven approach to contracts" thanks to increasing professionalism and independence of the judiciary (p. 13). Private firms are also increasingly appealing to local governments for dispute resolution as well, not surprising given China's millennium-long history of reliance on local elites and officials for resolution of local disputes (von Glahn, 2018, p. 285).¹³

Not all private firms are treated equally by these formal systems, however. Long et al. (2022) find that politically connected firms use the court system more than other private firms, which suggests that the legal system favors these connected firms. Similarly, earlier data analyzed by Ang and Jia (2014) found that close to 85% of private firms relied on private mediation or negotiation rather than courts, but firms headed by a delegate to a Peoples' Congress or Conference were 22% more likely to use courts and those headed by a former government official were 86% more likely.¹⁴ Again, politically connected firms seem to expect a better outcome from courts than from private negotiations. The judicial reform that has been gradually rolled out since 2014 may have reduced the courts' bias toward politically connected firms. The reform increased the local court's independence by moving control over the financial and personnel decisions of county and prefectural courts from local to provincial governments. Lei and Li (2022, p. 10) argue that local officials often interfered in case decisions in favor of local litigants because local firms were (and are) an important source of tax revenue and officials were fearful that

¹³The share of firms relying on local governments for dispute resolution increased from 2% to 21% between 1995 and 2006 (Long et al., 2022, p. 14).

¹⁴Long (2010) also documents the perceived preferential treatments enjoyed by SOEs in the Chinese judicial system.

firms might move to a competing jurisdiction should they fail in court. The authors further suggest that thanks to insider information and personal connections, local firms were better at bribing local officials and judges than external firms (p. 10). The judicial reforms since 2014 have substantially reduced the win rates for local firms, especially for politically connected firms, whose win rate fell by twice as much as that of unconnected firms (Lei & Li, 2022, p. 17).

The rapid and continuing growth of the private sector in China testifies to the success of these informal, and increasingly formal, contract enforcement institutions. Indeed, based on World Bank firm-level data from around 2002, during China's period of "increasing support to private property," Cull and Xu (2005) show that private firms' reinvestment rates were boosted by improved protection of both contracting rights *and* property rights against government expropriation. As Acemoglu and Johnson theorized and subsequent studies on Chinese firms show, however, protection against expropriation has been stubbornly elusive over time, as discussed below.

35.5.2 Protection Against Expropriation

Protection against expropriation of private businesses showed little improvement in China from 1993 to 2012 according to Long et al. (2022, p. 8), as measured by the tax burden, extra legal fees and assessments collected by local governments, and expenses on entertaining local officials and cultivating public relationships (which are shown to be related to corruption; see Cai et al., 2011). One of the most important safeguards against expropriation has been personal relations with local governments built through *guanxi*, favors, and bribes (Ang, 2016, 2020; Bai et al., 2019, 2021; Cull et al., 2015; Huang, 2011; Nee & Oppen, 2012; Park & Luo, 2001). There is evidence that firms pay local officials bribes to build relational capital, prevent predation, or secure lower taxes or special deals (Cai et al., 2011). They also bribe through hidden equity stakes to family members (Bai et al., 2019). Political fortunes can change quickly over time, however, making political connections inherently fragile mechanism requiring constant investments in new political relationships.

Personal relations with local officials are not the only safeguard. Perhaps more important, firms are protected by being indispensable to regional development as a source of investment, employment, and taxes. Local officials have strong incentives to protect and promote important local firms. Since local governments only keep a proportion of the taxes collected from local firms while having to self-finance the provision of most local public goods, they are motivated to maximize their tax revenues (Lin, 2022, Chap. 11). Local officials were also motivated by prospects for promotion, which long depended on meeting growth targets set by the next higher strata of government (Li & Zhou, 2005). Increasingly they also face targets for education, infrastructure, social stability, and increasingly, personal conduct. Over time private firms have become critical to achieving these targets not only because of their importance to taxable revenues, but also, to jobs and economic development more broadly (Nee & Oppen, 2012, pp. 228–9). Local governments

also rely on sales of land under their control to real-estate businesses as a source of revenue, further motivating local officials to burnish the reputation of their locality as a favorable business environment with potential investors. Moreover, local government officials compete with their counterparts in other locations to attract private business, which gives firms the option to move should their local government try to confiscate too much of their rents. Ang (2016, 2020) suggests that the relationship between local governments and private firms is akin to a franchising arrangement or a profit-sharing scheme in which government officials use industry profits to top up salaries of rank-and-file bureaucrats and fund desirable projects, giving everyone in the administration a stake in the private firms' performance. Ang (2020) further argues that officials' drive to foster local businesses has motivated them to create more predictable, rules-based business environments, improve their bureaucracies in response to businesses' demands for more professionalism, curb petty bureaucratic corruption, and reduce their reliance on fees and fines as their tax revenues grow.

The importance of relations with local governments can best be understood in the context of China's extensively decentralized economic system. As Xu (2011) points out, China is not a *de jure* or *de facto* federal state. Rather, regional power is granted by the central government and the CCP, which control vital levers of power, including over the careers of provincial leaders, the personnel policies of sub-national governments, important economic sectors (banking, telecommunication, energy), and the mass media and ideology. Yet local governments "are tasked to steer their own economies, collect taxes, finance and supply public services, and maintain law and order" (Ang, 2016, p. 106). Administrative and economic authority is delegated through five layers of government (i.e., central, province, city, county, and township), and each level appoints the leaders and governs the next level. Xu (2011, p. 9) describes China in the early 2000s as the most fiscally decentralized country in the world.¹⁵ This decentralization endowed China with adaptability and allowed the experimentation so critical to the emergence and strength of the private sector. Chinese local governments have "enormous administrative capacity and use it to extend a 'helping hand' to favored firms. This helping hand includes exemptions from regulations, lobbying the central government for the right for favored firms to break rules, improving local infrastructure, providing land (and to a lesser extent credit) at below market prices, and blocking the entry of other firms that threaten the profits of the favored firms" (Bai et al., 2019, p. 3).

Private entrepreneurs intent on protection against expropriation have enhanced their security by becoming members of the CCP (Li et al., 2008; Long et al., 2022), joining the People's Congress or the Peoples' Consultative Conference (Hou, 2019; Long et al., 2022), joining local legislatures, and appointing board directors with ties to top government officials (Alonso et al., 2022). Survey respondents in Hou (2019, Chap. 4) suggest that private entrepreneurs view predation as a major problem, and a main motivation for running for political office is to secure their property rights.

¹⁵ Expenditures by subnational governments were 70% of the national total compared to 56% in the United States or 40% in Germany (Xu, 2011, p. 9).

Such connections have large payoffs. Long et al. (2022) show that firms whose CEOs belong to the Party, the Congress, or the Consultative Conference had higher profitability, more patents, and enjoyed higher social, political, and economic status than non-connected private firms, probably reflecting their privileged access to resources such as bank loans, subsidies, land, and licenses. And as we noted earlier, they also relied more on formal courts for dispute resolution. Entrepreneurs who serve in the Peoples' Congress spend 14.5% less on informal payments to local governments on average (Hou, 2019, p. 103). Private firms with directors who had school ties to Politburo members got 16% higher subsidies (as a percent of sales revenues) from the government (Alonso et al., 2022).¹⁶ Firms with politically connected chairmen or CEOs face lower corporate tax rates compared to unconnected firms (Wang & You, 2022; Wu et al., 2012). In sum, firms with political connections can be said to face a different, more supportive business environment compared to non-connected enterprises, a difference that has increased over time (Long et al., 2022, p. 2).

The anticorruption campaign had a profound effect on firm relationships with local governments, making local officials reluctant to strike special deals with private firms or afford them the sort of protection that made possible the rise of many private enterprises in the first place. Thus, an unintended consequence of the anti-corruption campaign has been to damage the informal relations with local governments that serve to buttress private enterprises against expropriation. In response, private firms developed other ways of securing property rights, including, as we mentioned, greater political connections, as well as by blurring the line between state-owned and private property.

Increasingly private firms have aligned themselves with SOEs or with private investors partly owned by SOEs to the point where the assets of state-connected owners reached 44% of total assets of all private owners by 2019 (Bai et al., 2021, p. 3).¹⁷ These connected investors form a hierarchy of ownership ties whereby SOEs increased their investments in joint ventures with private firms and those private owners in turn increased their investments in joint ventures with other, smaller private firms. This expansion is relatively recent: the share of registered capital of state-connected private owners increased by 20% between 2000 and 2019. Private owners in such joint ventures with SOEs likely benefit from greater access to financial resources from banks, since SOEs produce one-third of output but are allocated two-thirds of credit (Bai et al., 2021, p. 34). This helps explain why mixed

¹⁶ School ties are defined as attending the same university over a similar 4-year period (Alonso et al. p. 5). 95% of their sample of connected firms received subsidies amounting to on average 1.6% of sales (p. 7).

¹⁷ This increase in state ownership in private firms could reflect one of two possibilities: (i) private owners willingly aligned themselves with the state for protection from expropriation, (ii) government officials forced private owners to give up their equity share to these officials and/or related parties. Both possibilities reflect a rise in state power.

private/state ownership is associated with greater profitability and productivity (Allen et al., 2022).¹⁸

Political connections and interlocking ownership serve to protect Chinese firms from state expropriation, but these approaches have drawbacks. They invite collusion with government and rent-seeking, and create a strong vested interest group, weakening incentives to seek more fundamental institutional changes such as equal protection under the rule of law. By protecting connected firms from the incentive effects of market competition, political connections also reduce pressures for efficiency and innovation. Indeed, Long et al. (2022) find that China's politically connected firms are less likely to innovate, which is consistent with findings elsewhere in the world. Interlocking ownership could also lead to increased government and Party control over private corporate decision-making, reducing risk-taking and stifling market competition. Finally, reliance on government-business relationships for protection reduces the legitimacy and prestige of these connected firms in the eyes of non-connected firms, and perpetuates the uneven playing field between insiders and outsiders in China's business world.

As discussed earlier, an important safeguard for private firms was the importance of economic growth to local governments, but that safeguard has become less effective in China's current business environment. China's central government has changed its priorities and its goals have shifted since 2012 with increasing emphasis on conformity to Party ideology and loyalty (Ang, 2020, p. 202) and, relatedly, expanding support for SOEs despite their deteriorating performance (Lardy, 2019). According to a survey of the CCP's Central Organization Department, 68% of private firms and 70% of foreign firms had party bodies by 2016 and the goal is to expand even further (cited in McGregor, 2019). Other adverse factors, notably widespread lockdowns under China's Zero Covid policy and continued regulatory uncertainty, hit private firms especially hard and they are now lagging SOEs in revenue, investments, and profitability (Huang & Lardy, 2023). As we discussed earlier, growth of investment by private firms and new firm entry have slowed. Although official rhetoric strongly supports private firms, there is no evidence yet that the private sector dynamism has returned to its earlier levels.

35.6 What Does the Future Hold?

We have presented evidence that China's rapid economic development was largely driven by its vibrant private sector, which was heavily dependent on informal institutions and a supportive culture to counter the risks of weak formal institutions. We have also shown that China has achieved important progress in improving

¹⁸There are tradeoffs, however. A field experiment in 2019 found that Chinese private firms preferred venture capital equity partners who were not affiliated with the government (Colonnelli et al., 2022). The authors submit that this arises from concerns about government interference in decision-making.

formal, as well as informal, contracting institutions, but made much less progress in improving anti-expropriation institutions. Moreover, there are signs that institutional improvements are diminishing or even reversing, the business environment has worsened and new private entry and investment have slowed. Besides the failure of incumbent firms to improve their productivity, Brandt et al. (2020) report a worrisome slowdown in the entry of new firms. The small, innovative startups that generated so much dynamism in earlier years now confront a very different business environment dominated by mixed ownership structures and unpredictable risks from government-mandated shut downs and confiscation of prominent and commercially successful entrepreneurs and enterprises. How does our analysis inform our understanding about China's future? We see warning signs in four areas.

First, informal institutions can only take you so far; further improvements in formal institutions are also necessary. The informal institutions in Table 35.1 have served Chinese private firms well, but they suffer from well-known limitations that threaten future growth and investment. Personal or kinship relations, business networks, or political connections are highly imperfect substitutes for the formal protections provided in an open access system where rule of law is safeguarded by political systems accountable to an active and informed civil society (North et al., 2009). Not only are these substitutes fragile, they have serious economic costs. As Greif (2006) describes, by excluding nonmembers, networks and kinship groups place costly restrictions on the scope of alliances and markets, reducing specialization and gains from trade. Reliance on personal relationships, such as *guanxi*, to enforce contracts is also inherently limiting in scope, making it impossible to achieve full economies of scale (Long, 2018). More generally, institutions that rely on reputation and personal ties as enforcement mechanisms require monitoring and adequate information on cheating, which raises the information costs of adding new members, again limiting economies of scale and scope. Networks and relationships are also discriminatory by their very nature, reducing the choice of talent and resources. Membership in such groups constitutes a barrier to entry for non-members, limiting competition.

These limitations and weaknesses are not unique to China, but what is unusual is the extent to which China's private sector expansion continues to depend heavily on informal institutions. Economies with better rules and stronger formal enforcement such as reliable courts can support more complex and potentially riskier transactions (Williamson, 2002, p. 440 n 2). The evidence we presented suggests that the continual improvement of impersonal, formal forms of contracting institutions has allowed businesses in China to supersede the limits of informal institutions and engage in sophisticated and complex transactions (Cull & Xu, 2005; Long & Zhang, 2011; Long et al., 2022). But these past achievements are threatened and further improvements are not assured, jeopardizing China's private-led development. Continued reliance largely on informal institutions could slow future private growth, not least because networks and personal ties reduce competition, curbing firms' incentives to innovate and improve efficiency. Moreover, without the backing of a strong legal framework, China's informal mechanisms are not only inherently insecure, they are increasingly under threat; for example, ties to local governments are under

threat from the anti-corruption campaign. This insecurity makes private investment riskier; another drag on future growth.

Second, improving anti-expropriation institutions in China is a particularly necessary and daunting task ahead. The literature suggests that effective protection against elite expropriation requires institutions that bind powerful actors, including state actors, to submit themselves to rule of law and abide by legal judgments. North et al. (2009) argue that open access and entry to the economy should be complemented by open access and entry to politics, allowing citizens to form competing civic organizations, including political parties, to advance their interests against state encroachment. Similarly, Acemoglu and Robinson (2019) argue that a strong state must be balanced by a strong, mobilized society. While there have been marked improvements in China's legal system at the local level, described above, there have been no similar developments strengthening civil society vis-à-vis the state and the CCP. And, as we described, China has a long history of strong-state-weak-society since the centralization of the Qin Dynasty two millennia ago. Indeed, the recentralization of power from local governments to central government, the de-institutionalization of the two-term limit for presidents, the perception of declining checks and balance over time,¹⁹ and the current emphasis on ideological conformity have arguably increased state power vis-à-vis civil society. Again, the resulting insecurity will slow private investment and growth.

Third, the trend toward increasingly close ties between private and public sectors poses risks for the future. In the 1980s, China offered strong incentives for firms that focused on productive efficiency, investment, and innovation, but over the last decades the rewards for political connections and ownership ties to the state have become relatively stronger. The global economic crises in 2007 and the subsequent state stimulus package, the anti-corruption campaign with its erosion of local government ties, the abrupt loss of property and power of prominent entrepreneurs, the halt in activity, and the absolute state power evidenced in the Zero Covid shut downs, have all changed private investors' assumptions about the business environment and how best to secure property. Increasingly, private firms are seeking equity shares with SOEs and closer links with the central government and the CCP. While such interlocking ownership ties increase private access to resources, at what cost in efficiency? As described earlier, total factor productivity (TFP) in manufacturing SOEs in China is lower than in private firms and TFP in industrial SOEs is no longer converging with private firms, but falling behind (Brandt et al., 2020). Ownership linkages with less efficient SOEs could potentially shift resources to more efficient private firms, but the productivity numbers do not reflect that. Instead, TFP growth for all manufacturing firms has slowed to 0.9% from 2009 to 2018 compared to 2.8% in the previous decade (Brandt et al., 2020, p. 2). Tying more and more resources to

¹⁹Huang (2023, ► Chap. 6) suggests that the weakening of the role of political elders (such as charismatic leaders like Deng Xiaoping) as a constraint in the Chinese political system could mean a decline in checks and balance.

the least productivity sector of the economy endangers efficiency and future development.

Fourth, innovation will likely be the major driver of China's future economic success but the present incentives are not directed at the most innovative segments of the economy. China has achieved great progress in innovations, as manifested in having the highest number of patents (in 2018), the highest growth rate in patent application, and the second highest number of top-citation scientists based on Web of Science data in the world (Zhu, 2021, ► Chap. 6). However, an unhealthy development for future innovation is that the Chinese government seems determined to rely on SOEs for innovation, as reflected by the increasing share of subsidies going to SOEs. Millennia ago, China achieved spectacular innovation via government-related innovators (Huang, 2023, pp. 251–284), such as the invention of paper centuries ahead of the Muslim world and Europe. The prestige and benefits of innovations that the central government currently encourages could still lure important advances in these privileged areas. But depending on SOEs for innovations more broadly is risky given their low efficiency, and strong emphasis on loyalty to government policies and ideology.

China's future innovations will also depend critically on its success in fostering international collaboration which was hugely beneficial in past decades (Huang, 2023, pp. 273–275); its top technological giants were often registered overseas, obtaining the benefits of international rule of law; and, the top scientific articles (such as in *Nature* and its derivative journals) have often been international collaborations. The deterioration of China's relationship with the West could increase the uncertainty of its future innovation potential.

China's future development depends on heeding these warning signals and reversing the damaging trends. It would be risky to bet against China's dynamic private sector, which has performed so well against difficult challenges in the past. But the future is murky unless the formal, legal institutions that increasingly secured private activity in the past are again allowed to strengthen and expand, the current expropriation risk is eased, the emphasis on efficiency is restored, and the private sector's innovative spirit and collaborative approach is allowed to flourish.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., & Johnson, S. (2005). Unbundling institutions. *Journal of Political Economy*, 113(5), 949–995.
- Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail: The origins of power, prosperity, and poverty*. Crown.
- Acemoglu, D., & Robinson, J. A. (2019). *The narrow corridor: States societies, and the fate of liberty*. Penguin Press.
- Allen, F., & Qian, J. (2014). China's financial system and the law. *Cornell International Law Journal*, 47, 499–553.

- Allen, F., Qian, J., & Qian, M. (2005). Law, finance, and economic growth in China. *Journal of Financial Economics*, 57(1), 57–116.
- Allen, F., Qian, J., & Qian, M. (2018). *A review of China's Institutions* (SSRN working paper # 3269018).
- Allen, F., Cai, J., Xian, G., Qian, J., Zhao, L., & Wu, Z. (2022). *Centralization and decentralization? The evolution of state-ownership in China* (SSRN working paper 4283197).
- Alonso, M., Palma, N., & Simon-Yarza, B. (2022). The value of political connections: Evidence from China's anti-corruption campaign. *Journal of Institutional Economics*, 18(5), 1–21.
- Ang, Y. Y. (2016). *How China escaped the poverty trap*. Cornell University Press.
- Ang, Y. Y. (2020). *China's gilded age: The paradox of economic boom and vast corruption*. Cambridge University Press.
- Ang, Y. Y., & Jia, N. (2014). Perverse complementarity: Political connections and the use of courts among private firms in China. *Journal of Politics*, 76(2), 318–322.
- Bai, C.-E., Hsieh, C.-T., & Song, Z. M. (2019). *Special deals with Chinese characteristics* (NBER working paper No. 25839).
- Bai, C.-E., Hsieh, C.-T., Song, Z. M., & Wang, X. (2021). *Special deals from special investors: The rise of state connected private owners in China* (NBER working paper no. 28170).
- Berkowitz, D., Lin, C., & Ma, Y. (2015). Do property rights matter: Evidence from a property law enactment. *Journal of Financial Economics*, 116(3), 583–593.
- Brandt, L., Litwack, J., Mileva, E., Wang, L., Zhang, Y., & Zhao, L. (2020). *China's productivity slowdown and future growth potential* (World Bank policy research working paper 9298).
- Broadberry, S., Guan, H., & Li, D. D. (2018). China, Europe, and the great divergence: A study in historical National Accounting, 980-1850. *The Journal of Economic History*, 78(4), 955–1000.
- Cai, H., Fang, H., & Lixin Colin, X. (2011). Eat, drink, firms and government: An investigation of corruption from entertainment expenditures of Chinese firms. *Journal of Law and Economics*, 54(1), 55–78.
- Carey, J. M. (2005). Presidential versus parliamentary government. In C. Ménard & M. M. Shirley (Eds.), *Handbook of new institutional economics* (pp. 91–122). Springer.
- Chang, C., & Wang, Y. (1994). The nature of the township-village enterprise. *Journal of Comparative Economics*, 19(3), 434–452.
- Chen, D., Deakin, S., Siems, M., & Wang, B. (2017). Law, trust and institutional change in China: Evidence from qualitative fieldwork. *Journal of Corporate Law Studies*, 17(2), 257–290.
- Chen, Z., Ma, C., & Sinclair, A. J. (2022). Banking on the confusion clan: Why China developed financial markets so late. *Economic Journal*, 132(644), 1378–1413.
- Clarke, D. C., Murrell, P., & Whiting, S. H. (2007). The role of law in China's economic development. In T. Rawski & L. Brandt (Eds.), *China's great economics transformation* (pp. 375–428). Cambridge University Press.
- Coase, R., & Wang, N. (2012). *How China became capitalist*. Palgrave Macmillan.
- Colonnelli, E., Li, B., & Liu, E. (2022). *Investing with the government: A field experiment in China* (SSRN working paper 4016868).
- Cull, R., & Xu, L. C. (2005). Institutions, ownership, and finance: The determinants of investment among Chinese firms. *Journal of Financial Economics*, 77, 117–146.
- Cull, R., Lixin Colin, X., & Zhu, T. (2009). Formal finance and trade credit during China's transition. *Journal of Financial Intermediation*, 18(2), 173–192.
- Cull, R., Li, W., Sun, B., & Xu, L. C. (2015). Government connections and financial constraints: Evidence from a large representative sample of Chinese firms. *Journal of Corporate Finance*, 32, 271–294.
- Deininger, K., & Jin, S. (2009). Securing property rights in transition: Lessons from implementation of China's rural land contracting law. *Journal of Economic Behavior and Organization*, 70(1–2), 22–38.
- Du, J., Liu, X., & Zhou, Y. (2014). State advances and private retreats? Evidence of aggregate productivity decomposition in China. *China Economic Review*, 31, 459–474.
- Fang, H., Liu, M., Hou, L., Xu, L. C., & Zhang, P. (2023). Factions, accountability, and Long-term development: Theory and evidence. *Journal of Comparative Economics*, 51(1), 15–40.

- Giles, J., & Mu, R. (2018). Village political economy, land tenure insecurity, and the rural to urban migration decision: Evidence from China. *American Journal of Agricultural Economics*, 100(2), 521–544.
- Greif, A. (2006). *Institutions and the path to the modern economy. Lessons from medieval trade*. Cambridge University Press.
- Greif, A., & Tabellini, G. (2010). Cultural and institutional bifurcation: China and Europe compared. *American Economic Review: Papers & Proceedings*, 100(2), 1–10.
- Greif, A., & Tabellini, G. (2017). The clan and the City: Sustaining cooperation in China and Europe. *Journal of Comparative Economics*, 45(1), 1–35.
- Griffin, J. M., Liu, C., & Shu, T. (2021). *Is the Chinese anti-corruption campaign authentic? Evidence from corporate investigations*. Available at SSRN: <https://ssrn.com/abstract=2779429>
- Haggard, S., & McCubbins, M. D. (2001). Introduction: Political institutions and the determinants of public policy. In S. Haggard & M. D. McCubbins (Eds.), *Presidents, parliaments, and policy* (pp. 1–17). Cambridge University Press.
- Han, L., & Hare, D. (2013). The link between credit markets and self-employment choice among households in rural China. *Journal of Asian Economics*, 26(C), 52–64.
- Hou, Y. (2019). *The private sector in public office: Selective property rights in China*. Cambridge University Press.
- Hsieh, C.-T., & Song, Z. M. (2015). Grasp the large, let go of the small: The transformation of the state sector in China. In *Brookings papers on economic activity* (pp. 295–346). Brookings Institutions.
- Huang, Y. (2008). *Capitalism with Chinese characteristics. Entrepreneurship and the state*. Cambridge University Press.
- Huang, P. C. C. (2011). The theoretical and practical implications of China's development experience: The role of informal economic practices. *Modern China*, 37(1), 3–43.
- Huang, Y. (2023). *The rise and fall of the EAST: How exams, autocracy, stability, and technology brought China success, and why they might Lead to its decline*. Yale University Press.
- Huang, T., & Lardy, N. R. (2023, February 7). *China's support for the private sector is only lip service so far* (Peterson Institute for International Economics Realtime Economic Issues Watch).
- Huang, T., & V ernon, N. (2022). *The Private Sector Advances in China. The evolving ownership structures of the largest companies in the Xi Jinping Era* (Peterson Institute for International Economics working paper 22-3).
- Huang, Z., Li, L., Ma, G., & Qian, J. (2021). The reversal of privatization in China: A political economy perspective. *Journal of Corporate Finance*, 71, 102115.
- Huang, Z., Liu, J., Ma, G., & Xu, L. C. (forthcoming). The effects of privatization on firm productivity in China. *Journal of Law and Economics*.
- Keefer, P. (2011). *Collective action, political parties and pro-development public policy* (World Bank policy research working paper 5676).
- Klein, B., Crawford, R. G., & Alchian, A. A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21(2), 297–326.
- Lai, W., Xu, L. C., & Yang, L. (2023). *Property rights, state capacity, and society capacity: The lasting impact of the Taiping rebellion*. CKGSB.
- Lardy, N. R. (2019). *The state strikes Bank: The end of economic reform in China?* Peterson Institute for International Economics.
- Lei, Z., & Li, Y. (2022). *Making local courts work: The judicial recentralization reform and local protectionism in China* (SSRN working paper 4052769).
- Li, W. (1997). The impact of economic reform on the performance of Chinese state enterprises, 1980–1989. *Journal of Political Economy*, 105(5), 1080–1106.
- Li, H., & Zhou, L.-A. (2005). Political turnover and economic performance: The incentive role of personnel control in China. *Journal of Public Economics*, 89, 1743–1762.
- Li, H., Meng, L., Wang, Q., & Zhou, L.-A. (2008). Political connections, financing and firm performance: Evidence from Chinese private firms. *Journal of Development Economics*, 87(2), 283–299.

- Lin, J. Y. (1992). Rural reforms and agricultural growth in China. *American Economic Review*, 82(1), 34–51.
- Lin, S. (2022). *China's public finance: Reforms, challenges, and options*. Cambridge University Press.
- Long, C. (2010). Does the rights hypothesis apply to China? *Journal of Law and Economics*, 53(4), 629–650.
- Long, C. (2018). The China experience: An institutional approach. In C. Ménard & M. M. Shirley (Eds.), *A research agenda for new institutional economics* (pp. 135–142). Edward Elgar.
- Long, C., & Zhang, X. (2011). Cluster-based industrialization in China: Financing and performance. *Journal of International Economics*, 84, 112–123.
- Long, C., Xu, L. C., & Yang, J. (2022). The government, private firms, and dual-track private sector development: China's experience in two crucial decades. *Journal of Government and Economics*, 6, 1–19.
- Lu, Y., & Tao, Z. (2009). Contract enforcement and family control of business: Evidence from China. *Journal of Comparative Economics*, 37(4), 597–609.
- McGregor, R. (2019). How the state runs business in China. *The Guardian*. July, 25, 2019.
- Miao, M., Tang, D. Y., Xu, L. C., & Yan, X. (2023). Property rights, political connections, and corporate investment. *Review of Finance*, 28(2), 593–619.
- Nee, V., & Oppen, S. (2012). *Capitalism from below. Markets and institutional change in China*. Harvard University Press.
- North, D. C. (1981). *Structure and change in economic history*. Norton.
- North, D. C. (1990). *Institutions, institutional change, and economic performance*. Cambridge University Press.
- North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.
- Park, S. H., & Luo, Y. (2001). Guanxi and organizational dynamics: Organizational networking in Chinese firms. *Strategic Management Journal*, 22, 455–477.
- Peng, M. W., & Heath, P. S. (1996). The growth of the firm in planned economies in transition: Institutions, organizations, and strategic choice. *The Academy of Management Review*, 21(2), 492–528.
- Persson, P., & Zhuravskaya, E. (2016). The limits of career concerns in federalism: Evidence from China. *Journal of the European Economics Association*, 14(2), 338–374.
- Pincus, S. C. A., & Robinson, J. A. (2014). What really happened during the glorious revolution? In S. Galiani & I. Sened (Eds.), *Institutions, property rights and growth: The legacy of Douglass North* (pp. 192–222). Cambridge University Press.
- Pomeranz, K. (2000). *The great divergence: China, Europe and the making of the modern world economy*. Princeton, NJ: Princeton University Press.
- Qian, Y. (2002). *How reform worked in China* (William Davidson institute working paper 473).
- Rodríguez-Pose, A., & Zhang, M. (2020). *The cost of weak institutions for innovation in China* (C.E.P.R. Discussion papers 14399).
- Shirley, M. (2008). *Institutions and development*. Edward Elgar.
- Shirley, M., & Xu, L. C. (2001). Empirical effects of performance contracts: Evidence from China. *Journal of Law, Economics and Organization*, 17(1), 168–200.
- von Glahn, R. (2018). *The economic history of China: From Antiquity to the Nineteenth Century* (The Chinese version: 万志英, 剑桥中国经济史: 古代到19世纪. 北京: 中国人民大学出版社).
- Wang, L., & You, K. (2022). The impact of political connections on corporate tax burden: Evidence from the Chinese market. *Finance Research Letters*, 47(B), 102944.
- Wang, X., Xu, L. C., & Zhu, T. (2012). Foreign direct investment under a weak rule of law: Theory and evidence from China. *The Economics of Transition*, 20(3), 401–424.
- Wei, S.-J., Xie, Z., & Zhang, X. (2017). From 'made in China' to 'innovated in China': Necessity, Prospect, and challenges. *Journal of Economic Perspectives*, 31(1), 49–70.
- Weingast, B. R. (1997). The political foundations of democracy and the rule of law. *American Political Science Review*, 91(2), 245–263.

- Weitzman, M. L., & Xu, C. (1994). Chinese Township-Village enterprises as vaguely defined cooperatives. *Journal of Comparative Economics*, 18(2), 121–145.
- Williamson, O. E. (1985). *The economic institutions of capitalism*. The Free Press.
- Williamson, O. E. (2002). The lens of contract: Private ordering. *American Economic Review*, 92(2), 438–443.
- Wu, W., Chongfeng, W., Zhou, C., & Jun, W. (2012). Political connections, tax benefits and firm performance: Evidence from China. *Journal of Accounting and Public Policy*, 31(3), 277–300.
- Xu, L. C. (1998). How China's government and state enterprises partitioned property and control rights. *Economic Development and Cultural Change*, 46(3), 537–560.
- Xu, C. (2011). The fundamental institutions of China's reform and development. *Journal of Economic Literature*, 49(4), 1076–1151.
- Yao, L., Liang, Z., & Ye, X. (2022). *Property rights system reform and Private Enterprise financing—An empirical study based on China* (Social science research network working paper. SSRN #4307952).
- Yi, Y., Kohlin, G., & Xu, J. (2014). Property rights, tenure security and forest incentives: Evidence from China's collective forest tenure reform. *Environment and Development Economics*, 19(1), 48–73.
- Zhang, F. (2018). *The institutional evolution of China; government vs. market*. Edward Elgar.
- Zhang, Q., & Liu, M. (2019). *Revolutionary legacy, power structure, and grassroots capitalism under the red flag in China*. Cambridge University Press.
- Zhu, T. (2021). *Catching up to America: Culture, institutions, and the rise of China*. Cambridge University Press.
- 赵鼎新. (2022). 儒法国家: 中国历史新论. 杭州: 浙江大学出版社. (English version: Zhao, D. (2015). *The Confucian-Legalistic State: A new theory of Chinese history*. Oxford University Press.)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.





Jared Rubin

Contents

36.1	Introduction: What Is the “New Institutional Economics of Religion”?	954
36.2	Religious Institutions and Political Economy	955
36.3	Religion and Legal and Financial Institutions	959
36.4	Religious Institutions and Human Capital	963
36.5	Concluding Thoughts: Leveraging New Institutional Economics to Study Religion	969
	References	970

Abstract

This chapter surveys the growing literature that employs insights from New Institutional Economics to shed light on the role that religion and religious institutions have played in historical and contemporary economic development. Due to the focus on religious *institutions*, the chapter covers issues at the intersection of religion and: finance, law, political economy, and human capital. It primarily focuses on the monotheistic, Abrahamic faiths—Judaism, Christianity, and Islam—as these are the primary religions covered in the literature. It concludes with a discussion of the usefulness of New Institutional Economics frameworks for studying religion and suggestions for future research areas at the intersection of religion and institutions.

Keywords

Religion · Economic history · Institutions · Religious institutions · Legitimacy · Christianity · Islam · Protestantism · Reformation · Judaism

J. Rubin (✉)
Chapman University, Orange, CA, USA
e-mail: jrubin@chapman.edu

36.1 Introduction: What Is the “New Institutional Economics of Religion”?

In the last two decades, there has been an explosion of research in the economics of religion. This can easily be seen by scanning top economics journals, field journals, attendance at the ASREC conference (Association for the Study of Religion, Economics, and Culture), and books published at top presses. When the field was in its infancy in the 1990s, most scholars (most prominently Laurence Iannaccone, Rod Stark, and Roger Finke) were interested in using rational choice models to understand the dynamics of religious groups.¹ However, in the last two decades, a vast majority of the research in the field has focused on the economic *consequences* of religion. Most of these works by nature focus on some institutional cause or consequence of religion—the Catholic Church, political legitimacy, schools, and courts are just a few of the many phenomena studies in the literature. Such works are what I shall refer to as the “New Institutional Economics of Religion.”

There are many ways that religion affects institutions—and vice versa—and much of the literature studying these interactions is inherently historical in nature.² In many societies—especially those in which one of the Abrahamic faiths has been predominant—religion and religious authorities have historically impacted the way that political, legal, and economic institutions worked and evolved. Dating at least back to Weber’s (1905) famous *Protestant Ethic* hypothesis, several scholars have thereby sought a linkage between religion and economic growth.³ Although Weber proposed a linkage based on doctrine (specifically, the Calvinist notion of predestination serving as motivation to show that one was part of the “elect”), most of the recent literature has focused on institutions as a mediating factor connecting religion and growth. Such studies are the focus of the present survey.

Most of the papers in this literature focus on some aspect of the Abrahamic faiths: Judaism, Christianity, and Islam. While this is undoubtedly a shortcoming of this literature (and, consequently, the present survey), there are several reasons why this is the case. One is pragmatic. Most social scientists interested in studying questions of religion simply do not have the language skills or historical knowledge to study East Asian, South Asian, indigenous, or animistic religions. This limits the type of data they can collect and, even more importantly, limits their capacity to ask historically appropriate questions. A second reason is that many researchers seek to understand the contribution of religion to patterns of long-run growth. It is natural to look at Europe and its offshoots—the first places to take off—for answers. In the process, it is natural to wonder if Christianity played a role in Europe’s takeoff. Or, as

¹For more on this literature, see Iannaccone’s (1998) pioneering survey in the *Journal of Economic Literature* and Iyer’s (2016) follow-up survey.

²For a broad survey of the effect of religion on economic history (not just institutional development), see Becker et al. (2021b).

³Koyama and Rubin (2022, ch. 4) overview the linkages between culture and growth, with an emphasis on religion. Becker et al. (2024) explicitly survey the literature connecting religion and economic growth.

Weber posited, was it Protestantism that mattered?⁴ After all, the two leading economies in the world for centuries after the spread of the Reformation (the Dutch Republic and Great Britain) were both Protestant. On the contrary, did Islam play a role in the relative economic stagnation of the Middle East?⁵ If so, was it through doctrine, institutions, or some other mediator?

This survey lays out what the literature has to say on these and many other questions. Due to space constraints—and the fact that I have co-authored extensive surveys on the broader impact of religion on historical economic growth (Becker et al., 2021b, 2024)—this survey only focuses on some of the key *institutional* mechanisms through which religion has affected economic outcomes. These include political, legal, financial, and educational institutions, all of which have historically (and contemporarily) affected and been affected by religion and religious institutions.

It is worth noting what this survey will not cover. It does not address the large sociology of religion literature on religiosity, the church/sect distinction, doctrinal strictness, and so forth. This survey is meant to cover the institutional *consequences* of religion, not its causes. An unfortunate shortcoming of this and other surveys on the economics of religion (Becker et al. 2021b, 2024; Iannaccone, 1998; Iyer, 2016) is a relatively small emphasis on Eastern religion or indigenous/animistic religions. While this is clearly fruitful ground for future research, a survey can only cover what is actually researched. Where appropriate, I address relevant studies on Confucianism and Hinduism, but these are a relatively small portion of the survey. Finally, this survey only covers outcomes related to *institutions*. This means that I will not cover important features related to religion such as demography, rituals, work effort, fertility, and schism.

36.2 Religious Institutions and Political Economy

One of the most important mechanisms through which religion affects economic growth is political economy. The most straightforward mechanism is via state monopoly over religion. A pioneering paper by Barro and McCleary (2005) finds that state religions are highly persistent: nations that had state religions in the past are likely to have them in the present. But what determines when a state takes control of religion? Bentzen and Gokmen (2023) argue that state religions became institutionalized where rulers in the distant past used divine legitimation. Their results suggest that those states which used divine legitimation in the past still have more institutionalized religion today and are more likely to be autocratic. Coşgel and Miceli (2009) argue that the degree of religious competition in a society is a key determinant of whether political authorities attempt to co-opt religion and religious institutions. Their model suggests that when one religion has a monopoly—such as the medieval Catholic Church or

⁴For a survey of the rapidly growing field of research on the causes and consequences of the Protestant Reformation, see Becker et al. (2016).

⁵For a survey of the growing literature connecting Islam to a host of economic outcomes, see Kuran (2018).

parts of the Middle East in the last millennium—that religion is more effective at *legitimizing* the state, and there is therefore more incentive to co-opt it. They reason that legitimacy lowers the cost of tax collection, since subjects are more likely to voluntarily submit to taxation when it comes from a source they view as legitimate. There is a limit to how effective such efforts can be however. In a follow-up study, Coşgel et al. (2018) test this theory with empirical evidence and find support for the idea that religious competition matters for state legitimacy. However, they qualify this finding and note that it only holds when political authorities have the same religion as the population. In societies like Mughal India (Muslims ruling over a predominantly Hindu population), Ottoman Europe (Muslims ruling over a predominantly Christian population), or Russian expansion into central Asia (Christians ruling over a predominantly Muslim population), religious authorities can have a delegitimizing effect, and they tend to not be co-opted by political authorities.

What does this mean for economic growth? Why do political authorities co-opting religious authorities matter for a society's economic prospects? Rubin (2017) puts forth a theory that connects religious legitimacy and economic performance. In his framework, political authorities stay in power by either coercive means or legitimacy. The latter is an intrinsic feature by which subjects are more likely to follow an authority they view as legitimate. Rubin's framework indicates that this means that religious authorities can serve as *legitimizing agents*. In return for legitimizing the state, they expect some say in laws and policies. Hence, the identity of legitimizing agents matters for the types of policies put forth by the state. With respect to religious authorities, it is not necessarily the case that their preferences are antithetical to economic growth. Some of their preferred policies are (e.g., interest restrictions), while others favor economic growth (e.g., poor relief). The point is that religious authorities typically have preferences that are orthogonal to growth. Hence, when they have a seat at the political bargaining table, those with preferences more aligned with growth (for self-interested reasons), such as economic elites, have little say in the state's policies. According to Rubin, this is one of the most important mechanisms through which religion and religious authorities affect economic growth.⁶

Greif and Rubin (2024) expand on these insights, proposing a framework which accounts for different legitimating arrangements across time and space. In their framework, the cognitive basis for what qualifies as legitimate rule is based on a society's *legitimacy principle*, which they define as the common knowledge principle

⁶Platteau (2017) and Auriol and Platteau (2017) argue that the manner in which religious authorities are co-opted affects the likelihood of *conflict* between the political and religious establishments. They reason that political authorities must co-opt the marginal cleric in a decentralized system (e.g., Sunni Islam), whereas they must co-opt the average cleric in a centralized system (e.g., the Catholic Church), since the leader can ensure that their subordinates fall in line. Hence, Muslim rulers are more likely to have challenges from radical authorities who are outside the state-ulema alliance. Relatedly, Auriol et al. (2023) argue that co-opting the religious establishment is only optimal when the military is also co-opted. In this case, religious authorities can be bought off by slowing down reforms that counter their preferences, while co-opting the military prevents cleric-led rebellions from succeeding.

that (p. 662) “articulates why a particular individual (or organization) has legitimate authority or action-based legitimacy (including the right to legitimate an authority or her action) and how she should establish legitimate authority in some policy domain.” It follows that a society’s history and culture dictate what makes for legitimate rule and that any comparative analysis must be context-specific and historically driven.

Greif and Rubin (2024) apply this framework to Tudor and Stuart England. They argue that two key turning points were the culmination of the Wars of the Roses (in which the low-legitimacy Henry Tudor won the crown on the battlefield) and the English Reformation. Both events incentivized the Tudors to seek legitimacy from the House of Commons. In addition, the dissolution of the monasteries—the second largest wealth transfer in English history after the Norman Conquest—further contributed to the “rise of the gentry” who dominated the Commons. Heldring et al. (2021) show that because the land transferred to the gentry was used more efficiently, in the long run there was more innovation and higher yields in agriculture, a greater share of the population working outside of agriculture, and ultimately higher levels of industrialization.

Differences in legitimating arrangements have also had broader consequences for the comparative economic trajectories of societies. Rubin (2017), for instance, argues that the persistent use of religious legitimacy in the Middle East was one reason its economies stagnated in the late medieval and early modern periods. He argues that an equilibrium emerged in which religious authorities were highly effective at providing legitimacy, so they remained in the ruling coalition at the expense of the economic elite. Bisin et al. (2024) extend this intuition, arguing that such a political economy equilibrium should engender the broader population to remain more religious—as they reap benefits when religious authorities hold political power—which in turn makes religious legitimacy all the more effective. Both Rubin (2017) and Bisin et al. (2024) compare the long-run economic and political trajectories of the Middle East with those of Western Europe, where religious legitimacy began to unravel as a potent political force in the late medieval period. European political authorities turned to alternative means of legitimation—most importantly, the economic elites in their parliaments—which in turn impacted the laws and policies that states promoted. The states that most aggressively turned away from religious legitimation, especially the Protestant states such as England and the Dutch Republic, ended up with policies more favorable to commerce.

These insights have received substantial support from the economics and political science literatures. Like Rubin (2017), Kuru (2019) argues that the “ulema-state alliance” that emerged in the eleventh century had a retarding influence on Middle Eastern economic and political development. The alliance raised the prominence of religious scholars at the expense of other intellectuals (e.g., scientists) and merchants, thereby diverting resources to religious pursuits. The political alliance therefore explains the reversal of economic and scientific fortunes of the Middle East, which in the centuries after the spread of Islam had been at the scientific, technological, and economic frontier of Western Eurasia. Chaney (2013) shows how these legitimating arrangements were manifested in medieval Egypt. He finds that religious authorities gained political power when the Nile River was either too high

(meaning flooding) or too low (meaning drought). These are precisely the times when political authorities are most in need of legitimacy and should be willing to cede political power to obtain it. The timing of this transition is confirmed by Blaydes et al. (2018), who use text analysis to study advice texts given to European and Middle Eastern rulers. They find that texts emphasizing a ruler's religious bona fides began to rise significantly in the eleventh century in the Middle East, just as the religious establishment gained a source of independent power. On the other hand, European advice texts began to deemphasize religion in the late medieval period, just as other sources of legitimacy were beginning to emerge.

There were several other consequences of the use (or lack thereof) of religious legitimacy in medieval Europe and the Middle East. Rubin (2011) shows that differences in the evolution of restrictions on taking interest on loans (usury) were the result of these political economy differences. Using a similar logic to that laid out above, he argues that the Middle Eastern equilibrium in which religious authorities played a key role in legitimating rule meant that there was little demand for changes to laws that flew in the face of doctrine, and thus such changes in laws were not supplied. Coşgel et al. (2012a, b) argue that a political economy framework can likewise explain why the Ottomans did not adopt the printing press for nearly two and a half centuries after knowing of it. They argue that the press threatened the intellectual monopoly of the religious establishment, especially over the interpretation of Islamic doctrine. Since religious authorities were important legitimators of state power, sultans were incentivized to prohibit a technology that threatened their power. This was not the case in Europe, where the Church had already lost much of its legitimating power by the time Gutenberg's invention spread in the fifteenth century. Cantoni et al. (2018) employ a similar framework to gain insight into Protestant and Catholic differences following the spread of the Reformation. They argue that because Protestant rulers were less likely than Catholic rulers to seek religious legitimacy, the former needed secular administrators to run the state. They find that at Protestant universities, law became a much more popular major (relative to theology), which gave local authorities a larger group of capable administrators. Protestant rulers also built more secular buildings, rather than churches and cathedrals, because of the different political economy incentives faced by Protestants and Catholics. In South Asia, Ticku et al. (2023) find that Muslim desecrations of Hindu temples were more likely to happen during weather fluctuations. Like the logic that follows from Chaney's (2013) study of Egyptian religious power, Ticku, Shrivastava, and Iyer suggest that regime stability enabled by religious legitimization is the most likely explanation for this finding.

Religion does not only affect political economy outcomes via its capacity to legitimate political power. The very nature of the state and the type of regime that states have has historically been influenced by various aspects of religion and religious institutions. Perhaps most prominently, Woodberry (2012) finds that around *half* of the variation in democracy in Africa, Asia, Latin America, and Oceania can be accounted for by the spread of Protestant missionaries. He argues that this was an unintentional consequence of missionaries bringing education, health care, printing presses, and several other features that build strong civil societies essential for healthy democracies. On the other hand, Belloc et al. (2016) show that religion played a role in slowing the transition

from autocracy to democracy in late medieval Italy, where city states were among the first places to experiment with democratic forms of governance (at least, since classical Greece). They show that places that experienced earthquakes went through a slower transition to independent governance if they were ruled by religious authorities. They reason that earthquakes were a sign from God which gave religious authorities more political power, thereby slowing the democratic transition. Meanwhile, Gryzmala-Busse (2023) argues that the Church played a role in European state-building by playing states off each other and keeping Europe fragmented. Religious doctrine, law, and administration were therefore central to how medieval states functioned. Though unintentional, these innovations were later used by secular rulers to build larger states.

A growing literature in economics and political science also seeks to understand the rise of Islamist political parties (see Grewal et al. (2019) for an overview). One institutional cause is highlighted by Bazzi et al. (2020), who show that a rise in property confiscations in Indonesia in the 1960s led to a massive transfer of resources to *waqf* (religious trust) properties, as these were viewed as relatively safe from expropriation. These transfers increased funding for Islamic schooling and madrasas, and an unintended consequence was that, decades later, religious authorities who benefited from these transfers gained political power. They find that places most affected by *waqf* transfers had greater Islamist electoral support and higher adoption of sharia laws.

In sum, one of the most important and common ways that religion affects economic growth is via political economy. The relative efficacy of religious authorities at legitimating political rule is a central reason why this is the case, and in return for legitimation religious authorities often gain political power. There are other, less direct ways that religion and religious institutions affect economic and political outcomes, including state-building and support of religious parties.

36.3 Religion and Legal and Financial Institutions

Beyond political economy, there are several other institutional forms that have been shaped by religion and religious institutions. Foremost among these are legal and financial institutions. These institutions have played a central role in shaping the economic and political trajectories of societies, especially in the Middle East and Western Europe. This section surveys the growing literature connecting religion to the growth and evolution (or lack thereof) of legal and financial institutions.

Much of the literature at the intersection of religion and legal institutions analyzes the Islamic world. This is in part due to the pioneering work of Timur Kuran, who has spent much of the last two decades laying out the reasons that Islamic law persisted, as well as analyzing its economic and political consequences. One important question Kuran (2005) asks is why the corporation did not arise indigenously in the Middle East nor was it borrowed from Europe despite knowledge of it from Europeans who conducted business in the region. Kuran argues that there were two aspects of Islamic law that were jointly responsible. The first is that Islamic partnership law required that a partnership dissolve upon the death of any member. Partnerships could be re-formed,

but only with the approval of all heirs. The second is that Islamic inheritance law established pre-ordained heirs—who they were and what their share of the inheritance would be—depending on one's family structure. This would typically include many people, some of whom the deceased may not have been close to or even known. Kuran argues that these two features combined to discourage large partnerships. In a period in which life expectancy was low, the premature death of a partner was a non-trivial possibility. This would mean that the partnership would be split up among many heirs, any one (or two) of whom could refuse to re-form the partnership, leading to a potentially costly dissolution. This disincentivized Muslims from forming large partnerships in the first place: with every partner added, the probability of premature dissolution increased. Over time, the fact that Islamic partnerships remained small made adopting the European corporation all the less likely. The various institutional elements that are essential to the type of big business performed by corporations—double entry bookkeeping, tradable shares, managerial hierarchies—never emerged in a context where partnerships remained small. Hence, merely knowing about the corporation was not enough to allow for its adoption, as the associated features that made it work were also not present.

Kuran (2001) argues that one possible Islamic alternative to the corporation was the *waqf*. The *waqf* was a trust that was supposed to fund some non-movable public good in perpetuity (e.g., a drinking fountain, bridge, or madrasa), although by the sixteenth century *waqfs* of movable goods (cash) became popular. Kuran reasons that the *waqf* solved one problem that partnerships did not: long life. In theory, therefore, the *waqf* could have been transformed into an organizational form that served a similar purpose to the business corporation. However, Kuran argues that there were features of the *waqf* that made it unsuitable for this purpose. Most importantly, by law the *waqf* was inflexible. It was extraordinarily difficult to change the purpose of a *waqf* from that dictated by its founder. Therefore, the *waqf* did not have the flexibility needed to respond to changes in market conditions, limiting its use in business settings.

These insights raise the question: why did Islamic law persist in the first place? Kuran's (2011) magnum opus lays out the theory that the early dominance of Islamic law stifled *demand* for change to the law. With little demand for change, why should political authorities seek other sources of law, except when it suited their self-interest (e.g., in the Ottoman Empire a form of secular law called *kanun* had purview over issues such as taxation)? The lack of the corporate form noted above is a case in point. Islamic law of partnerships remained relatively stagnant not out of some lack of foresight in the supply of legal innovation, but due to the lack of demand for changes to the law incentivized by the very contours of the legal code itself.

In other works, Kuran argues that other aspects of Islamic legal and political institutions had a retarding effect on economic growth. Kuran (2004b) points to the Ottoman dictate that religious minorities could conduct business using their own legal systems (e.g., Greek), so long as they were not conducting business with a Muslim. Until the seventeenth century or so, this did not have much of an effect on business dealings, as European legal systems were not much more advanced than Islamic law. But as European business became bigger and more complicated, an associated set of legal codes arose. When this happened, non-Muslims had access to

more advanced organizational forms and financial instruments than Muslims did, thereby placing a large part of the population at a disadvantage.

Another aspect of Islamic law that impacted economic growth was that it was *biased* toward men, Muslims, and elites. Kuran and Lustig (2012) show that in the Ottoman Empire, these groups were much more likely to win court cases, all else equal. For men and Muslims, this was largely due to aspects of Islamic law that favored these groups, such as the number of witnesses needed to successfully make a case against a defendant. Elites were favored due to the way that judges were chosen—ruling against an elite could have serious negative career consequences for a judge. Kuran and Rubin (2018) find that men, Muslims, and elites paid a price for these privileges: they paid 1.9–3.4 percentage points surcharge on loans. Kuran and Rubin argue that the fact that these groups were favored by courts made it riskier to lend to them, since they could abscond with the funds and not face consequences. The surcharge was the price they paid for this additional risk. This may have had a significant dampening effect on economic growth, as it was precisely these capital-rich individuals who were in the best position to invest in business enterprises.

Another aspect of legal systems, both in the Islamic world and in Europe, that influenced economic growth were laws on taking interest on loans and how those laws were enforced. As discussed in the previous section, Rubin (2011, 2017) notes that doctrine and secular laws on taking interest on loans diverged between the Islamic world and Western Christendom in the late medieval period. He argues that this was due to a political equilibrium in which religious legitimacy remained a potent force in the Islamic world while it was undermined in many parts of late medieval Europe. This meant that European rulers had greater incentive to permit moderate interest, and it followed that interest rate caps became commonplace throughout medieval Europe. Eventually, the Church responded to what had become common practice by finding ways of justifying various transactions once considered usurious. Meanwhile, in the Islamic world, ruses that simulated interest-bearing loans were permitted from at least the early ninth century, and probably before. But open lending at interest was never (and is still not) permitted by the letter of the law.

Did these restrictions have any impact on the financial development of the regions? Several papers suggest that it did, often for unforeseeable reasons. Rubin (2010) argues that bills of exchange evolved differently in the two regions, and this had a major impact on the development of financial institutions. In Europe, bills of exchange were credit instruments that could be purchased in one region and remitted in another. They included a currency exchange, which allowed financiers to profit from differences in exchange rates. This incentivized financiers to expand their operations inter-regionally to have the capacity to engage on both sides of the exchange. The famous Medici Bank is an example of an organization that spread throughout Europe in part to engage in such transactions. In the Islamic world, a similar credit instrument (the *suftaja*) was widely used but did not contain a currency exchange, which was banned by Islamic jurists because it could be used to make close to a risk-free profit. This meant that Islamic financiers did not use the *suftaja* as a credit instrument, and there was no incentive to expand inter-regionally to engage in trading these instruments.

Interest restrictions had long-run consequences for financial development. In the fifteenth century, the Church accepted the validity of Monte di Pietà, which was an early “bank” that lent to the poor at moderate interest. Because it was supposed to make no profit and had a pious mission, it was a means of extending credit that skirted usury restrictions. It soon took on other features we now view as central to banking, such as taking on deposits and remunerating deposits. Pascali (2016) finds that those cities which had montes soon after they arose *still* have much greater access to credit, bank density, per capita GDP, and productivity. On the other hand, in the Ottoman Empire, interest was formally prohibited, even if workarounds were straightforward. This meant, however, that no banking system arose indigenously, given that this would have required open lending at interest. Grosjean (2011) shows that this delayed institutional formation had persistent effects on access to finance in places once under Ottoman rule. In a study of six southeastern European countries that were at one time partially under Ottoman rule, Grosjean finds that the parts of the countries the Ottomans ruled have 10% less bank penetration.⁷ One reason in addition to delayed institutional formation that may be the case were cultural norms associated with saving. Walker (2020) shows in a field experiment along an old Ottoman-Habsburg border in Romania that those living on the old Ottoman side have less access to banks and save significantly less.

Moving outside of the Abrahamic faiths, traditional Confucian religion may have affected the emergence and spread of banking in China. Chen et al. (2022) note that Confucianism encouraged kin groups to support each other, thereby providing informal insurance. They argue that this discouraged the growth of an indigenous banking system, since there was little demand for funds to either start projects or for consumption after a negative shock. Lending therefore remained personal, and impersonal finance did not arise in China until after it was imported from the West.

A final sector where religion and financial institutions have had an important intersection is the rise of Islamic banking. Now around a \$4 trillion industry, Islamic banking has many elements that are similar to conventional, Western banks, though there are many small differences. Interest-bearing loans are replaced with similar debt contracts where risk is shared, and “sinful” transactions (e.g., those involving pork, alcohol, or gambling) are prohibited (Khan, 2010). Kuran (2004a), and Khan (2010) argue that Islamic banking is mostly Western banking with an Islamic veneer; they achieve the same goals through very similar means, though the latter satisfies some spiritual criteria. This view is not universally shared. Beck et al. (2013) show that Islamic banks are better capitalized and have higher asset quality. As a result, Islamic banks survived the 2008 crash in a better position than most Western-style banks. Likewise, Baele et al. (2014) show that Islamic loans have a much lower default rate than conventional loans. This difference is exacerbated during Ramadan, suggesting there is either greater spiritual or social pressure to pay back loans made by Islamic banks.

⁷This result comes from a cross-country analysis. In within-country specifications (i.e., including country fixed effects), regions that were under Ottoman rule have 4% less bank penetration.

In sum, religion has affected the development and evolution of financial and legal institutions in several ways. This is most famously the case in the Islamic world, where Islamic law has pervaded many aspects of life (including commerce) since the consolidation of the clerical elite in the ninth century. There have been several intended and unintended consequences of these arrangements, including the long absence of the corporation in the Islamic world, the persistence of biased courts, and the spread (or lack thereof) of financial institutions capable of dealing in interregional exchange. Although the connection between religion and financial development is often indirect—working through legal or political institutions—it has played a key role in shaping the economic trajectories of several societies in the Islamic and Christian worlds.

36.4 Religious Institutions and Human Capital

Perhaps the most active line of inquiry connecting religion to economic and political outcomes views *human capital* as a central mechanism linking the two. Religion has affected human capital accumulation through several mechanisms historically, and it continues to do so in the present. The most important mechanism has been through the supply of schools. This, however, is a two-edged sword. Where publicly-provided schools have been historically missing, religious schools have provided an important service that increases the overall human capital of the population. On the other hand, where religious schools compete with secular schools, the former often teaches skills that do not have much applicability in the labor market.

A large literature has emerged linking the spread of the Protestant Reformation to human capital accumulation. This is largely due to the pioneering work of Becker and Woessmann (2009), who sought an alternative explanation to Weber (1904/05) for why Protestant areas seemed to historically be better off than Catholic areas. Moving away from Weber's preferred explanation centered on doctrinal differences, Becker and Woessmann argue that Protestants had an economic advantage over Catholics in large parts of Europe since the spread of the Reformation due to their human capital advantage. Luther wanted his followers to be able to read and interpret the Bible on their own. He therefore demanded that Protestant rulers set up schools, and many local denominations did the same. The maintenance and operation of schools was institutionalized early in the Reformation via local school and church ordinances (Dittmar & Meisenzahl, 2020). Although Luther's intent was not to have his flock obtain economically-relevant human capital, this was an unintended consequence of the early Protestant program. The spread of schooling gave Protestants a human capital advantage that persisted over time and was reflected in the nineteenth century by Protestants being more likely to have white collar jobs requiring some level of education. Becker and Woessmann (2008) point out that Luther's call to read the Bible applied to both men and women, and therefore Luther demanded that each town should have both a boy's school and a girl's school. In a period in which few females had the ability to read, this was a revolutionary idea. Becker and Woessmann (2008) find that this had a persistent effect on gender differences in education.

In nineteenth-century Prussia, the gender gap in elementary school enrollment was much lower in Protestant regions than in Catholic ones.

Human capital differences between Protestants and Catholics extended beyond Prussia (the setting of Becker and Woessmann's studies) and extended beyond just literacy. Boppert et al. (2014) test Becker and Woessmann's hypothesis in the study of Swiss pedagogical exams taken at the end of the nineteenth century. They find that students in Protestant areas did in fact have better literacy scores, and they also had higher scores in writing, numeracy, and history. In a related study, Boppert et al. (2013) find that the Swiss Protestant advantage was particularly stark in conservative regions (i.e., those that voted in favor of conservative reforms). They suggest that this may be due to either differences in educational spending or the possibility that cultural traits (progressive versus conservative) are endogenous to a society's education level. Izumi et al. (2023) also provide nuance to the link between Protestantism and human capital accumulation. They find that the spread of Protestant churches in Korea increased literacy for women in the short run but not in the long run. This is possibly due to Protestant churches providing educational opportunities in the early twentieth century that were eventually substituted for by public education in the latter half of the century.

The Protestant educational advantage had several other consequences. A key insight of unified growth theory (Galor, 2011) is that there is a tradeoff between the quantity of children one has and their "quality" (as measured in large part by human capital investments). Given the importance placed on human capital by Protestants, it stands to reason that Protestant areas may have gone through a demographic transition earlier. Becker et al. (2012) show that such a tradeoff was present as early as 1816 in Protestant areas of Prussia—this is, prior to the broader fertility transition that occurred in many parts of Western Europe in the mid-to-late nineteenth century. Likewise, Becker et al. (2010) show that a lower price of education in mid-nineteenth-century Prussia decreased fertility. In fact, education in 1849 predicts a fertility transition a generation later, in 1880–1905. The Protestant educational advantage also had a feedback effect onto religiosity itself. Becker et al. (2017) find that reductions in religious schooling (relative to secular education) decreased religious and church attendance in the next generation, which further lowered demand for religious education. Becker, Nagler, and Woessmann show this result for early twentieth-century Prussia, while Arold et al. (2022) show a similar result for late twentieth-century Germany.

Relatedly, the most important human capital-inducing innovation of the last millennium—the movable type printing press—had a significant impact on religion, religious institutions, and subsequent human capital development. Rubin (2014) shows that the spread of printing was strongly associated with the subsequent spread of the Reformation. This study brings empirical evidence to an association that was long believed to be true by historians and, indeed, Luther himself. The early Protestants were particularly adept at using the newly invented press for propaganda, allowing them to spread their controversial ideas quickly enough to not be suppressed by the Church. Early Protestant printing was also at the vanguard of vernacularization—Luther's translation of the Bible was responsible for much of

what would become written German. Binzel et al. (2023) argue that this vernacularization played a key role in spreading knowledge to the wider population. They show that the spread of the Reformation was strongly associated with printing in the vernacular, which itself was subsequently associated with greater numbers of famous creative people and innovators.

The link between human capital and religion extends well beyond Protestantism. A recent literature has emerged trying to understand the role that Islam and Islamic institutions have played in human capital development in the Muslim world. This is a particularly important and vexing issue because the Islamic world was by far the human capital leader of Western Eurasia for centuries following the spread of Islam. Chaney (2023) provides insight into this reversal of fortunes. He has collected a massive data set of all known publications from the Islamic world and Europe since not long after the spread of Islam. Chaney finds a massive drop in scientific publications and a similar rise in religious publications beginning in the eleventh and twelfth centuries. This timing is consistent with the “madrasa movement” noted previously in this chapter (Kuru, 2019), in which madrasas became important cultural and political centers, thereby increasing the political power of the religious elite. Chaney’s results suggest that this had a negative impact on economically relevant human capital, and it may be one of the driving factors underlying the broader economic reversal of fortunes between the Middle East and Western Europe.

Following the madrasa movement, religious schools have played a large role in dictating the type of human capital that Muslims attained. Historically, most Muslim schools focused predominantly on religious learning, only indirectly teaching skills that were economically valuable. Some studies have found that this had consequences for literacy and associated human capital development. In colonial India, Chaudhary and Rubin (2011) find that districts with a more recent collapse of Muslim political rule have lower Muslim literacy rates (though not lower Hindu literacy). Their analysis suggests that Muslim religious authorities were more powerful in these districts, and thus Islamic schooling was better funded. Since Islamic schools were typically worse at teaching basic secular human capital, fewer children learned to read in these districts. In a follow-up study in the Indian Princely States (those states not under the direct control of the British), Chaudhary and Rubin (2016) find that Hindus living in Hindu-ruled states had higher literacy than those who lived in Muslim-ruled states (each state had a prince who had a known religious affiliation), but there were no literacy differences for Muslims across states with different rule. They reason that Muslim rulers were less likely to fund public schooling, which harmed the ability of students of all religions to read, since the rulers gained legitimacy from Muslim religious authorities. This was not true in Hindu-ruled states, where rulers were more likely to invest in public education. A similar finding in the Spanish context is presented by Cinnirella et al. (2023), who show that places that were under longer periods of Muslim rule—various Muslim empires ruled parts of the Iberian Peninsula from 711 to 1492—had much lower literacy centuries after Muslim rule on the peninsula ended. They suggest a demand-side explanation, whereby Muslim rule discouraged autonomous government and an independent

merchant class, both of which have historically been large drivers of demand for education.

In the twentieth century, religious reforms have aimed at reducing the human capital gap in Muslim-majority countries. For instance, Saleh (2016) studies Egyptian reforms from the early 1950s that aimed to reduce the education gap between Muslims and (Coptic) Christians, which was a key source of the latter being over-represented in higher paying occupations. The reforms permitted graduates of religious primary schools (*kuttabs*) to attend higher-level schools, which had previously been only open to those graduating from secular schools. Saleh finds that this reform did in fact reduce the Christian-Muslim gap in labor market outcomes. In Indonesia, the government attempted a reform in the 1970s to expand public education at the expense of religious education. Bazzi et al. (2022) find an unanticipated consequence of this policy: it massively increased the demand for secondary schooling, which religious schools were in the best position to supply. The policy therefore had the opposite of the intended effect. Instead of increasing national identity at the expense of religious identity, it made religious identity more salient and religious authorities more politically powerful.

There has also been a close historical connection between Judaism and human capital. The pioneering work by Botticini and Eckstein (2005, 2007) has placed human capital accumulation at the center of Jewish history and economic history. They argue that the Jewish educational advantage stemmed from an event that happened nearly 2000 years ago, when the Romans destroyed the Second Temple in Jerusalem in 70 CE. This shifted the focus of Jewish leaders from the cult of sacrifice to emphasizing literacy. Over time, cultural norms favoring obtaining literacy emerged among Jews, and they were therefore in prime position to take advantage of urban jobs requiring human capital after the Muslim conquests of the seventh and eighth centuries gave them such an opportunity (Botticini & Eckstein, 2005). Hence, from this period onwards, Jews transitioned out of agriculture and into skilled urban professions. Botticini and Eckstein (2007) explain the persistence of Jewish occupational specialization with a model centered on the feedbacks between human capital and occupational choice. Since Jewish cultural norms implied that Jews received utility from educating their children, they were more likely to invest in education. This gave Jewish merchants—who could leverage literacy more than most other professions at the time—an advantage that persisted for centuries.

The Jewish educational advantage has been shown to arise across many historical cases. It is also true in the present; the average years of schooling is much greater for Jews than it is for Christians, Muslims, Buddhists, or Hindus. However, Abramitzky and Halaburda (2020) warn that the appearance of a Jewish educational advantage may merely be a compositional effect. In a study of Jews and Catholics in interwar Poland, they find that urban Jews had similar literacy rates to urban Catholics, and the same was true for rural Jews and Catholics. However, because Jews predominantly lived in urban areas (where literacy rates were higher across the board), the overall Jewish literacy rate was higher than the Catholic literacy rate.

Several recent papers have also proposed a link between practices of the Catholic Church and human capital accumulation. In the industrial period, this link has typically been found to be negative. In a study of France during the Second Industrial Revolution of the late nineteenth century, Squicciarini (2020) finds that religious schools were an impediment to skill acquisition just as the returns to those skills were increasing in the wake of the “second industrialization.” Since a more skilled labor force was required to work the new machinery, French secular schools responded by changing the curriculum to be more amenable to industry. Yet, Catholic schools were much slower to adopt a technical curriculum and, in some cases, *increased* the amount of religious education. Squicciarini finds that this carried negative long-term consequences for industrialization and employment in regions of France with many religious schools. In a related study, Lecce et al. (2021) find that scientists were less likely to be born in highly religious parts of nineteenth-century France. They provide evidence that people born in these regions acquired less scientific human capital early in life because the scientific curriculum was weaker.

Studies of earlier periods show a negative relationship between the Catholic Church and human capital accumulation. If Protestant areas were associated with high human capital accumulation, one possible reason why Catholic areas were not was the Church’s response during the Counter-Reformation. The Church censored hundreds of books deemed to be out of line with Catholic doctrine and threatened authors who produced such works. Becker et al. (2021a) show that this censorship had a strong negative effect on book production, while cities that defied the Church and printed prohibited books attracted famous people, including scientists.

Not all Catholic organizations were associated with low human capital. An earlier, pre-Reformation linkage between the Catholic Church and human capital is reported by Akçomak et al. (2016), who study the religious community called the Brethren of the Common Life. Founded in the Dutch city of Deventer, one of the Brethren’s missions was that each Christian be able to read the bible and other religious texts. They therefore undertook several initiatives, including founding schools and producing books, that may have affected local human capital accumulation. Akçomak, Webbink, and ter Weel show that the Brethren were effective: places where they produced books in the latter half of the fifteenth century had greater literacy in 1600, and cities with a Brethren presence grew more in the fifteenth century.

A final linkage studied extensively between religion and human capital accumulation is via the work of Christian missionaries. This was not necessarily an intentional outcome of missionary activity. Missionaries above all else sought to Christianize native populations, particularly in Latin America, sub-Saharan Africa, South Asia, and East Asia. They attracted converts (in part) by providing public goods like schools and health clinics. Teaching native populations to read was viewed as particularly important, as it would allow them to read the Bible. An unintended side effect of these efforts was that native populations gained human capital that was economically useful. To be clear, this is not to say that these efforts meant that colonization was a good thing for native populations. It was often brutal, extractive, and tore the social fabric of the

colonized regions. It is possible, however, that the unintended effects of missionary activity provide a silver lining to colonialism's worst excesses.⁸

The basic link between historical missionary activity and education was established by Gallego and Woodberry (2010). They show that the activities of Protestant missionaries in the past are associated with greater schooling in the present. The same is true for Catholic missionaries only where they were in competition with Protestant missionaries and thus were incentivized to provide schools. One reason that historical Protestant missions are associated with better human capital outcomes in the present, in addition to supplying schools, is that Protestant missionaries often brought printing presses to their missions. This was done in the spirit of Luther's desire that every one of his flock know how to read the Bible. Cagé and Rueda (2016) study the spread of printing presses among missions in sub-Saharan Africa and find that the link between Protestant missionary activity and educational outcomes is mediated by the presence of a printing press. Outside of Africa, several studies have shown linkages between Protestant missionary activity and human capital accumulation. Bai and Kung (2015) study the consequences of Protestant missionaries erecting schools and hospitals in China. They show that places with such missionary activity grew faster than those that did not. Protestant missionaries have also been shown to have long-run positive effects on educational attainment in India (Lankina & Getachew, 2012), and this is especially true for women (Calvi et al., 2022).

In some contexts, Catholic missionaries also played a role in fostering human capital. In colonial Benin, Wantchekon et al. (2015) study the effect of regional school production by Catholic missionaries. They show that students attending these schools ended up with higher standards of living, and these effects were present in the following generation as well. A similar effect of Catholic missionaries is found in India by Castelló-Climent et al. (2018). They show that Catholic missionaries opened colleges, and where they did so there is a positive effect on education in the twenty-first century. Perhaps the largest effects of Catholic missionaries came from Jesuit missions. Jesuits were one of the major Catholic bulwarks against the Reformation, and one way they achieved this goal was by educating children. Valencia (2019) studied the effects of these efforts in the Guaraní area in modern-day Argentina, Brazil, and Paraguay. In this area, the Jesuits played a major role in educating the indigenous population while they were permitted to do missionary work between 1609 and 1767. Yet, despite being kicked out 250 years ago, places that had a Jesuit presence still do better today on a host of human capital-related outcomes (relative to places without missions). Valencia Caicedo finds that places with Jesuit presence also have higher incomes today, and this can mostly be explained via human capital differences. Ma (2021) finds a similar effect of Jesuit presence in China, where Jesuits acted as a conduit for passing technical knowledge between Europe and China in the seventeenth and eighteenth centuries.

⁸For an extended discussion on this topic, see Koyama and Rubin (2022, ch. 6).

In sum, one of the key mechanisms through which religion has historically affected economic, political, and institutional development is via human capital accumulation. In the case of Protestantism and Judaism, the connection is mostly a positive one, as unique events in the histories of these religions promoted cultural traits and institutional changes favoring human capital. In Islam and Catholicism, the connection is mixed: in some historical settings the connection has been positive and in others negative.⁹

36.5 Concluding Thoughts: Leveraging New Institutional Economics to Study Religion

This survey has discussed some of the most important ways that religion affects economic and political development: political economy, legal and financial institutions, and human capital accumulation. Common to all these mediating factors is the important role that *institutions* play. Although doctrine sometimes matters (e.g., religious restrictions on usury), institutions tend to play the most important role in affecting outcomes of interest to social scientists. Religious institutions interact with economic, legal, and political institutions to shape the ways that the latter institutions influence behavior. Such influences are context-specific and differ across time, space, and religious denomination. While most of the work highlighted in this survey has focused on the three major Abrahamic faiths (Christianity, Islam, and Judaism)—this is due to the nature of the literature, not a predilection of this survey’s author—the scant insights in the literature from Eastern religions suggests that such institutional interactions are present there, as well.

Whether the authors of many of the studies covered in this survey know it, they have leveraged insights from New Institutional Economics to study various political, legal, and economic consequences of religion, religious behavior, and religious institutions. These studies are merely the tip of the iceberg. Several studies in New Institutional Economics provide frameworks for conceptualizing how institutions shape and constrain both macroeconomic and microeconomic behavior in a variety of historical and contemporary settings. While it is undoubtedly true that religion has been among the most important social forces historically shaping individual behavior, it has only been in the last decade or two that practitioners of New Institutional Economics have taken religion seriously.

This is good news for scholars familiar with New Institutional Economics frameworks and methodologies. Some of the important questions that remain either unexplored or underexplored include: How do Buddhist and Confucian institutions affect Chinese economic development, both historically and in the twentieth/twenty-

⁹There is also evidence that Confucianism has affected modern day human capital accumulation. Chen et al. (2020) find that parts of China that had more *jinshi*—the highest level of Confucian bureaucrat—during the Imperial period have greater levels of schooling today. There are several possible causes of this finding, including cultural transmission and the supply of schools.

first centuries? How have caste restrictions affected trade, education, and industrial development? How have conversions to Evangelical Protestant faiths affected the economic and political prospects of the parts of the developing world where it has taken hold? Will Islamic politics slow Middle Eastern economic growth in the twenty-first century, and what role will Islam play in the economic potential of the region if/when oil revenues slow down? As world population growth continues to decline, what role will religion and religious institutions play in shaping worldwide demographic patterns? Access to new data via digitization, new textual analysis techniques, and advances in computing should allow researchers to answer questions they had either previously considered unanswerable or had simply never considered in the first place. It is an exciting time to be studying religion from the social science perspective, and scholars of the New Institutional Economics tradition should be front and center in the major advances made over the foreseeable future.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Abramitzky, R., & Halaburda, H. (2020). Were Jews in interwar Poland more educated? *Journal of Demographic Economics*, 86(3), 291–304.
- Akçomak, İ. S., Webbink, D., & ter Weel, B. (2016). Why did The Netherlands develop so early? The legacy of the brethren of the common life. *Economic Journal*, 126(593), 821–860.
- Arold, B. W., Woessmann, L., & Zierow, L. (2022). *Can schools change religious attitudes? Evidence from German state reforms of compulsory religious education* (CESifo working paper 9504).
- Auriol, E., & Platteau, J.-P. (2017). Religious co-option in autocracy: A theory inspired by history. *Journal of Development Economics*, 127, 395–412.
- Auriol, E., Platteau, J.-P., & Verdier, T. (2023). The Quran and the sword. *Journal of the European Economic Association*, 21(5), 1772–1820.
- Baele, L., Farooq, M., & Ongena, S. (2014). Of religion and redemption: Evidence from default on Islamic loans. *Journal of Banking and Finance*, 44, 141–159.
- Bai, Y., & Kung, J. K.-s. (2015). Diffusing knowledge while spreading God's message: Protestantism and economic prosperity in China, 1840–1920. *Journal of the European Economic Association*, 13(4), 669–698.
- Barro, R. J., & McCleary, R. M. (2005). Which countries have state religions? *Quarterly Journal of Economics*, 120(4), 1331–1370.
- Bazzi, S., Koehler-Derrick, G., & Marx, B. (2020). The institutional foundations of religious politics: Evidence from Indonesia. *Quarterly Journal of Economics*, 135(2), 845–911.
- Bazzi, S., Hilmy, M., & Marx, B. (2022). *Religion, education, and development* (Working paper).
- Beck, T., Demirgüç-Kunt, A., & Merrouche, O. (2013). Islamic vs. conventional banking: Business model, efficiency and stability. *Journal of Banking and Finance*, 37(2), 433–447.
- Becker, S. O., & Woessmann, L. (2008). Luther and the girls: Religious denomination and the female education gap in 19th century Prussia. *Scandinavian Journal of Economics*, 110(4), 777–805.
- Becker, S. O., & Woessmann, L. (2009). Was Weber wrong? A human capital theory of Protestant economic history. *Quarterly Journal of Economics*, 124(2), 531–596.

- Becker, S. O., Cinnirella, F., & Woessmann, L. (2010). The trade-off between fertility and education: Evidence from before the demographic transition. *Journal of Economic Growth*, 15(3), 177–204.
- Becker, S. O., Cinnirella, F., & Woessmann, L. (2012). The effect of investment in children's education on fertility in 1816 Prussia. *Cliometrica*, 6(1), 29–44.
- Becker, S. O., Pfaff, S., & Rubin, J. (2016). Causes and consequences of the Protestant reformation. *Explorations in Economic History*, 62, 1–25.
- Becker, S. O., Nagler, M., & Woessmann, L. (2017). Education and religious participation: City-level evidence from Germany's secularization period 1890–1930. *Journal of Economic Growth*, 22, 273–311.
- Becker, S. O., Pino, F., & Vidal-Robert, J. (2021a). *Freedom of the press? Catholic censorship during the counter-reformation* (CEPR discussion paper 16092).
- Becker, S. O., Rubin, J., & Woessmann, L. (2021b). Chapter 20: Religion in economic history: A survey. In A. Bisin & G. Federico (Eds.), *Handbook of historical economics*. North Holland.
- Becker, S. O., Rubin, J., & Woessmann, L. (2024, forthcoming). Religion and growth. *Journal of Economic Literature*, 62(3), 1094–1142.
- Belloc, M., Drago, F., & Galbiati, R. (2016). Earthquakes, religion, and transition to Self-Government in Italian Cities. *Quarterly Journal of Economics*, 131(4), 1875–1926.
- Bentzen, J., & Gokmen, G. (2023). The power of religion. *Journal of Economic Growth*, 28(1), 45–78.
- Binzel, C., Link, A., & Ramachandran, R. (2023). *The Protestant reformation and language choice in the holy Roman empire* (CEPR discussion paper 17818).
- Bisin, A., Rubin, J., Seror, A., & Verdier, T. (2024, forthcoming). Culture, institutions, and the long divergence. *Journal of Economic Growth*, 29(1), 1–40.
- Blaydes, L., Grimmer, J., & McQueen, A. (2018). Mirrors for princes and sultans: Advice on the art of governance in the medieval Christian and Islamic worlds. *Journal of Politics*, 80(4), 1150–1167.
- Boppart, T., Falkinger, J., Grossmann, V., Woitek, U., & Wüthrich, G. (2013). Under which conditions does religion affect educational outcomes? *Explorations in Economic History*, 50(2), 242–266.
- Boppart, T., Falkinger, J., & Grossmann, V. (2014). Protestantism and education: Reading (the bible) and other skills. *Economic Inquiry*, 52(2), 874–895.
- Botticini, M., & Eckstein, Z. (2005). Jewish occupational selection: Education, restrictions, or minorities? *Journal of Economic History*, 65(4), 922–948.
- Botticini, M., & Eckstein, Z. (2007). From farmers to merchants, conversions and diaspora: Human capital and Jewish history. *Journal of the European Economic Association*, 5(5), 885–926.
- Cagé, J., & Rueda, V. (2016). The long-term effects of the printing press in sub-Saharan Africa. *American Economic Journal: Applied Economics*, 8(3), 69–99.
- Calvi, R., Hoehn-Velasco, L., & Mantovanelli, F. G. (2022). The Protestant legacy: Missions, gender, and human Capital in India. *Journal of Human Resources*, 57(6), 1946–1980.
- Cantoni, D., Dittmar, J., & Yuchtman, N. (2018). Religious competition and reallocation: The political economy of secularization in the Protestant reformation. *Quarterly Journal of Economics*, 133(4), 2037–2096.
- Castelló-Climent, A., Chaudhary, L., & Mukhopadhyay, A. (2018). Higher education and prosperity: From Catholic missionaries to luminosity in India. *Economic Journal*, 128(616), 3039–3075.
- Chaney, E. (2013). Revolt on the Nile: Economic shocks, religion, and political power. *Econometrica*, 81(5), 2033–2053.
- Chaney, E. (2023). *Religion and the rise and fall of Islamic science* (Working paper).
- Chaudhary, L., & Rubin, J. (2011). Reading, writing, and religion: Institutions and human capital formation. *Journal of Comparative Economics*, 39(1), 17–33.
- Chaudhary, L., & Rubin, J. (2016). Religious identity and the provision of public goods: Evidence from the Indian princely states. *Journal of Comparative Economics*, 44(3), 461–483.

- Chen, T., Kung, J. K.-s., & Ma, C. (2020). Long live Keju! The persistent effects of China's civil examination system. *Economic Journal*, 130(631), 2030–2064.
- Chen, Z., Ma, C., & Sinclair, A. (2022). Banking on the Confucian clan: Why China developed financial markets so late? *Economic Journal*, 132(644), 1378–1413.
- Cinnirella, F., Naghavi, A., & Prarolo, G. (2023). Islam and human capital in historical Spain. *Journal of Economic Growth*, 28(2), 225–257.
- Coşgel, M. M., & Miceli, T. J. (2009). State and religion. *Journal of Comparative Economics*, 37, 402–416.
- Coşgel, M. M., Miceli, T. J., & Rubin, J. (2012a). Political legitimacy and technology adoption. *Journal of Institutional and Theoretical Economics*, 168(3), 339–361.
- Coşgel, M. M., Miceli, T. J., & Rubin, J. (2012b). The political economy of mass printing: Legitimacy, revolt, and technology change in the ottoman empire. *Journal of Comparative Economics*, 40(3), 357–371.
- Coşgel, M. M., Histen, M., Miceli, T. J., & Yıldırım, S. (2018). State and religion over time. *Journal of Comparative Economics*, 46(1), 20–34.
- Dittmar, J., & Meisenzahl, R. (2020). Public goods institutions, human capital, and growth: Evidence from German history. *Review of Economic Studies*, 87(2), 959–996.
- Gallego, F. A., & Woodberry, R. (2010). Christian missionaries and education in former African colonies: How competition mattered. *Journal of African Economics*, 19(3), 294–329.
- Galor, O. (2011). *Unified growth theory*. Princeton University Press.
- Greif, A., & Rubin, J. (2024, forthcoming). Endogenous political legitimacy: The Tudor roots of England's constitutional governance. *Journal of Economic History*, 84(3), 655–689.
- Grewal, S., Jamal, A. A., Masoud, T., & Nugent, E. R. (2019). Poverty and divine rewards: The electoral advantage of Islamist political parties. *American Journal of Political Science*, 63(4), 859–874.
- Grosjean, P. (2011). Long term institutional persistence: Ottoman rule and financial development in the regions of Europe. *Journal of Comparative Economics*, 39, 1–16.
- Gryzmala-Busse, A. (2023). *Sacred foundations: The religious and medieval roots of the European state*. Princeton University Press.
- Heldring, L., Robinson, J. A., & Vollmer, S. (2021). The long-run impact of the dissolution of the English monasteries. *Quarterly Journal of Economics*, 136(4), 2093–2145.
- Iannaccone, L. R. (1998). Introduction to the economics of religion. *Journal of Economic Literature*, 36(3), 1465–1495.
- Iyer, S. (2016). The new economics of religion. *Journal of Economic Literature*, 54(2), 395–441.
- Izumi, Y., Park, S., & Yang, H. (2023). The effects of South Korean Protestantism on human capital and female empowerment, 1930–2010. *Journal of Comparative Economics*, 51(2), 422–438.
- Khan, F. (2010). How 'Islamic' is Islamic banking? *Journal of Economic Behavior and Organization*, 76(3), 805–820.
- Koyama, M., & Rubin, J. (2022). *How the world became rich: The historical origins of economic growth*. Polity Press.
- Kuran, T. (2001). The provision of public goods under Islamic law: Origins, impact, and limitations of the Waqf system. *Law and Society Review*, 35(4), 841–897.
- Kuran, T. (2004a). *Islam and mammon: The economic predicaments of Islamism*. Princeton University Press.
- Kuran, T. (2004b). The economic ascent of the Middle East's religious minorities: The role of Islamic legal pluralism. *The Journal of Legal Studies*, 33, 475–515.
- Kuran, T. (2005). The absence of the Corporation in Islamic law: Origins and persistence. *American Journal of Comparative Law*, 53, 785–834.
- Kuran, T. (2011). *The long divergence: How Islamic Law Held Back the Middle East*. Princeton University Press.
- Kuran, T. (2018). Islam and economic performance: Historical and contemporary links. *Journal of Economic Literature*, 56(4), 1292–1359.

- Kuran, T., & Lustig, S. (2012). Judicial biases in ottoman Istanbul: Islamic justice and its compatibility with modern economic life. *Journal of Law and Economics*, 55(3), 631–666.
- Kuran, T., & Rubin, J. (2018). The financial power of the powerless: Socio-economic status and interest rates under partial rule of law. *Economic Journal*, 128(609), 758–796.
- Kuru, A. T. (2019). *Islam, authoritarianism, and underdevelopment: A global and historical comparison*. Cambridge University Press.
- Lankina, T., & Getachew, L. (2012). Mission or empire, word or sword? The human capital legacy in postcolonial democratic development. *American Journal of Political Science*, 56(2), 465–483.
- Lecce, G., Ogliari, L., & Squicciarini, M. P. (2021). Birth and migration of scientists: Does religiosity matter? Evidence from 19th-century France. *Journal of Economic Behavior and Organization*, 187, 274–289.
- Ma, C. (2021). Knowledge diffusion and intellectual change: When Chinese literati met European Jesuits. *Journal of Economic History*, 81(4), 1052–1097.
- Pascali, L. (2016). Banks and development: Jewish communities in the Italian renaissance and current economic performance. *Review of Economics and Statistics*, 98(1), 140–158.
- Platteau, J.-P. (2017). *Islam Instrumentalized: Religion and politics in historical perspective*. Cambridge University Press.
- Rubin, J. (2010). Bills of exchange, interest bans, and impersonal exchange in Islam and Christianity. *Explorations in Economic History*, 47(2), 213–227.
- Rubin, J. (2011). Institutions, the rise of commerce, and the persistence of Laws: Interest restrictions in Islam and Christianity. *Economic Journal*, 121(557), 1310–1339.
- Rubin, J. (2014). Printing and Protestants: An empirical test of the role of printing in the reformation. *Review of Economics and Statistics*, 96(2), 270–286.
- Rubin, J. (2017). *Rulers, religion, and riches: Why the west got rich and the Middle East did not*. Cambridge University Press.
- Saleh, M. (2016). Public mass modern education, religion, and human Capital in Twentieth-Century Egypt. *Journal of Economic History*, 76(3), 697–735.
- Squicciarini, M. (2020). Devotion and development: Religiosity, education, and economic Progress in 19th-century France. *American Economic Review*, 110(11), 3454–3491.
- Ticku, R., Shrivastava, A., & Iyer, S. (2023). *Economic shocks and Temple desecrations in medieval India* (Cambridge working papers in economics 1862).
- Valencia Caicedo, F. (2019). The Mission: Human capital transmission, economic persistence, and culture in South America. *Quarterly Journal of Economics*, 134(1), 507–556.
- Walker, S. (2020). Historical legacies in savings: Evidence from Romania. *Journal of Comparative Economics*, 48(1), 76–99.
- Wantchekon, L., Novta, N., & Klačnja, M. (2015). Education and human capital externalities: Evidence from colonial Benin. *Quarterly Journal of Economics*, 130(2), 703–757.
- Weber, M. (1904/05). Die protestantische Ethik und der „Geist“ des Kapitalismus. *Archiv für Sozialwissenschaft und Sozialpolitik* 20, 1–54 and 21: 1–110. Reprinted in: *Gesammelte Aufsätze zur Religionssoziologie*, 1920: 17–206. [English translation: *The Protestant Ethic and the Spirit of Capitalism*, translated by Talcott Parsons, 1930/2001, London: Routledge Classics.]
- Woodberry, R. D. (2012). The missionary roots of Liberal democracy. *American Political Science Review*, 106(2), 244–274.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Trust and the New Institutional Economics 37

Philip Keefer

Contents

37.1	Introduction	976
37.2	Variations in Trust and Trustworthiness Across Countries and Individuals	977
37.3	The Political and Economic Consequences of Trust and Trustworthiness	978
37.3.1	The Economic Costs of Mistrust	979
37.3.2	The Political Costs of Mistrust	982
37.4	Social Capital, Trust, and Institutions	986
37.4.1	Institutions and Norms of Trust	987
37.4.2	Social Networks and Trust	990
37.4.3	Intrinsic Norms of Trust and Social Distance	992
37.5	Conclusion	994
	References	996

Abstract

This essay reviews a paradox at the heart of the New Institutional Economics: institutions resolve problems of mistrust, but mistrust undermines the efficacy of institutions. Mistrust and opportunistic behavior drive the principal-agent and collective action dilemmas that are central to the New Institutional Economics, but also interfere with the institutional and organizational reforms that can resolve these dilemmas. The extrinsic and intrinsic drivers of untrustworthy, opportunistic behavior are key to addressing the paradox. The NIE agenda focuses most on reforms that reduce extrinsic incentives (reputational, financial, judicial, or social) to be untrustworthy and act opportunistically. Intrinsically motivated trustworthy behavior is also key to solving the two dilemmas, however. Institutions that support social networks and reduce social distance strengthen this intrinsic motivation.

P. Keefer (✉)

Inter-American Development Bank, Washington, DC, USA

e-mail: pkeef@iadb.org

© The Author(s) 2025

C. Ménard, M. M. Shirley (eds.), *Handbook of New Institutional Economics*,
https://doi.org/10.1007/978-3-031-50810-3_37

975

Keywords

Trust · Credible Commitment · Opportunistic behavior · Institutions · Political Economy · Growth · Development

37.1 Introduction

Scholarship in the NIE and the key contributions of Ronald Coase, Douglass North, Elinor Ostrom, and Oliver Williamson emphasize that cooperation is essential to social and economic progress; opportunistic behavior makes cooperation infeasible; but institutions can limit the threat of opportunism. These institutions have proven difficult to build, however. The paradox of trust, the focus of this essay, helps to explain why: institutions resolve problems of mistrust, but mistrust undermines the efficacy of institutions. The institutional reform agenda is therefore most difficult to implement in low-trust societies, where it is most needed.

In low-trust societies, individuals believe that others are untrustworthy and will take advantage—act opportunistically toward them. Williamson (1985, p. 47) wrote that opportunism is “self-interest seeking with guile.” Trustworthy individuals, in contrast, do not use guile in pursuit of their self-interest nor seek opportunities to renege on their commitments to others. The threats of opportunistic behavior are therefore more urgent in low-trust societies.

The NIE focuses on institutional solutions to this threat. Cook et al. (2005), for example, emphasize that successful societies do not rely on generalized trust among their citizens to solve the problem of opportunistic behavior. Instead, they depend on institutions, from courts to elections and political checks and balances, that make such behavior costly.

The institutional solution to mistrust is far from straightforward, however. Institutions emerge from and are sustained by the collective effort of individuals in a society. Collective action, though, is itself vulnerable to mistrust and opportunistic behavior. Governments, for example, are at the apex of third-party arrangements to limit untrustworthy behavior by citizens and firms. However, the collective action of citizens is an important check on government decision-makers who might otherwise shirk on their obligations to implement these third-party arrangements. Political parties can facilitate collective action by citizens to guide government performance but are more difficult to construct when party members mistrust each other or party leaders. In economic markets, production often entails collective action that is costly to coordinate in markets. Firms solve these collective action problems and internalize transactions that mistrust in the marketplace would render difficult. However, mistrust among firm employees, between owners and managers, and between managers and workers also makes internal exchange within firms more difficult.

The next sections of this essay demonstrate that trust and trustworthiness vary significantly across countries and review research showing that this variation has significant economic and political consequences. These norms directly influence the willingness of individuals to engage in political and economic exchanges with others. They also have indirect consequences: organizations (firms, political parties,

government agencies) that people build to circumvent the costs of market exchanges are more difficult to construct in low-trust environments. The last half of the essay reviews several lines of research that explore the sources of trust and trustworthiness: institutions themselves; social capital, and social distance. Although institution-building can contribute to greater trust, societies seeking to reinforce intrinsic norms of trust must also strive to close social distance and encourage social networks.

37.2 Variations in Trust and Trustworthiness Across Countries and Individuals

Development outcomes and institutional quality vary significantly across countries. So do trust and trustworthiness. To say that societies exhibit low trust is not to say that individuals in these societies trust everyone less. Rather, as Fukuyama (1995) observes, it implies that norms of trust and trustworthiness have a narrow radius. In societies where the radius is narrowest, the norms of trust limit opportunistic behavior only within tightly identified groups, such as intimate friends and the nuclear family. People in such societies are least likely to say that “people in general can be trusted.” When the norms of trust have a broader radius, they might reduce opportunistic behavior within kinship groups, social classes, or ethnic groups. The circle of trust is all-encompassing when all individuals in a society are reluctant to act opportunistically toward any other individual and correspondingly trust that all other individuals will refrain from opportunistic behavior toward them.

A standard approach to measuring the radius of trust in countries is to assess answers to the World Values Survey question, “Generally speaking, would you say that most people can be trusted, or that you can’t be too careful in dealing with people?” Across the 90 countries in the latest 2017–2022 wave of the survey, 27.5% of respondents said that most people can be trusted; 75% responded that you can’t be too careful. Fewer than 10% of Brazilians, Egyptians, Filipinos, and Peruvians responded that most people can be trusted. At the other extreme, more than 65% of Nordic respondents agreed that most people can be trusted. Thirty-seven percent of US respondents said others could be trusted (WVS Database, <https://www.worldvaluessurvey.org/WVSONline.jsp>).

Trustworthiness is complementary to trust and should be strongly correlated with it. Trusting attitudes cannot survive for long in a society with few trustworthy people. Consistent with this, measures of trust and trustworthiness from the World Values survey are also highly correlated: Knack and Keefer (1997) find that socially trustworthy behavior, measured by their index of civic cooperation, is significantly correlated (at 0.39) with the trust measure at the country level, again controlling for per capita income.

Enormous variations in survey measures of trust, both between and within countries, raise a natural question. Do the answers that respondents give to trust questions reflect their actual behavior? Do individuals who indicate that they are trusting in surveys also behave as if others are trustworthy? In fact, trusting attitudes predict trustworthy behavior.

Glaeser et al. (2000) conduct an experiment in which they pair individuals who meet and then are separated. One member of the pair, the sender, is given 15 dollars, and can send up to 15 dollars to the other member, the recipient. For each dollar sent, the “experimenter” gives an additional dollar to the recipient. The recipient can then return money to the sender. All rules of the game are known to all players. The amount of money sent is an indication of trusting behavior; the amount of money returned is an indication of trustworthy behavior. The amount sent averaged \$12.41 in the experiment, but the standard deviation was quite high (\$4.54), indicating that many participants sent much different amounts. Similarly, of the amount they received, senders on average returned 45.5%, but the standard deviation was 26.7 percentage points.¹ Glaeser et al. (2000) find that participants in the experiment who respond to the survey question by indicating that “most people be trusted” are significantly more likely to behave in a trustworthy manner: to return money to senders.

Barr (2003) conducted similar experiments in 24 Zimbabwean villages with 141 pairs of players but did not allow the subjects to know the identity of their partners prior to or during the experiment. The Zimbabwe data indicate a high correlation between trusting and trustworthy behavior across villages.

Another real-world experiment yields similar evidence. Cohn et al. (2019) deposited wallets containing money on the streets of the capital cities of 38 countries. The wallets contained sufficient information for the finders to return them to their owners. The percent of wallets returned in the 38 countries was highly correlated with country responses to the World Values Survey trust question. In countries where 20% or fewer of the wallets were returned, 15% or fewer of survey respondents indicated that others could be trusted; in countries where 60% or more of wallets were returned, approximately 30% of respondents expressed trust in others (Keefer & Scartascini, 2022).

37.3 The Political and Economic Consequences of Trust and Trustworthiness

Unbridled opportunistic behavior has significant negative consequences for economic growth and political stability. Directly, it interferes with transactions in economic and political markets: investors mistrust managers and refrain from committing capital to them and voters ignore political promises, believing that politicians will renege on them. Opportunistic behavior also has indirect effects, by making it more difficult to build the organizations (firms, government agencies, political parties) that individuals use to reduce the costs of exchange with each other.

Opportunistic behavior emerges in economic and political transactions in which parties are imperfectly informed about whether their counterparts are fulfilling their commitments and when they have limited ability to enforce those commitments. These conditions are pervasive across all the basic transactions of economic markets: financial

¹In the original version of this experiment (Berg et al., 1995), players did not meet each other, and the amounts sent and returned were somewhat lower than in Glaeser et al. (2000).

contracts, where creditors loan money to debtors on the promise of future repayment; employment contracts, where managers hire employees to accomplish tasks that are difficult to monitor or measure; and investment contracts, where investors rely on assurances by firms and governments that their assets will not be expropriated.

Imperfect information and lack of enforceability are also pervasive in political markets, where societies address their collective challenges, from crime and air pollution to national defense.² They prevent spontaneous cooperation among citizens to reduce pollution or provide for the common defense. They also undermine the delegation of responsibility for these matters to governments. Opportunities for governments to act opportunistically are greater when citizens cannot easily observe government actions nor easily enforce commitments made by politicians to increase their welfare.

Mistrust and opportunistic behavior indirectly influence economic and political markets by undermining organizations. Organizations coordinate individuals and limit free riding (see, e.g., North et al., 2009). Firms allow for production that would otherwise not occur. Political parties allow like-minded individuals to mobilize for policy change that they could otherwise not achieve; and government agencies allow societies to achieve collective goals that collective action dilemmas would otherwise render unattainable.

When trust among organization members is high, organizations can concentrate on coordination. Opportunistic behavior—free riding—is less of a threat and an organization can afford to be less hierarchical, invest less in performance management, and relax costly fiduciary controls. However, as the next sections discuss, this is not the case in organizations plagued by internal mistrust. When managers, party leaders, and government administrators mistrust workers, voters, and citizens, they replicate within organizations the types of controls and transactions that govern exchange outside of the organization. Managers preferentially hire family members; party leaders resort to vote buying; and government agencies impose stringent fiduciary controls on government-citizen exchanges.

37.3.1 The Economic Costs of Mistrust

Whether in political or economic markets, the narrower is the radius of trust, the larger the set of advantageous transactions that fall outside each person's radius of trust. Exchanges outside of the radius incur higher transaction costs as individuals seek to reduce the risks of opportunistic behavior to which they are exposed when

²Knack and Keefer (1997) refer to norms that have a wide radius and build trust at the level of communities or societies (as opposed to between pairs of traders or other small-numbers settings) as norms of "civic cooperation," or "civic norms." They develop a measure of this from the World Values Survey data based on survey respondents' beliefs about whether the following actions can ever be justified: claiming government benefits to which respondents were not entitled, avoiding a fare on public transport, cheating on taxes if they had the chance, keeping money that they had found, or failing to report damage they had accidentally done to a parked vehicle. There is wide individual-level variation within countries in the 50-point civic cooperation index.

they interact with individuals outside of their circle of trust. These exchanges are therefore less likely to occur.

In societies where the radius of trust is wide, contracting parties can dispense with costly monitoring of performance. Written contracts are less likely to be needed and they do not have to specify every possible contingency. The costs of production fall since firms and individuals can devote fewer resources to protecting themselves—through tax payments, bribes, political contributions, or private security services and equipment—from violations of their property rights. Because the risk of opportunistic behavior is lower, the returns to innovation and investment are higher.

When the radius of trust is narrow, segmented markets emerge, where individuals exchange within their narrow circle of trust, but not across circles. For example, intra-ethnic trust in an ethnically heterogeneous society leads to segmentation across ethnic groups (Greif, 1993). In economic markets, this reduces gains from specialization and economies of scale. Community members from outside of the circle, often disadvantaged, are excluded from the benefits of collective action by those from within the circle (e.g. Pantoja, 2002).

Relational contracts can allow for contracting between individuals occupying different circles of trust (see Macchiavello, 2022). They are based on long-term relationships between firms that rely on profits from future exchange to disincentivize opportunistic behavior in current transactions. Relational contracts increase welfare by allowing for investment, production, and exchange that would not otherwise occur. However, the more competitive are the markets in which firms operate, the lower are expected future profits and the more difficult it is to structure these contracts. Hence, the welfare gains from relational contracts are in tension with the potential welfare gains that societies could achieve by removing barriers to entry and increasing the competitiveness of markets.

Ample evidence indicates that the threat of opportunistic behavior suppresses investment, production, and exchange that would have taken place in the absence of this threat, lowering incomes. Keefer and Knack (1997) and Acemoglu et al. (2001) present cross-country evidence supporting the claim that growth is faster in countries with more secure property and contract rights, stronger rule of law, and higher quality of democracy. They reasonably attribute these results to the influence of formal institutions that limit the scope for opportunistic behavior in societies. However, the radius of trust is also key to securing economic rights, both directly and, because of its role in supporting formal institutions, indirectly.

Experimental evidence from the laboratory indicates that trust effects on economic markets could be large. In Glaeser et al. (2000), recipients who expressed trusting attitudes (the belief that most people can be trusted) returned 10% more money to senders than did other recipients (i.e., they were more trustworthy) controlling for other recipient characteristics.³

³The controls were the amount originally transferred by the sender to the recipient, whether the pair of participants were of different genders, the gender and race of recipient, and whether recipients were in their first year of college (the participants were all Harvard undergraduates).

Evidence from outside of the laboratory also points to the economic costs of a narrow radius of trust. In early cross-country tests of the proposition that trust supports growth, Knack and Keefer (1997) find that a ten-percentage point increase in the number of country respondents to the World Values Survey who express trusting attitudes is associated with an increase in per capita economic growth of almost one percentage point per year. The effects on growth of a one standard deviation increase in trust turn out to rival those of a one standard deviation increase in the fraction of children enrolled in primary education. Using a broader sample and different specifications, Zak and Knack (2001) report similar results.⁴ Algan and Cahuc (2013) present a wide range of estimates using more and more recent data and more robust identification strategies. These yield similar conclusions: capital accumulation, employment, and productivity are all higher and grow faster in countries with higher levels of trust.

Evidence also points to the importance of trust for organization-building in economic markets. Bloom et al. (2012) exploit evidence from 4000 firms around the world and find that firms with headquarters in high-trust regions are significantly more likely to decentralize. Within firms, higher levels of bilateral trust between the headquarter and subsidiary countries lead to higher levels of delegation to subsidiaries. This is associated with higher productivity because decentralization permits CEOs to delegate the allocation of resources by the subsidiary to better-informed subsidiary managers. When bilateral trust between headquarter and subsidiary countries is low, CEOs believe they face higher risks of opportunistic behavior by subsidiary executives and are reluctant to delegate resource allocation decisions.

Keefer and Scartascini (2022, p. 66) report even more direct evidence of the intra-firm effects of trust. The data come from a survey of the managers or owners of 3000 firms from 17 countries in Latin America and the Caribbean. Survey questions asked respondents about their trust in others, whether they delegated any of several key decisions to others inside the firm, and whether they delegated to individuals who were from their family or not from their family. Since family members are within a manager's narrow radius of trust, delegation to family members is correspondingly insensitive to managers' trust in others. Non-relatives lie outside that radius and, consistent with the importance of trust for firm organization, 88% of the correlation between trust and delegation is accounted for by delegation to employees who are not relatives.

The decision to delegate to non-family members appears to be economically important: respondents who delegate key decisions were also more likely to have made capital investments, found new customers, or begun to export. The effect is

⁴These results could be affected by reverse causation: the trust coefficient could be biased upward if growth increases trust (for example, by making people more optimistic), or downward if growth decreases trust (for example, by disrupting traditional social and community ties as in Olson, 1963 or Miguel et al., 2002). However, Zak and Knack (2001), following La Porta et al. (1997), use religious composition variables as instruments and find that the exogenous component of trust remains significantly related to growth.

also greater in larger firms, where manager capacity to closely supervise non-family members is attenuated.

Although the ability to delegate to non-family members appears to be associated with greater firm success, the costs of delegation outside the family are higher in low-trust societies. In these societies, the probability should be higher that large firms are family-controlled. Consistent with this, Aminadav and Papaioannou (2020) find that large firms in a low-trust region, Latin America, are more likely to be family controlled.⁵

Firms can explicitly choose to build trust, allowing them to expend fewer resources on worker supervision. Ouchi (1980), for example, describes how Japanese corporations relied on lifetime contracts and the socialization of their workers to align worker and firm objectives, allowing firms to spend less on evaluating worker contributions to firm profitability (see also Zucker, 1988; Kramer & Tyler, 1996). However, the success of this lifetime employment strategy depends on the credibility of those contracts to workers. The fact that Japan is a high-trust country may give Japanese firms an advantage in using this strategy to align firm and worker interests.

37.3.2 The Political Costs of Mistrust

The radius of trust and the threat of opportunistic behavior also have significant consequences for how groups and societies govern themselves. The narrower the radius, the more acute is the threat of free riding and the dilemma of collective action (Olson, 1963). Individuals in a society could all be better off if they each contributed to the provision of public goods or adhered to policies that limit the negative externalities that individual behavior can impose on others. However, individuals each have an incentive to act opportunistically toward the rest and free ride on their contributions, escaping the burden of policy compliance or public good contributions. The greater is mistrust among citizens, the more imposing is this dilemma of collective action.

Keefer and Scartascini (2022) present two pieces of evidence that point to the association between trust and free riding at the social level. One relates to the willingness of citizens to pay taxes: the ratio of tax to GDP is significantly higher in countries that exhibit high levels of interpersonal trust. For example, the ratio is well under 20% in countries where fewer than 30% of respondents say that others can be trusted. The ratio is closer to 40% in countries where more than 40% of respondents say that others can be trusted. The tax to GDP ratio naturally reflects many factors, ranging from citizen demand for government-funded services to state

⁵This is not to say that transactions within the narrow radius of trust are problem-free. Even within an extended family, norms of sharing resources can reduce the incentive for a family member to start a business (Portes, 1998). Banfield (1958) attributed incompetent government and poverty in a poor Italian village to “amoral familism” preventing cooperation among citizens.

capacity. However, holding constant these concerns, by far the most compelling indicator of the willingness of individuals in a society to contribute to the collective welfare of all is the monetary contributions that they in fact make.⁶ Those evidently rise with the level of inter-personal trust.

The second piece of evidence is the significant association between mistrust and informality. Firms and workers are informal when they are out of compliance with tax, labor, and other regulations. Non-compliance could certainly be the consequence of excessive and unreasonable regulation. Again, though, if one assumes that regulatory requirements and tax laws reflect broad social preferences and hold constant enforcement levels, the best indicator of opportunistic behavior and free riding on individual obligations to the collective good is the extent of non-compliance with these. Keefer and Scartascini (2022, p. 11) demonstrate a strong negative correlation between the percent of respondents in a country who indicate that they trust others and the size of the shadow economy.

Thomas Hobbes argued in *Leviathan* (1651) that government is the sole source of trust between strangers, implying that we should turn to government to solve the collective action dilemmas of citizens. He was not wrong: individuals delegate decision-making to government hoping that it will adopt and enforce society-wide policies that individuals would have collectively adopted in the absence of free riding. However, the relationship between citizens and government is far from immune to problems of trust: voters are as sensitive to the risks of opportunistic behavior by politicians as banks are to the risks of non-payment of loan obligations.

This has substantial implications for the political markets where members of a society compete for the right to control government. Opportunistic behavior and mistrust distort the functioning of these markets. Individuals know that politicians will pursue individuals' policy priorities only if other individuals express those same priorities. Hence, individuals circumscribe their policy demands to those that they are sure that others will also make. However, they can only count on those within their circle of trust to incur the costs of collective action in pursuit of these policies. Unfortunately, the narrower is the circle of trust, the lower the incentives of individuals to demand policies that provide benefits to everyone in society. By the same token, politicians have incentives to make promises that benefit those who believe them. As the radius of trust around a politician narrows, these promises tend to benefit ever smaller segments of society.

Keefer (2018) and Keefer and Scartascini (2022) extend this logic by noting that trust in government therefore depends not only on citizens' perceptions of whether politicians will refrain from opportunistic behavior, but also on their perceptions of whether other citizens will join them in monitoring and punishing such behavior. The less that citizens themselves can commit collectively to removing politicians who act opportunistically, the less able are politicians themselves to make credible promises.

⁶See Levi (1988) and Scholz and Lubell (1998) for the importance of informal norms and trust in revenue collection.

For example, key to citizens' oversight of government officials is their willingness to collect and assess information about government performance and their willingness to take costly action (such as voting, writing letters, signing petitions, demonstrating, etc.) to convey their preferences to officials and to expel poor performers. The purely self-interested citizen neglects both tasks and free rides on the efforts of others. Hence, Boix and Posner (1998) and Knack (2002) observe that norms prescribing cooperation and trustworthiness enhance governmental effectiveness by helping voters overcome the collective action problem of monitoring and sanctioning public officials. Nannicini et al. (2013) identify plausibly exogenous instruments for social trustworthiness (blood donations per capita) and use them to show that electoral punishment of Italian members of Parliament, for criminal prosecutions or absenteeism, is greater in districts in which citizens are more civic-minded.

It follows that trust in government should be higher in societies where generalized trust is higher: citizens can rely on each other to punish governments that shirk. Knack and Keefer (1997) report that citizens' confidence in government—the credibility of government—is significantly greater in higher-trust nations.⁷ Paxton (1999) and Inglehart (1999) argue that a culture of trust is necessary for governments to be willing to surrender power to the opposition, and therefore for the survival of democracy. Inglehart (1999) finds a strong correlation between trust and stability of democratic institutions, using cross-country data. In a classic comparative study of the United States, UK, Germany, Italy, and Mexico, Almond and Verba (1963) argue that a stable democratic political system depends on a strong “civic culture” with high interpersonal trust and active involvement in voluntary associations.

Scholars have also shown that the narrower is the radius of trust, the more that political promises tilt away from the broad pursuit of social welfare. For example, Persson and Tabellini (2000) show that the shift from majoritarian to proportional electoral rules increases the rents that politicians can extract for themselves when political competitors cannot make credible pre-election promises to voters. It reduces rents when they can.⁸ Keefer and Vlaicu (2008, 2017) describe the phenomenon of clientelism as one in which politicians have strong incentives to advocate for transfers to narrow groups and weak incentives to provide public goods. Public goods benefit those inside and outside the radius of trust and, therefore, have less political value. If the radius of trust is wide, in contrast, politicians make promises that solve the collective action dilemmas of a broad range of citizens so that they can receive widely distributed benefits that would, without government, be unattainable.

Voters understand the implications of politicians' radius of trust and modify their political demands accordingly. Using evidence from seven Latin American

⁷This test controlled for income per capita and primary and secondary educational enrollment. The dependent variable was a composite of citizens' confidence in the civil service, legal system, police, and education system.

⁸Majoritarian electoral rules are those where the number of seats per electoral district is small and citizens vote for candidates rather than parties; proportional representation rules are those where electoral districts are larger and citizens vote for party lists.

countries, Keefer et al. (2022) find that survey respondents who believe politicians keep their promises are significantly more likely to prefer raising taxes to finance public education over using these resources for private education. Individuals with lower trust levels also prefer government to prioritize direct cash transfers to citizens instead of public education investments.

The radius of trust also affects both the micro- and macroeconomic policies that voters demand. On the microeconomic side, Aghion et al. (2010) argue that costly regulation is greater in low-trust environments. This is surprising, since voters trust neither firms nor governments, and believe that regulation will be associated with corrupt deals between them. Voters nevertheless prefer greater regulation since this will at least remove the highest cost corrupt firms from the market.

Ardanaz et al. (2023) present evidence from a large survey experiment in four Latin American countries that shows a striking relationship between trust and preferences for macroeconomic—fiscal—policy. In these countries, politicians do not enjoy a broad radius of trust: most respondents (more than 75%) say that politicians are unlikely to fulfill their promises. Hence, politicians can only credibly commit to narrow groups in society. In the context of fiscal adjustment, this means they gain little from promising fiscal stability, which benefits everyone in society. Instead, they commit to targeting extra spending in boom times to those within their narrow radius and to protecting those within this radius from spending cuts in bad times. We expect, in these circumstances, that those few individuals who express trust in politicians should prefer welfare-reducing procyclical policy, exactly the relationship that Ardanaz et al. (2023) find in their data.

Cross-country evidence also reveals that in countries where the radius of trust is narrow, fewer public goods, like high-quality government, are provided. La Porta et al. (1997) and Knack and Keefer (1997) demonstrate that countries with more trusting citizens exhibit higher ratings in foreign investor risk assessments on subjective measures of governmental efficiency, corruption, and infrastructure quality.⁹ Knack (2002) reports similar results for the American states, using more “technocratic” measures of governmental quality.

Keefer (2007) compares younger and older democracies. Formal institutional arrangements, such as elections, are common to both. Nonetheless, younger democracies provide fewer public goods, such as the rule of law. The peculiarities of political promises in young democracies can account for this puzzle. In younger democracies, the radius of trust enjoyed by politicians is more likely to be narrowly confined to a few voters who have personally interacted with the politicians. Public goods, however, provide benefits that extend far beyond this narrow radius of voters. Under these circumstances, following the logic in Keefer and Vlaicu (2008), political competition need not improve public policy since politicians have weak incentives to promise and provide policies that broadly increase social welfare.

⁹La Porta et al. control for per capita income, include all countries with available data, and use trust values from the early 1990s. Knack and Keefer control for income and education, exclude formerly communist nations, and use the earliest-available observation on trust (typically, the early 1980s).

Just as mistrust affects the organization of firms in economic markets, it also affects the emergence of organizations that solve citizen collective action challenges in political markets. Even when the civic culture described by Almond and Verba (1963) is strong and robust, it is not a trivial task for like-minded individuals to cooperate to shift public policy and to hold politicians accountable for the promises that they make. A key organization that they rely on to overcome the dilemma of collective action is the political party. Parties can enable collective action by like-minded individuals seeking to achieve common political goals, including holding governments accountable for the promises they make. Evidence from Latin America and the Caribbean summarized in Keefer and Scartascini (2022) indicates that survey respondents are much more likely to express trust in government when political parties are more institutionalized. However, democracies vary greatly in the degree to which institutionalized parties exist. It is often the case that parties in democracies are structured neither to sanction free riding by members nor to allow members to replace leaders if they shirk (Keefer, 2018).

Government itself also solves citizens' collective action problems. Government agencies supply the public goods that citizens cannot individually procure for themselves. They enforce individual contributions to collective welfare that would otherwise be lost to free riding. However, the extent to which government agencies perform these tasks effectively depends on whether public sector workers collaborate effectively and are open to innovation. Trust within the public sector is crucial for both, but is often lacking.

Keefer and Vlaicu (2022) analyze evidence from a survey of 2449 public servants from 18 Latin American countries. Respondents who express greater trust in coworkers and other public employees also report that cooperation and information-sharing with their co-workers is important for them to complete their work successfully. Low-trust respondents are more resistant to innovation and organizational change. These results have a causal interpretation since they are robust to using society-wide trust in others, from other cross-country surveys, as an instrument for intra-organizational trust.

Because government performance has significant effects on citizen welfare, international development agencies invest considerable effort in working with countries to improve their government organizations (their "state capacity"). However, state capacity is itself a public good and, like all public goods, more difficult to improve in societies plagued with mistrust. Consistent with this, Cruz and Keefer (2015) document that when political parties are less institutionalized and therefore less able to overcome the problems of mistrust among citizens, World Bank public sector projects are less likely to achieve their reform objectives.

37.4 Social Capital, Trust, and Institutions

The importance of norms of trust and trustworthiness raises important questions about their origins. Individuals may adhere to a social norm for intrinsic reasons: the norm is a "moral value" (Uslaner, 2002) and violating it reduces their utility.

Alternatively, they may behave in a trustworthy manner because of the potential extrinsic consequences of not doing so, such as judicial punishment, social ostracism, or reputational loss.¹⁰ This section reviews three sources of trust, institutions, social capital (social networks), and social distance. Though all of these can operate on both extrinsic and intrinsic margins, intrinsic norms of trust are of greatest interest: the costs of economic and political exchange are lowest when people want to behave in a trustworthy manner, not when societies build costly institutions to compel them to behave in that way.¹¹

37.4.1 Institutions and Norms of Trust

Formal institutions shape the rules of the game to influence the incentives of economic and political actors to engage in opportunistic behavior. They are therefore, in the first instance, concerned with extrinsic norms of trust. Indeed, Ostrom (1990) argues that wherever people have voluntarily organized themselves to manage common-pool natural resources effectively over a long period of time, a key reason for success is that participants invest in monitoring and sanctioning free-riding behavior. In countries where legal codes, enforcement agencies, and courts are sufficiently well-developed, the prospect of legal sanctions reduces incentives to cheat. Regulatory agencies (such as the Securities and Exchange Commission), stock exchange memberships, and professional associations restrain cheating by instituting financial disclosure rules or licensing requirements (e.g., for accountants or realtors), or by promulgating formal ethical codes (e.g., bar and medical associations). Credit bureaus protect lenders from opportunistic debtors.¹² The Better Business Bureaus (formal or third-party institutions) of the United States increase the reputational consequences for opportunistic behavior by firms. They facilitate the dissemination of information about firms' compliance with contractual commitments. In the political sphere, formal institutions can tie the hands of state actors and make it difficult for them to renege on their commitments, strengthening trust.

Zucker (1986) argues that internal and external immigration weakened informal norms (or, in Fukuyama's terms, narrowed individuals' radius of trust) in the United States between 1840 and 1920. During this period, the cultural heterogeneity of immigrants and internal migration both increased, disrupting social ties. She suggests that formal institutions subsequently emerged, either in response or

¹⁰Williamson (1993) distinguished calculative, personal, and institutional trust; intrinsic trust is closest to his description of personal trust.

¹¹The discussion abstracts from the role of religion. It is, however, an obvious and important source of intrinsic motivation. Religious institutions can also enforce trustworthy behavior (as in Greif, 1993). And ties among believers can build social capital and reduce social distance (though possibly reducing social capital and increasing social distance in religiously heterogeneous communities). See Rubin (2024) Chap. ► 36 in this book.

¹²In the United States, employers often use credit bureaus to investigate job applicants. Bad credit is viewed as a predictor of shirking and thievery.

coincidentally, that offset the disruption of social ties and changed the extrinsic incentives of individuals to behave in a trustworthy manner.¹³

Cross-country evidence reveals a strong correlation between generalized trust and the strength of economic rights. Keefer and Scartascini (2022) report survey results from more than 100 countries revealing that the rule of law and the security of property rights are both greater in countries where survey respondents express greater generalized trust. Knack and Keefer (1997) look directly at formal institutional arrangements that are intended to limit opportunistic behavior. They find that interpersonal trust is higher when the executive branch of government is constrained from acting arbitrarily and when courts are regarded as independent.

These correlations indicate that generalized trust and the strength of economic rights should be mutually supportive. On the one hand, the stronger are economic rights, the less concerned are individuals about opportunistic behavior by others and the more trusting they can be. On the other, the greater is their trust in others, the more willing are individuals to delegate the enforcement of economic rights to third parties. Ample research supports both arguments.

One avenue through which institutions affect intrinsic norms is through their effect on parental behavior and the inter-generational transmission of norms. As Tabellini (2008) argues, formal institutions that raise the costs of opportunistic behavior—for example, that strengthen economic rights—shift parental calculations about the norms that are likely to be optimal for their children’s well-being. When the society around them is anarchic, parents expect that their children will regularly encounter untrustworthy behavior and opportunism. Hence, they inculcate intrinsic norms that tend toward suspicion and mistrust of strangers. These norms impose costs on their children, requiring them to exercise greater care and exert more effort in their interactions with others. Hence, when the society around them is well-ordered and economic rights are secured, parents instead prefer to inculcate intrinsic norms of trust and trustworthiness in their children.

Tabellini’s (2008) argument also accounts for the findings of research showing that traumatic past events echo through the ages: forced labor in the Peruvian colony of Spain (Dell, 2010) or the slave trade in West Africa (Nunn & Wantchekon, 2011) reduce trust and cooperative behavior centuries later. The impact of these traumas on parents and the norms that they therefore instill in their children may help to account for these long-term consequences.

¹³ At the same time that immigration was undermining informal modes of contract enforcement, letters of credit and later, credit ratings were introduced (Zucker, 1986, p. 87); requirements of financial soundness for listing on stock exchanges became more stringent; banks devoted more resources to investigating borrowers, and increased collateral requirements (Zucker, 1986, pp. 88–89); the ratio of managers to workers in manufacturing rose, as monitoring worker effort and output became a greater concern (Zucker, 1986, pp. 91–92); the proportion of transactions occurring within hierarchies, as opposed to within markets, increased (Zucker, 1986, p. 93); licensing standards (e.g., certification of accountants) emerged and professional associations were created (Zucker, 1986, p. 94); and third party enforcement increased, as with the increased use of escrow accounts.

This work shows that institutions influence intrinsic norms. However, it does not fully account for the emergence of institutions nor exclude the possibility that norms influence institutions. Ample empirical evidence indicates that intrinsically motivated trust plays a large role in human behavior and is not easily explained by institutions. Even in single-play trust and public goods experiments, where non-trusting or non-trustworthy behavior has no extrinsic (reputational or institutional) costs, most participants exhibit at least some degree of trust and trustworthiness (Ostrom, 2000; Berg et al., 1995). Intuitively, as well, the role of guilt, fear of eternal damnation, or shame—consequences that are intrinsic to an individual's utility function and are linked distantly, if at all, to reputation or formal institutions—should have a significant role in motivating trusting and trustworthy behavior.

Knack and Keefer (1997) find that cross-country variation in formal institutions leaves much of the variation in trust and trustworthiness unexplained (education and income, for example, are also highly correlated with trust, after controlling for formal institutions). Differences in formal institutions also do not explain the large variations in trust and trustworthiness across the US states (Knack, 2002) and among individuals within countries, nor the strong downward trend from about 1965 to 1990 for the United States overall (Putnam, 2000; Knack, 1992).

Other research supports the argument that intrinsic norms of trust support economic rights. Indeed, in a low-trust environment, individuals may actively oppose formal institutional arrangements to control opportunistic behavior. Greif et al. (1994) observe that although it is critical for the security of economic rights, third-party enforcement exposes citizens to the risk of opportunistic behavior by the third party. Generalized mistrust reduces citizen willingness to delegate this enforcement authority.

Keefer et al. (2023) establish the importance of trust and social relationships as crucial complements to formal institutions by analyzing data from 90 traditional food markets in Lima, Perú with identical institutional arrangements. All have identical formal rules regarding the election and responsibilities of market boards and the delegation to boards of responsibility for regulating the behavior of market vendors (e.g., regarding the upkeep of their stands or their contributions to the financing of market infrastructure). Vendor willingness to trust boards to exercise this authority seems to vary significantly across markets. Markets with stronger social ties among board members are more likely to enforce the rules and, consequently, exhibit better infrastructure and greater resilience to the emergence of supermarket competitors.

Intrinsic norms may also support extrinsic motivations that are socially and not institutionally generated, such as avoiding social ostracism. Survey evidence points to social disapproval as a significant deterrent to voter abstention in American elections (Knack, 1992). Moreover, experimental evidence indicates that many people are willing to bear sizeable costs to punish free-riding behavior by others, even in one-shot games where there are no future rewards for the punisher (Fehr & Gächter, 2000). Social ostracism may be a “second-order” norm enforcing social obligations to contribute to public goods or to behave in a trustworthy manner (e.g., Coleman, 1990: ch. 11; Elster, 1989; Hardin, 1982, pp. 172–9).

Scholars have studied two further explanations for the presence of intrinsic norms of trust and trustworthiness. One is that where social networks (an element of social

capital) are deeper and broader, the radius of trust is wider. Another is that where the social distance between individuals is lower, trust between them is greater. Both strands of literature offer important insights into norms of trust, although we are still far from having a complete understanding of when they emerge and why they vary.

37.4.2 Social Networks and Trust

A central question in research on social capital is whether social networks affect the intrinsic motivations of individuals to behave cooperatively and in a trustworthy manner.¹⁴ Coleman argues that the number, intensity, and structure of “horizontal” interactions among individuals in a community facilitate the emergence of desirable norms and trust (1990, pp. 318–319). Putnam emphasizes voluntary associations because they “instill in their members habits of cooperation, solidarity, and public-spiritedness” with positive spillovers for trust and cooperative behavior in the larger social arena (Putnam, 1993, pp. 89–90). Common membership reflects and nurtures common interests and increases the value that members place on the opinions that others have of them. They are correspondingly more reluctant to act opportunistically toward other members.

Hypotheses linking networks and associational activity, trust, and economic and political outcomes are difficult to test: social activity builds trust, but more trusting individuals are more likely to engage in social interactions and join associations. Nevertheless, regarding the first proposition, that associational activity builds trust, Brehm and Rahn (1997) find that membership in groups and trust are strongly related in US survey data, and that causation runs in both directions. Using survey data for the United States, Sweden and Germany, Stolle and Rochon (1998) conclude that membership in all types of associations is conducive to generalized trust, but do not correct for the possibility that more trusting individuals are more likely to be active in groups. Using data from the Michigan Socialization Studies from 1965 to 1982, Claibourn and Martin (2000) find that lagged trust levels are unrelated to contemporaneous group memberships, and that lagged memberships are only weakly related to contemporaneous levels of trust.

The second proposition is that social networks enhance cooperative behavior. If it is valid, groups with greater social capital (denser social networks) should confront less free riding and be better able to mobilize for collective action. This was the landmark finding in Putnam (1993), as he sought to account for the persistent differences in economic and political outcomes in southern and northern Italy. In

¹⁴Scholars have different definitions of social capital that often embrace both social networks and norms of trust; the focus here is on the effects of the first on the second. Putnam (1993, p. 167) defines social capital as “features of social organization, such as trust, norms, and networks, that can improve the efficiency of society by facilitating coordinated actions.” Coleman (1990, ch. 12) defines social capital in terms of the quantity of obligations or informal “credit slips” between parties that are likely to be repaid, thus implicating both networks (extent of obligations) and norms (which affect likelihood of repayment).

villages in Sri Lanka and India with more active community groups and associations, household participation in the design of community-based water was higher, monitoring more intense, less time was dedicated to collecting water, and water-associated health outcomes were better (Isham & Kähkönen, 2002).

Finally, if dense social networks increase trust, they should also lower the costs of economic exchange and increase incomes, as between northern and southern Italy (Putnam, 1993). Narayan and Pritchett (1999) find that Tanzanian villages with higher levels of associational membership exhibit higher household incomes. Using 21 billion friendships from Facebook, Chetty et al. (2022) document that low socioeconomic status individuals who have more friendships with high-status individuals are significantly more economically mobile.

Still, not all studies yield results consistent with Putnam's earlier findings. In Bangladesh, the neighborhood average of trust in one's neighbors was unrelated to each neighborhood's average membership in civic associations (Pargal et al., 2002). Moreover, associational activity in a community only weakly predicted whether that community was successful in organizing collective action. Knack (2002) finds that membership in groups is unrelated to measures of government performance across the American states. Knack and Keefer (1997) are also unable to identify a relationship across countries, controlling for income, education, and other variables.¹⁵

The lesson researchers have drawn from these mixed results is that the type of association, the nature of its membership, and the intensity of its activities all matter (Stolle & Rochon, 1998 and Varshney, 2002 discuss these distinctions). For example, in cross-country data, group memberships are significantly associated with trust (Knack, 2003). However, the association depends on the type of group. One type focuses on redistributionist concerns, such as professional associations and unions (as in Olson, 1982).¹⁶ The other type is social and engages in little or no lobbying on economic issues (as in Putnam). Trust is significantly higher in countries where membership in "Putnam" groups is higher but membership in unions and professional or trade associations is unrelated to trust (Knack, 2003).

One way in which group composition matters is whether it reinforces social cleavages and corresponding narrow particularistic interests or bridges them. In Weimar Germany, civil society organizations reinforced them: "socialists, Catholics, and bourgeois Protestants each joined their own choral societies and bird-watching clubs[.]" (Berman, 1997: 425). Satyanath et al. (2017) find that German towns with greater association density during the interwar period experienced significantly faster growth of Nazi Party membership. This effect was driven by towns in those federal states that exhibited less stable governments.

However, perhaps because associational activity is so diverse across countries and over time, there is little evidence that it is systematically associated with support

¹⁵These results do not directly contradict those of Putnam (1993, 2000), because his indexes of social capital mix measures of associational activity with other dimensions of social capital.

¹⁶Olson (1965, 1982) discusses the role of social ties and social sanctions in generating collective action on behalf of narrow interests.

for populism. Algan et al. (2017) and Keefer et al. (2021) argue that trust reduces support for populist—anti-institutional—candidates. To the extent that associational activity increases trust, therefore, it should reduce support for populism. In fact, Giuliano and Wacziarg (2020), looking at populist voting in the United States, examine vote shares for Donald Trump in the 2016 elections and find that they were lower in counties where the density of organization memberships was higher.

37.4.3 Intrinsic Norms of Trust and Social Distance

Research on social networks makes a convincing case that they can support trust and cooperative activity depending on differences in activity, intensity of interaction, and composition of members. However, even outside of social networks, the social distance among individuals in a society should affect trust. The more numerous the dimensions along which individuals differ and the greater are those differences, the less they interact (Akerlof, 1997) and the less able they are to trust each other (Zak & Knack, 2001).

Dimensions along which social distance might be salient include blood and ethnic ties, language, culture, education, income, wealth, occupation, social status, political and economic rights, and geographic distance. Zucker (1986, p. 63) summarizes the argument linking social distance to trust:

Just as ethnicity, sex, or age may be used as an index of job skills by employers, they can be used as an index of trust in a transaction. They serve as indicators of membership in a common cultural system, of shared background expectations. In general, the greater the number of social similarities (dissimilarities), the more interactants assume that common background expectations do (do not) exist, hence trust can (cannot) be relied upon.

At the extreme, if social distance matters, then trust and trustworthy behavior should be most evident among friends. Glaeser et al. (2000) find higher levels of trust and trustworthiness among pairs who were socially closer. Binzel and Fehr (2013) find similar patterns in experiments in Cairo, Egypt, even though their experimental design has important features that were intended to attenuate the influence of friendship.¹⁷ Etang et al. (2011) investigate a social distance within and between villages. Their experimental data from rural Cameroon reveals that participants in the trust game send significantly more money to fellow villagers; the magnitude of the effect is similar to the effects of gender, education, and membership in rotating credit groups. Cross-country studies have looked at social distance in terms of social characteristics and income. Ethnic and linguistic homogeneity is associated with greater trust, while income inequality decreases it (Knack & Keefer, 1997; Zak & Knack, 2001; Alesina & La Ferrara, 2000, 2002).

¹⁷ First, even when partners act in a trustworthy manner, subjects receive the untrustworthy payoff with a certain probability. Second, the payoffs that participants are randomly based on either the trust game with the stranger or with their friend.

If social distance increases mistrust, it should also reduce civic cooperation. In fact, cooperation with the census and participation in groups are lower where ethnic heterogeneity and income inequality are higher (Alesina & La Ferrara, 2000, 2002; Vigdor, 2004; Costa & Kahn, 2003a). Desertion in the Union Army in the U.S. Civil War was higher in companies with greater diversity in age and occupation (Costa & Kahn, 2003b). Karlan (2003) finds that more culturally homogeneous rotating credit associations have lower default rates on loans. Similarly, Kähkönen (1999) and Grootaert (1999) document greater collective action (e.g., to secure water supplies) in more socially homogeneous villages (e.g., along caste, kinship, religious or ethnic lines).¹⁸

If social heterogeneity interferes with social cooperation, its effects should be detected in government performance, as well, recalling that it, also, ultimately depends on social cooperation to hold government accountable for its actions. Easterly and Levine (1997) find that ethnic heterogeneity in countries is correlated with a range of indicators of inefficient policies, including a high black-market currency premium, high corruption levels, low schooling rates, a lack of financial development, and poor infrastructure. Using cross-city and cross-county data for the United States, Alesina et al. (1999) find lower levels of public good provision in more ethnically divided areas. Miguel and Gugerty (2005) show similar results across Kenyan communities. Keefer and Knack (2002) conclude that property rights are more uncertain in highly polarized societies, as measured not only by ethnic tensions and heterogeneity but also by income and land inequality.

Empirical investigations of the effects of social distance do not pin down the exact mechanisms through which it might operate. Altruism may be greater in more homogeneous groups since homogeneity may reduce the willingness to act opportunistically toward others. Social distance also affects individuals' beliefs about others. They may regard themselves to be inherently trustworthy and project this belief on others socially close to them, with similar backgrounds and experiences. Their confidence in the inherent trustworthiness of others weakens as social distance with others increases.

Related to this mechanism, people may also believe that they share similar preferences and goals with others who are socially close to them, along whichever dimension of social distance is most salient (religion, geography, ethnicity, etc.). This is important because the costs of collective action are lower when group members share the same objectives. When preferences differ, as when social distance is greater, mistrust grows because incentives to free ride on collective obligations are stronger.

¹⁸ Exogenous factors appear to influence how much social distance matters. Van Bastelaer (2000) finds no effect of social homogeneity on repayment rates in credit associations and argues that this is because, among the associations he studies, income shocks covary more in homogeneous groups. Glennerster et al. (2013) estimate a precise null effect of local diversity on collective action in post-war Sierra Leone, possibly because the experience of war itself changed the salience of pre-war measures of social distance.

These mechanisms, all rooted in the beliefs about others, point to a key role for information and media. Individuals can seek out and consume information that reinforces or changes their beliefs about those who are socially more distant. Campante et al. (2022) and Keefer and Scartascini (2022, ch. 9) review these effects, which raise complex and difficult issues. Among the most important: whether greater access to information increases trust and makes beliefs about others more accurate depends on the complex interaction of the behavioral and cognitive characteristics of consumers and the strategic behavior of the suppliers of information. Blouin and Mukand (2019) present evidence that gives reason for optimism: government-sponsored radio broadcasts in post-genocide Rwanda engendered greater interethnic trust.

Another plausible mechanism through which social distance might affect trust relates to fairness and agreement about how much one should contribute voluntarily to a public good. The collective action literature has long posited that conformity with cooperative social norms is motivated by a sense of fairness or reciprocity (e.g., Hardin, 1982; Fehr & Gächter, 2000). Social heterogeneity can reduce agreement about what constitutes a fair set of contributions, for example across individuals with different incomes. Here, it is not differences in preferences regarding which public goods should be supplied, but rather differences in norms regarding who should pay for them, that reduce trust.

The risk of social opprobrium or ostracism in the event of improper behavior toward another may also increase as social distance declines. John Stuart Mill (1848, pp. 135–136, 444) wrote that “. . . much of the security of person and property in modern nations is the effect of manners and opinion” and of “the fear of exposure.” However, the most interesting findings in the literature on social distance relate to its effects among strangers, where fear of ostracism should matter less.

Research on social distance does not clearly distinguish the potentially independent effects of distance and social networks. In fact, the work of Greif (1993) illustrates that the two can overlap. This work also illustrates the indistinct boundary between intrinsic and extrinsic norms. The Maghribi traders managed large transactions over long distances and time periods, despite substantial threats of opportunistic behavior. They shared a common religion, supportive of intrinsic norms. However, their religious institutions also created a social network through which information on trader behavior could be reliably transmitted to other traders. The reputational effects of the Maghribi social network provided extrinsic motivation to behave in a trustworthy manner, but the willingness to incur costs to punish could have been an intrinsic value that was a product of a common group identity.

37.5 Conclusion

The NIE anticipated that the options available to limit opportunistic behavior include both formal rules and procedures and informal norms (North, 1990). However, the reform agenda that flows from the NIE has typically focused on formal institutional reforms that reduce the extrinsic incentives of actors to engage in opportunistic behavior. To limit opportunism in the public sector, reformers focus on elections and the strengthening of public sector organizations, such as tax administration.

To encourage private sector exchange and investment, reform efforts aim to establish strong judiciaries and regulatory institutions. In low-trust settings, these institutional arrangements function less well.

The effects of opportunistic behavior and mistrust on organization-building explain why these institutions have proven to be difficult to establish. Indeed, a growing body of research suggests that informal norms of trust and trustworthiness may play an even larger role than first suspected.¹⁹ For example, the efficacy of formal institutions may depend on informal norms of trust and trustworthiness, both because formal institutions are more difficult to build when mistrust prevails, and because it is difficult to devise formal institutions that can curb opportunistic behavior in all of the contexts in which it can arise.

However, although intrinsic norms of trust and trustworthiness are crucial, they are difficult to instill.²⁰ This is apparent in Woolcock (1998, p. 186)'s observation that

The challenge for development theorists and policy-makers alike is to identify the mechanisms that will create, nurture, and sustain the types and combinations of social relationships conducive to building dynamic participatory societies, sustainable equitable economies, and accountable developmental states.

Keefer and Shirley (2000) argue, in societies where formal and informal institutions of wide "radius" are missing it may be possible in the short and medium term to improve just the reach and functioning of informal norms that operate only within family, religious or ethnic groups, despite the risks that this poses for inter-group transactions and cohabitation. Still, even at the community level, it may be difficult to foster social norms or networks. Gugerty and Kremer (2002) examined the impact of a donor-funded program to strengthen rural women's groups in Kenya. They find that groups randomly chosen to receive donor funding experienced larger turnover in membership, and changes in leadership in favor of men or more educated women.

Even if trust is difficult to build, more can be done to take it into account in the design of political, economic, and institutional reforms. The importance of trust and trustworthiness is sufficiently well-documented, and ways to measure them are sufficiently reliable, that such efforts are amply justified. For example, tax reforms to improve auditing procedures are in principle aimed at preventing some individuals from free riding on the contributions of others. However, even those who pay their taxes may oppose these reforms if they worry that the tax administration acts

¹⁹ Sociologists have long been concerned with social norms. Norms "specify what actions are regarded by a set of persons as proper or correct, or improper and incorrect" (Coleman, 1990, p. 243). Norms and their accompanying potential rewards (for compliance) or punishments (for noncompliance) "affect the costs and benefits which individuals taken into account when exercising choice" (Coleman, 1987, p. 135). Norms have no legal or other formal basis and may sometimes even conflict with laws (Coleman, 1990, p. 243).

²⁰ Questions of how social norms emerge and evolve over time in response to changes in technology, population, the political environment, etc., the extent to which internalization of norms is individually rational, and their neural and cognitive underpinnings (Coleman, 1990; Elster, 1989; Gelfand et al., 2024), are beyond the scope of this chapter.

opportunistically. At the same time, the effect of technical reforms to increase the extrinsic costs of evasion could be strengthened with public campaigns that reinforce intrinsic norms regarding contributions. Reforms such as a land tax or the construction of rural roads can be evaluated not only for their economic merits and their impact on productivity and incomes in rural areas. Their non-economic effects could be even greater if they disrupt feudal social relations and limit the capacity of local leaders to act opportunistically toward rural households.

Reformers, especially in low-trust societies, can also turn to the primitive conditions that give rise to opportunistic behavior, especially imperfect information. Mistrust may make it difficult to strengthen the formal institutions that allow for easy enforcement of contracts and property rights, for example, or that limit illicit electoral behavior. Reforms that increase individuals' information about product quality, borrower behavior, or government performance may have more success in low-trust, weak-institution environments.

Eighteen years ago, Keefer and Knack (2005) reviewed the implications of social capital and social norms for the New Institutional Economics and identified three priorities for future research. One was identifying institutional mechanisms for defusing tensions among groups with few common norms or networks to unite them. The second was to describe the conditions under which the distrust associated with ethnic heterogeneity can be alleviated. Without having successfully addressed either priority, new and related challenges have emerged. The intervening decades have highlighted how little we know about the conditions under which social distance can not only grow, but do so on new dimensions, such as politics. While still far from understanding how to build inter-ethnic trust, we are confronted with similar challenges in building trust across political cleavages. The third priority they cited remains as relevant as ever and its answer will ultimately help us address the first two challenges: identifying ways in which governments or donors can support the emergence of trust and trustworthiness in adverse environments.

Acknowledgments I am exceedingly grateful for the extensive and insightful comments of the editors, Claude Ménard and Mary Shirley. The earlier edition of this book contained a chapter on social capital and institutions written by both Steve Knack and me. If not for Steve's tragic death, this chapter, too, would have been the—much superior—product of our joint efforts. The conclusions of this chapter are not intended to represent the views of the Inter-American Development Bank, its Executive Directors, or the countries they represent.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Acemoglu, D., Johnson, S., and Robinson, J.A. (2001). "The colonial origins of comparative development: An empirical investigation." *American Economic Review* 91(5), 1369–1401.
- Aghion, P., Algan, Y., Cahuc, P., & Shleifer, A. (2010). Regulation and distrust. *The Quarterly Journal of Economics*, 125(3), 1015–1049. <https://doi.org/10.1162/qjec.2010.125.3.1015>
- Akerlof, G. (1997). Social distance and social decisions. *Econometrica*, 65(5), 1005–1027.

- Alesina, A., & La Ferrara, E. (2000). Participation in heterogeneous communities. *Quarterly Journal of Economics*, 115(3), 847–904.
- Alesina, A., & La Ferrara, E. (2002). Who trusts others? *Journal of Public Economics*, 85(2), 207–234.
- Alesina, A., Baqir, R., & Easterly, W. (1999). Public goods and ethnic divisions. *Quarterly Journal of Economics*, 114(4), 1243–1284.
- Algan, Y., & Cahuc, P. (2013). Trust and growth. *Annual Review of Economics*, 5(1), 521–549. <https://doi.org/10.1146/annurev-economics-081412-102108>
- Algan, Y., Guriev, S., Papaioannou, E., & Passari, E. (2017). *Brookings papers on economic activity: The European trust crisis and the rise of populism*. Brookings Institution Press.
- Almond, G. A., & Verba, S. (1963). *The civic culture: Political attitudes and democracy in five nations*. Princeton University Press.
- Aminadav, G., & Papaioannou, E. (2020). Corporate control around the world. *The Journal of Finance*, 75(3), 1191–1246.
- Ardanaz, M., Hübscher, E., Keefer, P., & Sattler, T. (2023). *Voter responses to fiscal crisis: Experimental evidence on preferences for fiscal adjustment* (Inter-American Development Bank working paper). Inter-American Development Bank.
- Banfield, E. C. (1958). *The moral basis of a backward society* (Vol. 26, p. 185). The Free Press.
- Barr, A. (2003). Trust and expected trustworthiness: Experimental evidence from Zimbabwean villages. *Economic Journal*, 113(6), 14–30.
- Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10, 122–142.
- Berman, S. (1997). Civil society and the collapse of the Weimar Republic. *World Politics*, 49(3), 401–429.
- Binzel, C., & Fehr, D. (2013). Social distance and trust: Experimental evidence from a slum in Cairo. *Journal of Development Economics*, 103(July), 99–106.
- Bloom, N., Sadun, R., & Van Reenen, J. (2012). The organization of firms across countries. *The Quarterly Journal of Economics*, 127(4), 1663–1705.
- Blouin, A., & Mukand, S. W. (2019). Erasing ethnicity? Propaganda, nation building, and identity in Rwanda. *Journal of Political Economy*, 127(3), 1008–1062.
- Boix, C., & Posner, D. N. (1998). Social capital: Explaining its origins and effects on government performance. *British Journal of Political Science*, 28, 686–693.
- Brehm, J., & Rahn, W. (1997). Individual-level evidence for the causes and consequences of social capital. *American Journal of Political Science*, 41(3), 999–1023.
- Campante, F., Durante, R., & Tesei, A. (2022). Media and social capital. *Annual Review of Economics*, 14, 69–91.
- Chetty, R., et al. (2022). Social capital I: Measurement and associations with economic mobility. *Nature*, 608, 108–121.
- Claibourn, M. P., & Martin, P. S. (2000). Trusting and joining? An empirical test of the reciprocal nature of social capital. *Political Behavior*, 22(4), 267–291.
- Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science*, 365(6448), 70–73.
- Coleman, J. (1987). Norms as social capital. In G. Radnitzky & P. Bernholz (Eds.), *Economic Imperialism*. Paragon.
- Coleman, J. (1990). *Foundations of social theory*. Harvard University Press.
- Cook, K. S., Hardin, R., & Levi, M. (2005). *Cooperation without trust?* Russell Sage Foundation.
- Costa, D. L., & Kahn, M. E. (2003a). Understanding the decline in social capital, 1952–1998. *Kyklos*, 56(1), 17–46.
- Costa, D. L., & Kahn, M. E. (2003b). Cowards and heroes: Group loyalty in the American civil war. *Quarterly Journal of Economics*, 118(2), 519–548.
- Cruz, C., & Keefer, P. (2015). Political parties, clientelism, and bureaucratic reform. *Comparative Political Studies*, 48(14), 1942–1973.
- Dell, M. (2010). The persistent effects of Peru's mining *Mita*. *Econometrica*, 78(6), 1863–1903.

- Easterly, W., & Levine, R. (1997). Africa's growth tragedy: Policies and ethnic divisions. *Quarterly Journal of Economics*, 112(4), 1203–1250.
- Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4), 99–117.
- Etang, A., Fielding, D., & Knowles, S. (2011). Does trust extend beyond the village? Experimental trust and social distance in Cameroon. *Experimental Economics*, 14, 15–35.
- Fehr, E., & Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3), 159–181.
- Fukuyama, F. (1995). *Trust: The social virtues and the creation of prosperity*. The Free Press.
- Gelfand, M. J., Gavrilets, S., & Nunn, N. (2024). Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual Review of Psychology*, 75, 341.
- Giuliano, P., & Wacziarg, R. (2020). *Who voted for Trump? Populism and social capital* (CEPR discussion papers 15140). Institute of Labor Economics (IZA).
- Glaeser, E. L., Laibson, D., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *Quarterly Journal of Economics*, 115(3), 811–846.
- Glennester, R., Miguel, E., & Rothenburg, A. D. (2013). Collective action in diverse Sierra Leone communities. *The Economic Journal*, 123(568), 285–316.
- Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *The American Economic Review*, 83(3), 525–548.
- Greif, A., Milgrom, P., & Weingast, B. (1994). Coordination, commitment, and enforcement: The case of the merchant guild. *Journal of Political Economy*, 102(4), 745–776.
- Grootaert, C. (1999). *Social capital, household welfare and poverty in Indonesia* (Local Level Institutions working paper No. 6). The World Bank.
- Gugerty, M. K., & Kremer, M. (2002). The impact of development assistance on social capital: Evidence from Kenya. In C. Grootaert & T. van Bastelaer (Eds.), *The role of social capital in development* (Chap. 7, pp. 213–233). Cambridge University Press.
- Hardin, R. (1982). *Collective action*. Resources for the Future.
- Inglehart, R. (1999). Trust, well-being and democracy. In M. Warren (Ed.), *Democracy and trust*. Cambridge University Press.
- Isham, J., & Kähkönen, S. (2002). Institutional determinants of the impact of community-based water services: Evidence from Sri Lanka and India. *Economic Development and Cultural Change*, 50(3), 667–691.
- Kähkönen, S. (1999). *Does social capital matter in water and sanitation delivery? A review of literature* (Social Capital Initiative working paper No. 9). The World Bank.
- Karlan, D. (2003). *Social capital and group banking*. Dissertation chapter, Massachusetts Institute of Technology.
- Keefer, P. (2007). Clientelism, credibility, and the policy choices of young democracies. *American Journal of Political Science*, 51(4), 804–821.
- Keefer, P. (2018). Organizing for prosperity: Collective action, political parties, and the political economy of development. C. Lancaster & N. Van de Walle (Eds.), *The Oxford handbook of the politics of development* (Chap. 21). Oxford University Press.
- Keefer, P., & Knack, S. (2002). Polarization, property rights, and the links between inequality and growth. *Public Choice*, 111, 127–154.
- Keefer, P., & Knack, S. (2005). Social capital, social norms and the new institutional economics. C. Ménard & M. Shirley (Eds.), *Handbook of the new institutional economics* (Chap. 27, pp. 701–725). Springer.
- Keefer, P., & Scartascini, C. (2022). *Trust: The key to social cohesion and growth in Latin America and the Caribbean*. Inter-American Development Bank.
- Keefer, P., & Shirley, M. M. (2000). Formal versus informal institutions in economic development. In C. Ménard (Ed.), *Institutions, contracts and organizations*. Edward Elgar.
- Keefer, P., & Vlaicu, R. (2008). Democracy, credibility, and clientelism. *The Journal of Law, Economics, and Organization*, 24(2), 371–406. <https://doi.org/10.1093/jleo/ewm054.318>
- Keefer, P., & Vlaicu, R. (2017). Vote buying and campaign promises. *Journal of Comparative Economics*, 45(4), 773–792.

- Keefer, P., & Vlaicu, R. (2022). *Employee trust and performance constraints in public sector organizations* (Inter-American Development Bank working paper IDB-WP-1376). Inter-American Development Bank.
- Keefer, P., Scartascini, C., & Vlaicu, R. (2021). Trust, populism, and the quality of government. In A. Bågenholm, M. Bauhr, M. Grimes, & B. Rothstein (Eds.), *The Oxford handbook of the quality of government* (Chap. 12, pp. 249–267). Oxford University Press.
- Keefer, P., Scartascini, C., & Vlaicu, R. (2022). Demand-side determinants of public spending allocations: Voter trust, risk and time preferences. *Journal of Public Economics*, 206(104579), 104579.
- Keefer, P., Espinoza, Á., Espinoza, M., & Fort, R. (2023). *The rule of law, collective action, and growth: Micro evidence from Perú* (Inter-American Development Bank working paper). Inter-American Development Bank.
- Knack, S. (1992). Civic norms, social sanctions, and voter turnout. *Rationality and Society*, 4, 133–156.
- Knack, S. (2002). Social capital and the quality of government: Evidence from the U.S. states. *American Journal of Political Science*, 46(4), 772–785.
- Knack, S. (2003). Groups, growth and trust: Cross-country evidence on the Olson and Putnam hypotheses. *Public Choice*, 117, 341–355.
- Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics*, 112(4), 1251–1288.
- Kramer, R. M., & Tyler, T. R. (1996). *Trust in organizations*. Sage Publications.
- Levi, M. (1988). *Of rule and revenue*. University of California Press.
- Macchiavello, R. (2022). Relational contracts and development. *Annual Review of Economics*, 13, 337–362.
- Miguel, E., & Gugerty, M. K. (2005). Ethnic diversity, social sanctions, and public goods in Kenya. *Journal of Public Economics*, 89(11–12), 2325–2368.
- Miguel, E., Gertler, P., & Levine, D. I. (2002). *Did industrialization destroy social capital in Indonesia?* Unpublished working paper, University of California, Berkeley, CA.
- Mill, J. S. (1848). *Principles of political economy*. John W. Parker.
- Nannicini, T., Stella, A., Tabellini, G., & Troiano, U. (2013). Social capital and political accountability. *American Economic Journal: Economic Policy*, 5(2), 222–250.
- Narayan, D., & Pritchett, L. (1999). Cents and sociability: Household income and social capital in rural Tanzania. *Economic Development and Cultural Change*, 47(4), 871–897.
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders: A conceptual framework for interpreting recorded human history*. Cambridge University Press.
- Nunn, N., & Wantchekon, L. (2011). The slave trade and the origins of mistrust in Africa. *American Economic Review*, 101(7), 3221–3252.
- Olson, M. (1963). Rapid growth as a destabilizing force. *Journal of Economic History*, 23, 529–552.
- Olson, M. (1965). *The logic of collective action: Public goods and the theory of groups*. Harvard University Press.
- Olson, M. (1982). *The rise and decline of nations: Economic growth, stagflation, and social rigidities*. Yale University Press.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
- Ostrom, E. (2000). Collective action and the evolution of social norms. *Journal of Economic Perspectives*, 14(3), 137–158.
- Ouchi, W. G. (1980). Markets, bureaucracies, and clans. *Administrative Science Quarterly*, 25(1), 129–141.
- Pantoja, E. (2002). Qualitative analysis of social capital: The case of coal-mining areas in Orissa, India. In C. Grootaert & T. van Bastelaer (Eds.), *Understanding and measuring social capital* (Chap. 5, pp. 108–151). The World Bank.

- Pargal, S., Gilligan, D., & Huq, M. (2002). Social capital in solid waste management: Evidence from Dhaka, Bangladesh. In C. Grootaert & T. van Bastelaer (Eds.), *The role of social capital in development* (Chap. 6, pp. 188–212). Cambridge University Press.
- Paxton, P. (1999). Is social capital declining in the United States? A multiple indicator assessment. *American Journal of Sociology*, 105(1), 88–127.
- Persson, T., & Tabellini, G. (2000). *Political economics: Explaining public policy*. The MIT Press.
- Porta, L., Rafael, F. L.-d.-S., Shleifer, A., & Vishny, R. W. (1997). Trust in large organizations. *American Economic Review Papers and Proceedings*, 87, 333–338.
- Portes, A. (1998). Social capital: Its origins and applications in modern sociology. *Annual Review of Sociology*, 24, 1–24.
- Putnam, R. with Leonardi, R., & Nanetti, R. Y. 1993. *Making democracy work: Civic traditions in modern Italy*. Princeton University Press.
- Putnam, R. (2000). *Bowling alone: Collapse and revival of American community*. Simon & Schuster.
- Satyanath, S., Voigtländer, N., & Voth, H.-J. (2017). Bowling for fascism: Social capital and the rise of the Nazi Party. *Journal of Political Economy*, 125(2), 478–526.
- Scholz, J. T., & Lubell, M. (1998). Trust and taxpaying: Testing the heuristic approach to collective action. *American Journal of Political Science*, 42(2), 398–417.
- Stolle, D., & Rochon, T. (1998). Are all associations alike? Member diversity, associational type and the creation of social capital. *American Behavioral Scientist*, 42(1), 47–65.
- Tabellini, G. (2008). The scope of cooperation: Values and incentives. *The Quarterly Journal of Economics*, 123(3), 905–950.
- Uslaner, E. (2002). *The moral foundations of trust*. Cambridge University Press.
- Van Bastelaer, T. (2000). *Does social capital facilitate the poor's access to credit? A review of the microeconomic literature* (Social Capital Initiative working paper No. 8). The World Bank.
- Varshney, A. (2002). *Ethnic conflict and civic life: Hindus and Muslims in India*. Yale University Press.
- Vigdor, J. L. (2004). Community composition and collective action: Analyzing initial mail response to the 2000 census. *Review of Economics and Statistics*, 82(1), 251–271.
- Williamson, O. E. (1985). *The economic institutions of capitalism: Firms, markets, relational contracting*. The Free Press, Collier Macmillan.
- Williamson, O. E. (1993). Calculativeness, trust and economic organization. *Journal of Law and Economics*, 36, 453–486.
- Woolcock, M. (1998). Social capital and economic development: Toward a theoretical synthesis and policy framework. *Theory and Society*, 27, 151–208.
- Zak, P., & Knack, S. (2001). Trust and growth. *Economic Journal*, 111, 295–321.
- Zucker, L. G. (1986). Production of trust: Institutional sources of economic structure, 1840–1920. *Research in Organizational Behavior*, 8, 53–111.
- Zucker, L. G. (1988). *Institutional patterns and organizations*. Ballinger.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Alberto Simpser

Contents

38.1	Introduction	1002
38.2	Informal Norms as Drivers of Corrupt Behavior	1003
38.2.1	What Is a Norm?	1003
38.2.2	Norms and Behavior	1004
38.2.3	Empirical Evidence on Norms and Corrupt Behavior	1007
38.2.4	Are Informal Norms and Behavior “Proportional”?	1009
38.3	Norms and the Law	1011
38.3.1	Mutual Causation	1011
38.3.2	Norms as Moderators of the Effectiveness of the Law	1012
38.3.3	The Law as a Moderator of the Effectiveness of Norms	1013
38.3.4	Efficiency of Laws Vs. Norms	1015
38.4	Informal Norms in the Fight Against Corruption	1015
38.4.1	Personal Norms and Anticorruption	1016
38.4.2	Social Norms and Anticorruption	1017
38.5	Conclusion	1019
	References	1019

Abstract

Corruption, the abuse of entrusted power for private gain, has long been understood to respond to costs and benefits as shaped by the law. Increasingly, informal norms are also receiving attention as drivers of corrupt behavior and as potential policy targets to combat corruption. This chapter surveys existing scholarship about how informal norms affect behavior, how they interact with the law, and how they might be influenced through policy interventions to combat corruption. The reader seeking fully crafted answers and policy recipes will not find them here. Instead, this chapter aims to highlight promising ideas, suggestive evidence, and avenues for future research and policy development.

A. Simpser (✉)
ITAM, Mexico City, Mexico
e-mail: alberto.simpser@itam.mx

Keywords

Corruption · Dishonesty · Social norms · Moral norms · Informal norms · Formal institutions · Law

38.1 Introduction

In the broadest sense, corruption refers to moral decay (Buchan & Hill, 2014). For the contemporary policymaker, multilateral organization, or social scientist, however, corruption is more narrowly understood to mean the abuse of entrusted power for private gain, a definition that evokes a public servant or elected official receiving bribes in exchange for services or privileges, but also recognizes that corruption can and does arise among private-sector and other non-state organizations and persons¹.

The study of corruption has a rich intellectual history encompassing many disciplines in the humanities and social sciences (Buchan & Hill, 2014). Nevertheless, the predominant lens through which the causes and the consequences of corruption, as well as anticorruption policy, are understood today is an economic one. Efforts to combat corruption were, and continue to be, justified largely on the basis of corruption's potential to undermine economic development (World Bank, 1997a, b; Mauro, 1995, 1997, 1998; Tanzi, 1998; Olken & Pande, 2012).² The causes of corrupt behavior are also primarily understood through the lens of microeconomic theory (Buchanan & Tullock, 1962; Becker, 1968; Becker & Stigler, 1974; Rose-Ackerman, 1978). In this tradition, corrupt behavior is argued to arise from calculated opportunism. Typically, a “principal” hires an “agent” to perform a task. The principal wants the agent to act honestly, but the agent has discretion, and the principal cannot keep constant watch. The agent decides whether to act honestly or corruptly by comparing the material benefits of acting corruptly (bribe money, for example) vs. the expected cost of doing so (e.g., the cost of a fine or jailtime weighted by the probability of being caught and found guilty). When benefits exceed expected costs, the agent chooses corruption.

Without denying the importance of material incentives and the usefulness of the microeconomic approach, scholars and practitioners have long recognized that corrupt behavior is also influenced by moral, social, and normative forces. Banfield (1967) describes a norm of “amoral familism” as a cause of corrupt behavior by public officials in Southern Italy. Treisman (2000, 2007) discusses religious norms as one among a wide range of causes of corruption. Even the microeconomics

¹ <https://www.transparency.org/en/what-is-corruption#:~:text=We%20define%20corruption%20as%20the%20abuse%20of%20entrusted%20power%20for%20private%20gain>

² See for example: <https://www.worldbank.org/en/topic/governance/brief/combating-corruption>. An older literature argued that where elites are hostile to economic development corruption can enable economic activity (Leff, 1964; Zhang et al., 2023).

tradition recognized early on that people differ in their propensity to act honestly even when facing identical material incentives (e.g., Becker & Stigler, 1974).

Anticorruption policies emanating from the microeconomic approach aim to tilt the material cost-benefit calculus to render corrupt behavior less desirable (Becker, 1968; Klitgaard, 1988; Shleifer & Vishny, 1993). In practice, this idea takes the form of tighter performance monitoring, better accounting controls in procurement, and stricter laws and regulations, among others (Hanna et al., 2011).³ If norms matter, however, policies might also seek to select people with “good” norms, and to change norms themselves (Klitgaard, 1988; Barfort et al., 2019). Approaches to corruption that recognize the role of informal norms are gradually making inroads in the outlook of international bureaucracies (e.g., World Development Report, 2017, pp. 77–79).⁴

The purpose of this chapter is to survey scholarship relating to informal norms as causes of corruption and to the way informal norms interact with the law. This goal is very much in the spirit of the New Institutional Economics (NIE), which has depicted informal norms as on a par with formal institutions as drivers of economic development (North, 1990) and accorded a key role to social norms, beliefs, and cognitions in understanding human behavior (North, 2005). Corruption research in the NIE tradition has, for instance, emphasized the importance of informal norms as substitutes for formal commitment mechanisms, because corrupt agreements cannot be enforced in courts of law (Lambsdorff et al., 2005). The chapter draws on research from a variety of social science disciplines, an approach that I also take to be in the spirit of the NIE, which has long recognized the benefits of “reconciling differences between economics and other social sciences” (North, 1990, p. 5).

The chapter is organized as follows. Section 38.2 discusses norms as causes of corrupt behavior. Section 38.3 focuses on how formal laws and informal norms interact, substitute for, or reinforce each other to influence behavior. Section 38.4 discusses ideas and evidence about the role of informal norms in the fight against corruption. Section 38.5 offers brief concluding thoughts.

38.2 Informal Norms as Drivers of Corrupt Behavior

38.2.1 What Is a Norm?

I employ “informal norms” as a catch-all term that encompasses social norms as well as personal or moral norms. While there exists no universally accepted definition of social norms, these come in two broad types: *descriptive norms* refer to how others behave on average, while *injunctive norms* denote what others approve or

³See also <https://ida.worldbank.org/en/topics/cross-cutting/governance-and-institutions>

⁴Corruption’s noneconomic *consequences* are also receiving increasing attention. Corruption, for example, has been argued to undermine generalized trust and prosocial behavior (Spadaro et al., 2023) as well as trust in government (Sole-Olle & Sorribas-Navarro, 2014).

disapprove of on average (Cialdini et al., 1990; Schultz et al., 2007; Krupka & Weber, 2009). Implicit in the idea that social norms drive behavior is the notion that individual behavior is conditional on (an individual's perceptions of) what others think (injunctive norms) and/or how they act (descriptive norms).

Cristina Bicchieri's seminal work refines these ideas. In her account, an individual's social preferences are *conditional* when they depend on how others behave or think, and *unconditional* when they do not (2017, pp. 8–9). A person's beliefs about how others are likely to behave are termed *empirical expectations*, and a person's beliefs about whether others approve or disapprove of a specific behavior are labeled *normative expectations*. Bicchieri defines a *social norm* as “a rule of behavior such that individuals prefer to conform to it on condition that they believe that (a) most people in their reference network conform to it (empirical expectations) and (b) most people in their reference network believe they ought to conform to it (normative expectations)” (Bicchieri, 2017, p. 35). By a *reference network*, Bicchieri refers to the set of other people whose behavior and normative beliefs a person cares about (e.g., family, close friends, and colleagues).

Thus, descriptive norms resemble Bicchieri's notion of empirical expectations, while injunctive norms resemble normative expectations. Nevertheless, a norm in Bicchieri's account is a rule of behavior, while descriptive and injunctive norms simply describe others' average actions and normative viewpoints, without specifying how individual behavior depends on these.

Another kind of informal norm is a *personal* or *moral norm*: an internalized sense of duty or altruism that does not depend on how others behave or think and therefore is, as per Bicchieri, unconditional. Personal norms are often referred to as intrinsic motivations and are often understood to be a type of preference. Prominent realms of personal norms include altruism (understood as caring for others' wellbeing) and a sense of duty or morals (i.e., about what is right or wrong for its own sake).

38.2.2 Norms and Behavior

Social and personal norms have been found to influence behavior in a wide variety of realms including tax compliance, energy use, and voting, among others (Sugden, 1982; Andreoni, 1989; Cialdini et al., 1990; Gerber et al., 2008; Gerber & Rogers, 2009; Allcott & Rogers, 2014; Bicchieri, 2017). But what are the reasons for such influence? Adhering to, or deviating from, a *personal norm* can generate powerful emotions such as personal satisfaction, “warm glow,” guilt, and regret. *Social norms'* effect on behavior stems in part from the powerful psychological drive to imitate and conform (McDonald & Crandall, 2015; Over, 2020).

Different mechanisms linking individual behavior to social norms have been suggested. *Norm learning* describes a process in which people match their actions “to what they infer is normative among their peers through observation” (Brady et al., 2021; see also Mesoudi, 2008). *Reinforcement learning*, in contrast, occurs when people adjust their actions “in response to positive or negative social

feedback” (Brady et al., 2021). The link between norms and behavior can thus be automatic and devoid of conscious reflection or introspection (McDonald & Crandall, 2015) or, alternatively, rely on conscious and strategic anticipation of rewards and punishments. Actions inconsistent with an injunctive social norm, for instance, can elicit attention, social disapproval, ostracism, and reputational damage (Jacobson et al., 2011).⁵ Despite being informal, norms tend to be stable because they are acquired early in life and/or because they are reinforced through social feedback (Köbis et al., 2018). Nevertheless, both personal and social norms are amenable to influence and change, a topic to which the chapter will return in Sect. 38.4.

Benabou and Tirole (2011) integrate the insights of the microeconomic approach to corruption with those of norms-based approaches. They formally model an individual’s choice about whether or not to take a certain action as function of: the cost of taking the action (e.g., effort), the payoff to taking the action (e.g., a wage), intrinsic motivation (e.g., personal ethics or altruism), and a reputational payoff that depends on the action taken (which can also be interpreted as a self-image concern relating to social norms).⁶

An important class of models aims to explain why societies exhibit different levels of corruption. These models are premised on the idea of “frequency-dependent behavior,” meaning that individual behavior is influenced by the perceived societal rate of a behavior (i.e., by descriptive norms or empirical expectations). In these models, the existing level of corruption is itself a cause of individual corrupt behavior—in other words, “corruption may corrupt” (Andvig & Moene, 1990; Tirole, 1992; Aidt, 2003; Fisman & Golden, 2017; Shalvi, 2016; Gans-Morse et al., 2018; Stephenson, 2020).

Because these are effectively coordination games, they can give rise to multiple equilibria. The most common version in the literature produces two stable equilibria: one with low societal corruption and one with high societal corruption (Fig. 38.1).

The S-shaped solid line describes how individual behavior (vertical axis) depends on perceived aggregate behavior (horizontal axis). For example, if 60% of citizens (horizontal axis, A') are initially perceived to be acting corruptly (horizontal axis), the likelihood that an individual will act corruptly is approximately 70% (vertical axis, A'). If all individuals are similar and hold similar perceptions, this implies that the aggregate rate of corruption in the next period will be 70% (horizontal axis, B'). In that period, given that 70% are acting corruptly, the figure indicates that the likelihood that an individual acts corruptly is approximately 81% (vertical axis, B'), and this will become the new societal rate of corruption in the period after that one. By the same logic, this adjustment process will continue until the societal level of

⁵Descriptive norms can also influence behavior by conveying information about injunctive norms or normative expectations. In other words, observing that nobody “does X” can be understood to signal that others believe that “it is not acceptable to do X.”

⁶The model also considers a public-goods-like payoff stemming from the aggregate rate of the behavior.

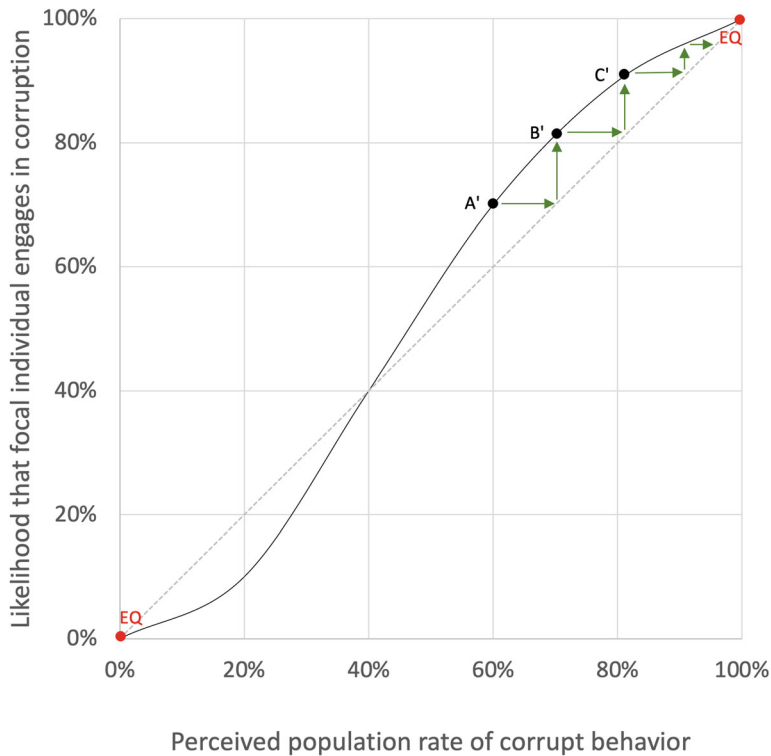


Fig. 38.1 Coordination model of corrupt behavior

corruption reaches the point where the curve crosses the 45° line to the right of the graph—which in this example happens at 100%. An aggregate level of corruption equal to 100% is a stable point. Small deviations away from 100% will again return, over time, to 100%. This is a high-corruption equilibrium.

Similarly, if the society starts out with a low-enough level of corruption (specifically, lower than 40% threshold, which is where the curve crosses the 45° line in the middle), it will converge toward the low-corruption equilibrium (0% corrupt behavior in this example). Suppose, for example, that 20% of people are initially corrupt. Knowing this, an individual has a 10% likelihood of choosing to engage in corruption, and therefore in the next period 10% will be the new societal level of corruption. Following this logic, the societal level of corruption will eventually reach the low-corruption equilibrium. Small deviations in societal corruption away from 0% will again converge to 0%, making this a stable, low-corruption equilibrium. It is in this sense—that the initial level of corruption determines the direction of over-time convergence—that corruption is “self-reinforcing” (Andvig & Moene, 1990; Aidt, 2003). This type of model can be used to think about potential policy interventions to reduce societal levels of corruption, and idea that we will return to in a later section of this chapter (but see Stephenson, 2020).

38.2.3 Empirical Evidence on Norms and Corrupt Behavior

A variety of empirical approaches have been utilized to study the relationship between informal norms and corrupt behavior. One approach compares behavior across individuals who reside in the same place but grew up in different societies and, therefore, internalized different social norms. Fisman and Miguel (2007), for example, compare the number of parking tickets accrued in New York City by United Nations diplomats from different countries. They find that diplomats from more-corrupt countries accumulated more parking tickets than those from less-corrupt countries, which they interpret as evidence that diplomats' moral norms, learned in their respective countries of origin, durably shaped their behavior while in New York City. Using a similar approach, Barr and Serra (2010) find that students from more-corrupt countries were more likely to bribe in a lab experiment; and DeBacker et al. (2015) show that US firms with owners born in more-corrupt countries engaged in more tax evasion than those from less-corrupt countries.

Those studies do not measure informal norms directly. Instead, they proxy for "corruption norms" using country-level corruption perceptions indices. Such indices are constructed by asking experts to rate country-wide levels of corruption, and thus may or may not accurately reflect a country's normative environment.⁷ A different way to measure norms is via survey questions. Surveys make it possible to measure descriptive and injunctive social norms as perceived by individuals (by asking how others are expected to behave or to judge a specific behavior) as well as to estimate the descriptive and injunctive norms that prevail in a society (by asking individuals about their behavior, or about how they judge a specific behavior, and averaging survey responses over a representative sample). Simpson (2020), for instance, uses European Social Survey items that ask people how wrong they consider bribery to be to construct country-level proxies of injunctive corruption norms. Focusing on second-generation immigrants, he finds that injunctive norms in the country of origin predict self-reported bribing behavior in the destination country. Survey questions can also be used to elicit personal norms.

Krupka and Weber (2013) introduce a general method to empirically elicit injunctive social norms. They depict a norm as a menu of possible actions, each associated with a rating about how acceptable the action is. To improve accuracy, participants are asked to pick the rating they believe best reflects the average across all participants, and correct responses are rewarded. In a laboratory experiment on bribing behavior, Banerjee (2016) applies this method to elicit participant approval of each of the possible bribery actions allowed in the corruption game in question (i.e., no bribery, bribe = 20 units, bribe = 40 units, and so on). Participants were asked to rate each action as: "very socially inappropriate," "socially inappropriate,"

⁷The World Bank's Control of Corruption indicator, for example, "captures perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as 'capture' of the state by elites and private interests" (World Bank, World Development Indicators).

“somewhat socially inappropriate,” “value neutral,” “somewhat socially appropriate,” “socially appropriate,” or “very socially appropriate.” Participants who picked the modal response were paid twice the participation fee. The purpose of that study was to compare whether framing the experiment as a bribery game, as opposed to framing it using neutral language, would yield different injunctive norms and different bribing behavior (they did). Other studies of corruption norms and behavior that employ this elicitation method include d’Adda et al. (2016) and Schram et al. (2022). Survey responses, of course, can be biased (e.g., by social desirability concerns) or inaccurate (e.g., when respondents do not pay attention). While incentives, as used in the Krupka-Weber method, can help, surveys are also not a perfect way to measure norms.

There is much to learn from scholarship focusing on dishonesty. While dishonesty and corruption are certainly not the same thing, they are conceptually related: corruption is one of many forms of dishonest behavior, while honest behavior, and norms that prescribe honesty, are incompatible with corruption. They are also empirically related. In a remarkable study, Cohn et al. (2019) handed more than 17,000 “lost” wallets to people in public buildings in 355 cities in 40 countries, and measured whether wallet recipients subsequently contacted the purported owners to return the wallets—an instance of honest behavior. Across countries, the rate of contact varied from under 10% in China to upwards of 80% in Denmark. A follow-up study (Sulitzeanu-Kenan et al., 2022) uses Cohn et al.’s (2019) data to investigate how wallet-return efforts correlated with corruption norms. They divide up the data according to whether a wallet was handed to a public- vs. a private-sector employee. They show that the level of wallet-return effort in the public sector correlates strongly with country-level corruption norms—proxied by two corruption perceptions indices.⁸

In a separate study, Gächter and Schulz (2016) measure individual-level dishonest behavior through a die-rolling experiment where the number of points obtained can be misreported by participants to increase their personal payout (Fischbacher & Föllmi-Heusi, 2013). They find that country-level rates of dishonesty predict cross-national variation in aggregate corruption (as well as in tax evasion and fraud) in a sample of 23 countries. These studies furnish evidence for a close relationship between dishonesty and corruption norms and behaviors.

General injunctive norms, that is, norms that do not specifically refer to corruption or honesty, have also been shown to correlate with dishonest or corrupt behavior. Consider, for example, the sociological characterization of informal norms of universalism vs. particularism as a continuum, where universalism holds

⁸Sulitzeanu-Kenan et al. (2022) regard wallet-return efforts not only as an honest behavior but also as indicative of personal norms of honesty. In addition to the result reported in the text, they document *within-country* variation in norms of honesty between the public and the private sector. Specifically, they compute the difference between the rate of wallet-return effort in the public sector minus the same rate in the private sector within every country (which they term “public-sector culture”), and show that this varies considerably across countries, being largest in magnitude in Sweden and Switzerland and smallest (and negative) in Russia and Serbia.

that what is correct is correct in every circumstance, and particularism posits that personal relationships take precedence over social conventions and obligations (Parsons & Shils, 1951).⁹ Insofar as particularistic norms value the wellbeing of friends and family over the public interest, they have been argued to provide fertile grounds for corrupt behavior. Rotondi and Stanca (2015) operationalize norms of particularism/universalism as the difference between agreement with the statements “It is occasionally alright to ignore the law and do what you want” and “You should always strictly obey the law, even if it means missing good opportunities.” They find that norms of particularism predict higher levels of self-reported bribery in comparison with norms of universalism.

To move beyond correlations and credibly identify causal effects, scholars have implemented randomized experiments. Köbis et al. (2015), Abbink et al. (2018), and Schram et al. (2022) randomly assign descriptive norms (empirical expectations) in the lab by informing participants about how many others engage in bribery. They find that participants are more willing to engage in bribery when they are told that bribery is widespread. Corbacho et al. (2016) conduct a field experiment where they manipulate exposure to information about the prevalence of corruption in Costa Rica. Participants exposed to information that corruption was increasing were more likely to report to be willing to offer a bribe to avoid paying a traffic ticket than control participants not exposed to the information. Interventions that inform about the rate of change of a behavior (as opposed to its level) are known as “dynamic norms” treatments (Sparkman and Walton, 2017). Overall, evidence from randomized experiments is consistent with the idea that descriptive norms influence individual corrupt behavior. It is important to keep in mind that descriptive norms could act on behavior directly as well as indirectly, e.g., by influencing perceived injunctive norms or normative expectations: observing that most of my peers bribe traffic policemen could lead me to infer that they also believe that bribing traffic policemen is an acceptable action. There is experimental evidence, in fact, that empirical expectations can shape normative ones. Banerjee et al. (2016), for example, present participants with a vignette describing a person who goes to a government office to obtain a driver’s license and is asked for a bribe. Within the vignette, they randomize information about how common bribery is in that situation. They find that bribery is considered more acceptable when it is depicted as being common.

38.2.4 Are Informal Norms and Behavior “Proportional”?

Any attempt to influence corrupt behavior by influencing norms must be underpinned by an understanding of the relationship between norms and behavior.

⁹Enke (2019) makes a similar distinction between universal vs. communal/relational values. The former “are said to reflect ‘universal’ impersonal principles, such as fairness, individual rights, and justice. These principles emphasize the welfare of all individuals in society equally.” At the same time, “morality is said to also include ‘communal’ or ‘relational’ values (such as in-group loyalty or the moral relevance of betrayal and respect) tied to particular groups or relationships” (p. 962).

In some cases, the relationship appears to be proportional or monotonic, meaning that stricter norms are associated with less corruption. This possibility comports with empirical findings from cross-national studies (e.g., Fisman & Miguel, 2007; Barr & Serra, 2010; DeBacker et al., 2015; Gächter & Schulz, 2016; Simpson, 2020).

Proportionality is also a feature of theoretical models about the relationship between individual behavior and descriptive norms, such as those discussed earlier in this chapter. Threshold models are a related class of models of individual behavior that rely on frequency-dependent behavior (Granovetter 1978). Such models predict that behavior changes when a descriptive or normative threshold is crossed. If thresholds vary from person to person, a “snowball” dynamic can arise (Kuran, 1989; Berger et al., 2023). Consider, for example, a judge j who is willing to act corruptly only if he believes that a fraction $Y_j\%$ of all judges is acting corruptly, and suppose that the threshold varies across judges. Initially, only a handful of judges act corruptly. Upon learning about this, judges with a low threshold $Y_j\%$ who had not previously acted corruptly now begin to do so. This increases the fraction of judges acting corruptly and, in turn, motivates judges with a slightly higher threshold to begin acting corruptly. Eventually, many judges will be corrupt. In general, models of frequency-dependent corrupt behavior predict proportionality between norms and behavior.¹⁰

While intuitively appealing, the assumption that behavior is monotonic in the underlying social norms is ultimately an empirical question. A recent study (Aycinena et al., 2022) provides evidence for a specific form of nonmonotonicity in the relationship between *injunctive* social norms and dishonest behavior. The authors elicit individual normative expectations (injunctive norms) about dishonesty and measure individual dishonest behavior in an exercise where individuals roll a die and are rewarded in proportion to the number of die-points they report to the researcher. They categorize respondents as Consequentialists if they believe that the extent of lying (i.e., the degree to which one overreports the number of points shown by the die, found in Gächter and Schulz (2016) to be associated with corruption) matters, so that lying to a greater extent is less acceptable than lying to a lesser extent. Respondents for whom the extent of lying does not matter—that is, those for whom lying a little bit is just as unacceptable as lying a lot—are called Deontists. Aycinena and coauthors find that those who espouse the strictest injunctive norms, namely the Deontists, are substantially *more* likely to lie to the maximal possible extent (i.e., report the number of points that yields the greatest reward, irrespective of their die-roll) than those who believe that the unacceptability of lying is proportional to the extent of the lie (the Consequentialists). This finding is consistent with a model where the cost of lying is constant irrespective of the extent of the lie. In such a model, it is optimal to lie to the greatest possible extent. I am not

¹⁰Note that, embedded in Bicchieri’s very definition of a social norm, there is an assumption of proportionality: A norm is a rule of behavior such that I am willing to engage in behavior X on condition that at least 50% of my reference group are likely to engage in X and to approve of X . Behavior changes when empirical and normative expectations cross the 50% threshold.

aware of an analogous study where the behavior of interest is corruption rather than lying. Nevertheless, some studies of corruption (e.g., Frank & Schulze, 2000) document a similar empirical pattern of the outcome variable, with some people engaging in a little corrupt behavior and others pursuing corruption to the maximal possible extent. Whether that pattern correlates nonmonotonically with underlying norms, as in Aycinena et al. (2022), is unknown.

In practice, the precise mapping between norms and behavior can be difficult to determine not only because of potential non-monotonicity in the relationship between norms and behavior, but also because individuals simultaneously espouse a wide range of personal and social norms that can pull their behavior in conflicting directions (McDonald et al., 2013). Urinboyev and Svensson's (2018) ethnographic study of an Uzbek village, for example, shows that injunctive social norms whereby individuals are expected to share economic and political resources with relatives and friends, put state officials in a position of choosing between fulfilling this norm and behaving according to norms of honesty. In a case study of the German corporation Siemens AG, Klinkhammer (2013) describes a similar conflict of norms where corrupt behavior constituted a failure of compliance with organizational norms of legality and, at the same time, was useful to the company's objectives. Importantly, when norms about corruption prescribe conflicting courses of action, it is not possible to infer norms directly from observed behavior.

Another pitfall to avoid is the labeling of specific norms as inherently "good" or "bad." Many norms can produce both desirable and undesirable behaviors depending on context. Norms of trust, reciprocity, and prosociality, generally regarded as good for society, can simultaneously underpin corrupt behavior. Finan and Schechter (2012), for example, find that norms of reciprocity in Paraguay support vote buying, an illegal practice commonly regarded to be corrupt. Reciprocity norms make the vote-buying exchange viable in a context where the vote is secret. Similarly, norms of altruism and generalized trust—also widely regarded as socially desirable—can render one-shot corrupt transactions, which are unenforceable in a court of law, viable (Rosenbaum et al., 2013).

38.3 Norms and the Law

Corrupt behavior is influenced by both norms and laws. The relationship between norms and laws is multifaceted. This section discusses the different ways in which norms and the law interact to produce behavior: by mutually influencing other and by complementing or substituting for each other.

38.3.1 Mutual Causation

Informal norms underlie much legal activity. The political process, for example, can lead lawmakers to pass laws banning or permitting environmental protections or governmental mechanisms of accountability, among other things, based on how the

public feels and votes—that is, based on social or moral norms. Laws can also be codified to reflect prevailing community practices and standards.

Laws can also cause norms. Laws, for example, can socialize people into internalizing personal and social norms, through what is known as the expressive power of the law (Sunstein, 1996). McAdams and Rasmusen (2007) explain: “By saying what people *should* do, even if there is no penalty, the law tries to shift or maintain tastes, and to educate a society’s newcomers—children and immigrants—in its norms.” The law can influence norms by communicating what society esteems. Benabou and Tirole (2011) model the expressive power of the law as stemming from an informational asymmetry where legislators have better information about the preferences of society than individual citizens. In this account, an expressive law conveys information about what society values. Empirically, Tankard and Paluck (2017) demonstrate the expressive power of the law in a setting unrelated to corruption. They show that a 2015 Supreme Court ruling in favor of gay marriage influenced perceived social norms of support for gay marriage.

The expressive power of the law can also be harnessed to stigmatize practices that might otherwise not have been recognized as unacceptable, including corruption. A major goal of Hong Kong’s Independent Commission Against Corruption (ICAC), established in 1973 in the wake of a grand corruption scandal, was to “educate the public against the evils of corruption” and “enlist and foster public support” (Clark 1993, cited in Skidmore, 1996 p.123). And threshold laws, such as those establishing blood-alcohol thresholds for driving, have been shown to influence injunctive norms attached to legal vs. illegal behaviors (Lane et al., 2023).

Finally, the law can influence norms by changing behaviors, in turn influencing habits and eventually tastes. One way in which laws can influence behavior is by changing public perceptions about the frequency of corrupt behavior. Laws that increase the rate of prosecutions or convictions for corrupt behavior, for example, can lead the public to perceive that corruption is more widespread than they thought, potentially “normalizing” corrupt behavior—an unintended consequence.

38.3.2 Norms as Moderators of the Effectiveness of the Law

Norms can influence the effectiveness of the law. Acemoglu and Jackson (2017) argue that the effectiveness of a law in regulating behavior depends on the degree to which it reflects community norms. And when social norms and laws sanction the same behavior, the norm may increase the level of compliance above that which the law could have attained on its own. For this reason, optimal legal punishment may need to be adjusted to reflect the existence of private punishment stemming from social norms (Cooter & Porat, 2001).

Personal and social norms about obeying the law are especially important for corrupt behavior, which is almost everywhere illegal. Norms can motivate individuals to follow the law because it is a law, even if the behavior in question is not considered inherently immoral or unacceptable. Thus, norms dictating obedience to the law can help to curb corrupt behavior in instances when social norms favor the

behavior. For example, social norms encouraging officials to help friends or relatives to obtain a government job could be counteracted by social norms dictating that laws must be obeyed (provided nepotism is illegal). The Italian saying that “in Milan traffic lights are instructions; in Rome, they are suggestions; in Naples, they are decorations,” reflects the notion that norms vary in the extent to which they dictate obedience to laws (Lisciandra, 2023). Gelfand et al. (2011) categorize societies as loose/tight depending on the degree to which they have weak (/strong) norms and tolerate (/fail to tolerate) behavior that strays from the norm. One would expect laws against corruption to wield more power in tight societies.

Community norms that impact a community’s capacity for collective action may also impact the effectiveness of the law in relation to corruption. Community monitoring interventions (CMI) consist of measures that enhance the power of communities to monitor their governments. CMIs have been found to be effective at curbing corruption, and the World Bank, the United Nations, and many NGOs devote substantial resources to supporting CMIs (Molina et al., 2017). A meta-analysis of CMI effectiveness finds that local norms that enable a community to act collectively and organize more efficiently—such as norms of interpersonal trust—significantly improve the community’s ability to support the rule of law and curb corruption (Molina et al., 2017; see also Nunn & Wantchekon, 2011).

Norms of professionalism also matter for obedience to the law, particularly for bureaucrats and elected officials, such as judges, police, and employees of agencies of the executive power. Evans (1992, 1995) ascribes the success of economic bureaucracies in developmental states, such as Japan’s Ministry of International Trade and Industry, to meritocratic recruitment and promotion practices, which strengthen esprit de corps and adherence to professional norms, providing a bulwark against corrupt temptations.

It is important to note that corrupt behavior differs from many other kinds of illegal or criminal behavior in that one of its main purposes is to circumvent or undermine the formal legal system (Muthukrishna et al., 2017). Construction companies, for example, frequently use corruption to obtain government contracts, and corporations of all kinds often attempt to influence the legislative process, and the political process in general, to enhance their market power at the expense of competitors. For these reasons, norms that render corruption acceptable, insofar as they increase the incidence of corrupt behavior, can undermine the very legitimacy of the process of enactment of laws and of law enforcement. In turn, when the creation and enforcement of laws is perceived to be corrupt, norms of obedience to the law are unlikely to be effective (or to even exist in the first place) (Acemoglu, 1995).

38.3.3 The Law as a Moderator of the Effectiveness of Norms

The law can also influence the effectiveness of informal norms. Community standards, for example, are norms that the law can invoke: “If a custom is a law, then legal officials are obligated to enforce it” (Cooter & Fikentscher, 1998, p. 315).

Laws can also crowd out personal norms. Crowding out takes place when a legal incentive—for example, the possibility of being fined or jailed—undermines intrinsic ethical motives or personal norms: “incentives that appeal to self-interest may fail when they undermine the moral values that lead people to act altruistically or in other public-spirited ways” (Bowles, 2008; see also Bohnet et al., 2001; Bénabou & Tirole, 2006; Underhill, 2016; Wang & Keith Murnighan, 2016).

In one of the few experimental studies of crowding out in the realm of corruption, Lambsdorff et al. (2023) find, in a laboratory setting, that introducing the possibility of punishment for bribe-taking has two effects that impact overall welfare in opposite ways. For some subjects, the introduction of punishment enhances the intrinsic motivation to perform, consistent with the idea that the law has expressive power. For other subjects, however, the introduction of punishment decreases intrinsic motivation, consistent with the law crowding out personal norms.

Field evidence for a crowding-out effect of punishment on honesty is furnished by Galeotti et al. (2021). They describe an experiment where French users of public transportation were randomly assigned to an inspection condition, where an inspector checked whether they had a valid ticket, or a no-inspection control. They regard the inspection condition as an instance of exposure to legal enforcement. They subsequently measured honest behavior.¹¹ Passengers exposed to the inspection condition behaved *less* honestly than the controls (who were not inspected). In other words, exposure to legal enforcement reduced (i.e., crowded out) intrinsic honesty.

Independent of the crowding-out argument, a number of studies suggest that informal norms matter most where formal institutions and the rule of law are weak. The underlying idea is that laws and formal institutions tend to be more powerful than informal norms. The study of corruption norms and parking tickets by Fisman and Miguel (2007) mentioned earlier, for example, finds that the importance of norms on parking tickets was greatly weakened by the introduction of legal punishment—specifically, by the removal of diplomatic immunity from the legal consequences of accruing parking tickets. This finding is echoed in studies of norms and the law in realms other than corruption. Grosjean (2014), for example, argues that persistent norms of honor among Scots-Irish migrants to the United States and their descendants—norms that underpinned violent behavior—faded as institutional quality in the South converged to that in the North. Consistent with this idea, models of intergenerational transmission of informal norms indeed suggest that a strong rule of law will discourage parents from passing along permissive corruption norms to their children (Bisin & Verdier, 2000; Doepke & Zilibotti, 2017). Of course, the strength of the rule of law may itself be endogenous to the prevalence of informal corruption-friendly norms, as argued earlier in this chapter. Ascertaining the conditions under which informal norms matter most as drivers of behavior merits further research.

¹¹ To measure honest behavior, they had an actor pretend to pick up a bank note from the ground and ask passengers (the study subjects) whether the bill was theirs. Saying that the bill was theirs (which it was not) counted as dishonest behavior, while saying the bill was not theirs counted as honest behavior.

38.3.4 Efficiency of Laws Vs. Norms

Are formal laws more efficient than informal norms at preventing corrupt behavior? Or are norms more efficient than laws? The NIE has emphasized that corrupt exchanges have the disadvantage, compared to formal legal exchanges, that they cannot be enforced in the courts, opening the door to opportunism and betrayal—and consequently increasing transaction costs. Along similar lines, some scholarship argues that obsolete norms, even if they were initially efficient, are difficult to change and to adapt to changing circumstances and therefore likely to become less efficient and effective over time (Posner, 1996; Kahan, 2000).

Game theoretic analysis has likewise pointed to a possible tension between efficient norms and those that can be sustained in equilibrium. In some circumstances, norms of cooperation (which hold efficiency advantages) are fragile to the presence of even small amounts of noncooperative behavior, and thus not sustainable in equilibrium (Mahoney & Sanchirico, 2001). An example of an inefficient norm that is nevertheless sustained in equilibrium is the norm of “amoral familism” (Banfield, 1967), which amounted to the collective expectation that “no one will further the interest of the group or community except as it is to his private advantage to do so” (pp. 83–84). Where efficient norms are fragile, the law can potentially help to sustain them (Mahoney & Sanchirico, 2001).

Conversely, it may be possible for informal norms to best formal laws in terms of efficiency. Transactions based on informal norms do not require police, courts, or monitoring and enforcement agencies, potentially providing some efficiency advantages. In general, the field of law and economics recognizes that the relative efficiency of formal laws vs. informal norms depends on context.

38.4 Informal Norms in the Fight Against Corruption

Looking along human history, only a minority of countries have successfully transitioned from patrimonial, extractive, or limited-access orders where status and connections largely determine outcomes and corruption is rampant, to universalist, meritocratic, egalitarian, and law-abiding systems where corruption is the exception (North et al., 2009; Acemoglu & Robinson, 2012). Such transitions threaten many entrenched interests and are therefore difficult to attain (Rothstein, 2011). Countries or territories that have successfully transitioned include Sweden, Denmark, Norway, Singapore, Taiwan, Hong Kong, and Japan, among others.

How exactly such momentous transitions might be attained remains the subject of ongoing debate and research, but the importance of informal norms is emphasized in various historical accounts. According to Rotberg, historical country-level transitions from particularism and high corruption to universalism and low corruption have ensued from energetic political will focused on “deep teaching, committed persuasion, and the effective mobilization of large arrays of people behind a clearly defined and intelligible project attractive to whole communities and legions of voters” to attain a shift “from a deep acceptance of corrupt behavior toward a robust

approval of new noncorrupt norms” (Rotberg, 2018, p. 13). Sweden’s patrimonial bureaucracy, for example, developed into a modern Weberian one over the course of the nineteenth century (Bågenholm, 2013). Rothstein observes that “not only formal but also informal institutions” were “important drivers” of this transformation (2011, p. 245), hinting at the importance of interpersonal trust and trust in the law. The spread of liberal ideas might have also played an important role (Rothstein & Teorell, 2015). Nevertheless, even if the historical record accords an important role to informal norms in the fight against corruption, the question of whether policy might effectively harness informal norms to that end remains an open one.

Purposefully engineering transitions out of patrimonialism has proven very challenging. In the modern world, legislation, institutional reform, and even social revolutions more often than not have failed to subvert the underlying set of entrenched relationships and political-economic incentives. A recent study of typical policy tools favored by the United Nations Convention against Corruption finds that such tools are unlikely to work well in the absence of the rule of law: “Countries with no proper rule of law have a problem of impunity and of selective, if not discriminatory legal enforcement” (Mungiu-Pippidi & Dadăsov, 2017; see also Mungiu-Pippidi, 2018, and Dadăsov, 2017).¹²

In light of the difficulty of tackling corruption, scholars and policymakers have displayed increasing openness to nontraditional and eclectic approaches. In what follows, I review a variety of potentially promising ideas and avenues for further research on anticorruption policy.

38.4.1 Personal Norms and Anticorruption

One category of approaches is based on *personal norms*. Influencing personal norms to render citizens more public-minded could potentially result in more-honest public servants and contribute to generating grassroots political pressure against corruption. A growing, but still young, body of research suggests that it is in fact possible to intervene to change personal norms (Paluck et al., 2021). Enduring change in personal norms is believed to require deep and active mental engagement (Goldberg et al., 2020), as in “perspective-taking” exercises, where individuals are encouraged to take another’s point of view (Gehlbach & Brinkworth, 2012; Broockman & Kalla, 2016; Kalla & Broockman, 2023).

Personal norms also lie at the basis of hiring policies aimed at curbing corruption. Such policies aim to screen individuals for intrinsic ethical motivation. Hannah and Wang (2017) find evidence of naturally occurring “negative selection” of university students into the Indian public service: students who displayed lower levels of

¹² Other scholarship is more optimistic about the potential effectiveness of anticorruption policies. Jancsics (2019), for example, suggests that customizing policy to specific contextual conditions can help.

honesty (in a Fischbacher & Föllmi-Heusi, 2013 die-rolling game) were more likely to express interest in a government job.

Screening policies can mitigate negative selection. Hannah and McIntyre (2018) suggest that potential hires can be tested for personality, civic-mindedness, and honesty. The Perry Public Sector Motivation Questionnaire, for example, has been shown to correlate with ethical behavior (Perry & Hondeghem, 2008). Field experiments have also found that it is possible to attract higher-quality applicants (as measured by the Perry Questionnaire) by offering higher wages (Dal Bó et al., 2013; but see Deserranno, 2015 for different evidence from Uganda) and by appealing to pro-social motivations when advertising positions (Ashraf et al., 2020).

38.4.2 Social Norms and Anticorruption

Social norms might also be harnessed to engineer change out of corruption. Consider the coordination models of corruption reviewed earlier in this chapter. A key insight of those models is that changing descriptive norms can change behavior society-wide by inducing convergence to a new equilibrium. Nyborg et al. (2016), among others, argue that public policy can be used to tip a society from one coordination equilibrium to another, by providing “reasons for people to change their expectations” (p. 43). In some instances, publicizing the facts might help. Specifically, when corruption levels are low, but the public erroneously believes that they are high, correcting the misperception can effectively shift descriptive norms in the desired direction (Shamir & Shamir, 1997) and jump-start an equilibrium adjustment process such as the one described in Fig. 38.1.

Social-norms nudges (henceforth “norms nudges”) are interventions that provide information aimed at influencing empirical and normative expectations (Bicchieri & Dimant, 2022), thereby harnessing preexisting social norms toward desired ends. Norms nudges have successfully been used to influence electricity consumption (Schultz et al., 2007; Allcott, 2011), environmental behavior (Goldstein et al., 2008), tax compliance (Castro & Scartascini, 2015), and voter turnout (Gerber et al., 2008), among other behaviors.

Evidence about the potential for norms nudges to durably change corrupt behavior, however, is still scant. Dimant et al. (2020) experimentally studied the effectiveness of norms nudges aimed at increasing honest behavior. To measure honest behavior, they asked participants to self-report the result of a private die roll and paid them an amount corresponding to the number of points, creating an incentive to misreport to maximize payment. They then experimentally nudged, in the treatment group, empirical expectations by stating that “the vast majority of participants were honest about the reported number in the dice task,” and normative expectations by stating that “the vast majority of participants said that one should be honest about the reported number in the dice task.” Discouragingly, they found precisely estimated null effects of these nudges. The intervention failed to change honest behavior. In fact, the norms nudges were unable to influence even empirical and normative expectations (potentially because of self-serving belief distortions; Bicchieri et al., 2023).

Two features of that experiment may have contributed to the ineffectiveness of the norm's nudges. First, a behavioral norm was previously in place (dishonesty was already viewed as unacceptable). Norms nudges could potentially be more successful where there is no clear, widespread norm. Second, their experiment by design precluded the possibility that dishonesty might be detected by a third party. Therefore, the reputational costs and social disapproval that commonly undergird the connection between social norms and individual behavior were absent. It is possible that the same nudges could have been effective in an environment where detection was a possibility.

Also on the research frontier, Dimant and Shalvi (2022) propose meta-nudging, whereby nudges are targeted toward social influencers who are in a position to enforce normative behavior among ordinary individuals. In this approach, ordinary individuals are indirectly affected by the nudging, via the influencers' enforcement work. In contrast with norms nudging, which aims to change the social norm, meta-nudging targets the *enforcement* of "good" norms.

The role of socially influential individuals is also highlighted by Bicchieri (2017), who describes influencers as social trendsetters who are more willing than the average person to bear the risk of behaving differently and initiating change. Prentice and Paluck (2020) similarly suggest that targeting "social referents" is a promising way to exploit the structure of a social network to effectively influence social norms. In this vein, Hoffman et al. (2017) recommend "empowering trendsetters that can galvanize behavioural change" as an anticorruption measure in present-day Nigeria. On the flip side, Berger and Heath (2008) warn that norms can be abandoned when an unpopular group adopts them.

One reason that social norms drive behavior is their effect on what psychologists term "honest-image concerns." Honest-image concerns refer to caring that one comes across to others as an honest person (Weisel & Shalvi, 2015). Honest-image concerns can play an important role when a corrupt transaction depends on one party making an offer to a second party—that is, in one-shot collaborative corruption. Concern about being judged negatively or rebuffed will naturally influence the first party's willingness to make a corrupt offer.

In a lab experiment focusing on corrupt behavior, Salmon and Serra (2017) experimentally manipulate the degree to which behavior can be socially observed. They find that observability influences corrupt behavior. They further hypothesize that average levels of honest-image concerns vary from country to country. Consistent with this, they find that the effect of social observability on corrupt behavior varies across countries. In a related study, Dorrough et al. (2023) find that the social stigma attached to bribery varies cross-nationally. This perspective suggests that interventions that stigmatize corrupt behavior, and thus render honest-image concerns more prominent, hold the potential to reduce the incidence of collaborative corruption.

A cautionary note is in order when designing anticorruption interventions based on social norms, because such interventions can sometimes backfire. Publicizing the fact that corrupt behavior is common, for example, can have the unintended effect of conveying the message that the behavior is socially approved of or that it will not elicit significant social sanctions, undoing the effect of the social norm against the behavior (Efferson et al., 2008; Muthukrishna et al., 2017; Bicchieri & Dimant, 2022).

38.5 Conclusion

Social and personal norms emerge as important determinants of corrupt behavior and as potential targets for policy interventions. Norms can underpin the enactment of laws. Norms can also enhance or undercut the effectiveness of laws. At the same time, formal laws, institutions, and material incentives have the potential to crowd out personal norms. How to change behavior by intervening on norms remains an active area of research. Much remains to be learned as well about the circumstances under which informal norms are important drivers of behavior in general and corrupt behavior in particular.

A single chapter cannot do justice to the full breadth of relevant knowledge. The hope is to contribute to catalyzing future theoretical and empirical work on corruption and anticorruption policy that seriously engages concepts and findings about informal norms from a variety of disciplines and research traditions. The foreseeable future promises to yield exciting new insights about the causes and consequences of social norms, their interaction with formal laws and institutions, and their joint effects on corrupt behavior at the level of the individual, the group, and society.

Acknowledgments Simpson thanks Mary M. Shirley and Claude Ménard for very helpful comments and suggestions, and acknowledges IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements de l'Avenir Program).

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Abbink, K., Freidin, E., Gangadharan, L., & Moro, R. (2018). The effect of social norms on bribe offers. *Journal of Law, Economics, and Organization*, 34(3), 457–474.
- Acemoglu, D. (1995). Reward structures and the allocation of talent. *European Economic Review*, 39, 17–33.
- Acemoglu, D., & Jackson, M. O. (2017). Social norms and the enforcement of laws. *Journal of the European Economic Association*, 15(2), 245–295.
- Acemoglu, D., & Robinson, J. A. (2012). *Why nations fail*. Crown.
- Aidt, T. S. (2003). Economic analysis of corruption: A survey. *Economic Journal*, 113(491), F632–F652.
- Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, 95(9–10), 1082–1095.
- Allcott, H., & Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, 104(10), 3003–3037.
- Andreoni, J. (1989). Giving with impure altruism: Applications to charity and Ricardian equivalence. *Journal of Political Economy*, 97(6), 1447–1458.
- Andvig, J. C., & Moene, K. O. (1990). How corruption may corrupt. *Journal of Economic Behaviour and Organization*, 13(1), 63–76.
- Ashraf, N., Bandiera, O., Davenport, E., & Lee, S. S. (2020). Losing prosociality in the quest for talent? Sorting, selection, and productivity in the delivery of public services. *American Economic Review*, 110(5), 1355–1394.

- Aycinena, D., Rentschler, L., Beranek, B., & Schulz, J. F. (2022). Social norms and dishonesty across societies. *Proceedings of the National Academy of Sciences*, 119(31), e2120138119.
- Bågenholm, A. (2013). The case of modern Sweden. In M. F. Jensen & James Kennedy (Coords.), *Fighting corruption in modernity: A literature review: The Danish history of corruption and anti-corruption, 1660–1900*. https://pure.au.dk/ws/portalfiles/portal/56074642/fighting_corruption_in_modernity_a_literature_review_1.pdf
- Banerjee, R. (2016). On the interpretation of bribery in a laboratory corruption game: Moral frames and social norms. *Experimental Economics*, 19, 240–267.
- Banerjee, R., Baul, T., & Rosenblat, T. (2016). Social norms regarding bribing in India: An experimental analysis. *Journal of Contextual Economics–Schmollers Jahrbuch*, 136(2), 171–197.
- Banfield, E. C., & with assistance of Banfield, L. F. (1967). *The moral basis of a backward society*. Free Press.
- Bardhan, P. (1997). Corruption and development: A review of issues. *Journal of Economic Literature*, 35, 1320–1346.
- Barfort, S., Harmon, N. A., Hjorth, F., & Olsen, A. L. (2019). Sustaining honesty in public service. *American Economic Journal: Economic Policy*, 11(4), 96–123.
- Barr, A., & Serra, D. (2010). Corruption and culture: An experimental analysis. *Journal of Public Economics*, 94(11–12), 862–869.
- Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, 76(March/April), 169–217.
- Becker, G. S., & Stigler, G. J. (1974). Law enforcement, malfeasance, and compensation for employees. *The Journal of Legal Studies*, 3(January), 1–18.
- Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5), 1652–1678.
- Benabou, R., & Tirole, J. (2011). *Laws and norms* (No. w17579). National Bureau of Economic Research.
- Berger, J., & Heath, C. (2008). Who drives divergence? Identity signaling, outgroup dissimilarity, and the abandonment of cultural tastes. *Journal of Personality and Social Psychology*, 95(3), 593.
- Berger, J., Efferson, C., & Vogt, S. (2023). Tipping pro-environmental norm diffusion at scale: Opportunities and limitations. *Behavioural Public Policy*, 7(3), 581–606.
- Bicchieri, C. (2017). *Norms in the wild: How to diagnose, measure, and change social norms*. Oxford University Press.
- Bicchieri, C., & Dimant, E. (2022). Nudging with care: The risks and benefits of social information. *Public Choice*, 191(3–4), 443–464.
- Bicchieri, C., Dimant, E., & Sonderegger, S. (2023). It's not a lie if you believe the norm does not apply: Conditional norm-following and belief distortion. *Games and Economic Behavior*, 138, 321–354.
- Bisin, A., & Verdier, T. (2000). Beyond the melting pot. *Quarterly Journal of Economics*, 115, 955–988.
- Bohnet, I., Frey, B. S., & Huck, S. (2001). More order with less law: On contract enforcement, trust, and crowding. *American Political Science Review*, 95(1), 131–144.
- Bowles, S. (2008). Policies designed for self-interested citizens may undermine 'the moral sentiments': Evidence from economic experiments. *Science*, 320(5883), 1605–1609.
- Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33), eabe5641.
- Broockman, D., & Kalla, J. (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science*, 352(6282), 220–224.
- Buchan, B., & Hill, L. (2014). *An intellectual history of political corruption* (Political corruption and governance series). Palgrave Macmillan.
- Buchanan, J. M., & Tullock, G. (1962). *The calculus of consent*. University of Michigan Press.
- Castro, L., & Scartascini, C. (2015). Tax compliance and enforcement in the pampas evidence from a field experiment. *Journal of Economic Behavior & Organization*, 116, 65–82.

- Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, 58(6), 1015.
- Cohn, A., Maréchal, M. A., Tannenbaum, D., & Zünd, C. L. (2019). Civic honesty around the globe. *Science*, 365(6448), 70–73.
- Cooter, R. D., & Fikentscher, W. (1998). Indian common law: The role of custom in American Indian Tribal Courts (part I of II). *American Journal of Comparative Law*, 46(2), 287–337.
- Cooter, R., & Porat, A. (2001). Should courts deduct nonlegal sanctions from damages? *The Journal of Legal Studies*, 30(2), 401–422.
- Corbacho, A., Gingerich, D. W., Oliveros, V., & Ruiz-Vega, M. (2016). Corruption as a self-fulfilling prophecy: Evidence from a survey experiment in Costa Rica. *American Journal of Political Science*, 60(4), 1077–1092.
- d’Adda, G., Drouvelis, M., & Nosenzo, D. (2016). Norm elicitation in within-subject designs: Testing for order effects. *Journal of Behavioral and Experimental Economics*, 62, 1–7.
- Dadásov, R. (2017). European aid and governance: Does the source matter? *European Journal of Development Research*, 29, 269–288.
- Dal Bó, E., Finan, F., & Rossi, M. (2013). Strengthening state capabilities: The role of financial incentives in the call to public service. *Quarterly Journal of Economics*, 128(3), 1169–1218.
- DeBacker, J., Heim, B. T., & Tran, A. (2015). Importing corruption culture from overseas. *Journal of Financial Economics*, 117(1), 122–138.
- Deserranno, E. (2015). *Financial incentives as signals: Experimental evidence from the recruitment of health workers* (Working paper).
- Dimant, E., & Shalvi, S. (2022). Meta-nudging honesty: Past, present, and future of the research frontier. *Current Opinion in Psychology*, 47, 101426.
- Dimant, E., Van Kleef, G. A., & Shalvi, S. (2020). Requiem for a nudge: Framing effects in nudging honesty. *Journal of Economic Behavior & Organization*, 172, 247–266.
- Doepke, M., & Zilibotti, F. (2017). Parenting with style: Altruism and paternalism in intergenerational preference transmission. *Econometrica*, 85(5), 1331–1371.
- Dorough, A. R., Köbis, N., Irlenbusch, B., Shalvi, S., & Glöckner, A. (2023). Conditional bribery: Insights from incentivized experiments across 18 nations. *Proceedings of the National Academy of Sciences*, 120(18), e2209731120.
- Efferson, C., Lalive, R., Richerson, P. J., McElreath, R., & Lubell, M. (2008). Conformists and mavericks: The empirics of frequency-dependent cultural transmission. *Evolution and Human Behavior*, 29(1), 56–64.
- Enke, B. (2019). Kinship, cooperation, and the evolution of moral systems. *Quarterly Journal of Economics*, 134, 953–1019.
- Evans, P. (1992). The state as problem and solution: Predation, embedded autonomy, and structural change. In S. Haggard & R. R. Kaufman (Eds.), *The politics of economic adjustment: International constraints, distributive conflicts and the state*. Princeton University Press.
- Evans, P. B. (1995). *Embedded autonomy: States and industrial transformation*. Princeton University Press.
- Finan, F., & Schechter, L. (2012). Vote-buying and reciprocity. *Econometrica*, 80(2), 863–881.
- Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise – An experimental study on cheating. *Journal of the European Economic Association*, 11(3), 525–547.
- Fisman, R., & Golden, M. A. (2017). *Corruption: What everyone needs to know*. Oxford University Press.
- Fisman, R., & Miguel, E. (2007). Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets. *Journal of Political Economy*, 115(6), 1020–1048.
- Frank, B., & Schulze, G. G. (2000). Does economics make citizens corrupt? *Journal of Economic Behavior & Organization*, 43(1), 101–113.
- Frisk Jensen, M., & Kennedy, J. (Coords.). (2013). *Two literature reviews on the pre-modern and modern categories of cases respectively*. Manuscript, ANTICORRP project (EU Grant 290529).
- Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595), 496–499.

- Galeotti, F., Maggiani, V., & Villeval, M. C. (2021). Fraud deterrence institutions reduce intrinsic honesty. *Economic Journal*, 131(638), 2508–2528.
- Gans-Morse, J., Borges, M., Makarin, A., Mannah-Blankson, T., Nickow, A., & Zhang, D. (2018). Reducing bureaucratic corruption: Interdisciplinary perspectives on what works. *World Development*, 105, 171–188.
- Gehlbach, H., & Brinkworth, M. E. (2012). The social perspective taking process: Strategies and sources of evidence in taking another's perspective. *Teachers College Record*, 114, 1–29.
- Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., et al. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, 332(6033), 1100–1104.
- Gerber, A. S., & Rogers, T. (2009). Descriptive social norms and motivation to vote: Everybody's voting and so should you. *Journal of Politics*, 71(1), 78–191.
- Gerber, A. S., Green, D. P., & Larimer, C. W. (2008). Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review*, 102(1), 33–48.
- Goldberg, M. H., Gustafson, A., & Van Der Linden, S. (2020). Leveraging social science to generate lasting engagement with climate change solutions. *One Earth*, 3(3), 314–324.
- Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research*, 35(3), 472–482.
- Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83(6), 1420–1443.
- Grosjean, P. (2014). A history of violence: The culture of honor and homicide in the US South. *Journal of the European Economic Association*, 12(5), 1285–1316.
- Hanna, R., Bishop, S., Nadel, S., Scheffler, G., & Durlacher, K. (2011). *The effectiveness of anti-corruption policy: What has worked, what hasn't, and what we don't know—a systematic review*. Technical report. EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.
- Hannah, R., & McIntyre, V. (2018). *New possibilities for cutting corruption in the public sector*. Harvard Kennedy School. <https://epod.cid.harvard.edu/article/new-possibilities-cutting-corruption-public-sector>
- Hannah, R., & Wang, S.-Y. (2017). Dishonesty and selection into public service: Evidence from India. *American Economic Journal: Economic Policy*, 9(3), 262–290.
- Hoffman, L. K., Patel, R., & Arkwright, H. E. P. (2017). *Collective action on corruption in Nigeria: A social norms approach to connecting society and institutions*. Chatham House.
- Jacobson, R. P., Mortensen, C. R., & Cialdini, R. B. (2011). Bodies obliged and unbound: Differentiated response tendencies for injunctive and descriptive social norms. *Journal of Personality and Social Psychology*, 100, 433–448.
- Jancsics, D. (2019). Corruption as resource transfer: An interdisciplinary synthesis. *Public Administration Review*, 79(4), 523–537.
- Jensen, M. F. (2018). The building of the Scandinavian states: Establishing Weberian bureaucracy and curbing corruption from the mid-seventeenth to the end of the nineteenth century. In *Bureaucracy and society in transition* (Comparative Social Research, Vol. 33, pp. 179–203). Emerald Publishing Limited. <https://doi.org/10.1108/S0195-631020180000033013>
- Kahan, D. M. (2000). *Gentle nudges vs. hard shoves: Solving the sticky norms problem* (Vol. 536, pp. 607–645). University of Chicago Law Review.
- Kalla, J. L., & Broockman, D. E. (2023). Which narrative strategies durably reduce prejudice? Evidence from field and survey experiments supporting the efficacy of perspective-getting. *American Journal of Political Science*, 67(1), 185–204.
- Klinkhammer, J. (2013). On the dark side of the code: organizational challenges to an effective anti-corruption strategy. *Crime, Law and Social Change*, 60, 191–208.
- Klitgaard, R. E. (1988). *Controlling corruption*. University of California Press.
- Köbis, N., van Prooijen, J.-W., Righetti, F., & Van Lange, P. A. M. (2015). “Who Doesn't?” – The impact of descriptive norms on corruption. *PLoS One*, 10(6), e0131830. <https://doi.org/10.1371/journal.pone.0131830>

- Köbis, N., Carter, D. I., & Starke, C. (2018). A social psychological view on the social norms of corruption. In I. Kubbe & A. Engelbert (Eds.), *Corruption and norms: Why informal rules matter* (Political corruption & governance series). Palgrave Macmillan.
- Krupka, E., & Weber, R. A. (2009). The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, 30(3), 307–320.
- Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3), 495–524.
- Kubbe, I., & Engelbert, A. (Eds.). (2018). *Corruption and norms: Why informal rules matter* (Political corruption & governance series). Palgrave Macmillan.
- Kuran, T. (1989). Sparks and prairie fires: A theory of unanticipated political revolution. *Public Choice*, 61(1), 41–74.
- Lambsdorff, J. G., Taube, M., & Schramm, M. (Eds.). (2005). *The new institutional economics of corruption*. Routledge.
- Lambsdorff, J. G., Grubiak, K., & Werner, K. (2023). *Intrinsic motivation vs. corruption? Experimental evidence on the performance of officials*. Unpublished manuscript, University of Passau.
- Lane, T., Nosenzo, D., & Sonderegger, S. (2023). Law and norms: Empirical evidence. *American Economic Review*, 113(5), 1255–1293.
- Leff, N. (1964). Economic development through corruption. *American Behavioral Scientist*, 8, 8–14.
- Lisciandra, C. (2023). Norms2: Norms about norms. *Erkenntnis*, 88(6), 2673–2694.
- Mahoney, P. G., & Sanchirico, C. W. (2001). Competing norms and social evolution: Is the fittest norm efficient? *University of Pennsylvania Law Review*, 149(6), 2027–2062.
- Mauro, P. (1995). Corruption and growth. *Quarterly Journal of Economics*, 110(August), 681–712.
- Mauro, P. (1997). The effects of corruption on growth, investment, and government expenditure: A cross-country analysis. In K. A. Elliott (Ed.), *Corruption in the global economy*. Institute for International Economics.
- Mauro, P. (1998). Corruption and composition of government expenditure. *Journal of Public Economics*, 69, 263–279.
- McAdams, R. H., & Rasmusen, E. B. (2007). Norms and the law. *Handbook of Law and Economics*, 2, 1573–1618.
- McDonald, R. I., & Crandall, C. S. (2015). Social norms and social influence. *Current Opinion in Behavioral Sciences*, 3, 147–151.
- McDonald, R. I., Fielding, K. S., & Louis, W. R. (2013). Energizing and demotivating effects of norm-conflict. *Personality and Social Psychology Bulletin*, 39, 57–72.
- Mesoudi, A. (2008). An experimental simulation of the “copy-successful-individuals” cultural learning strategy: Adaptive landscapes, producer–scrounger dynamics, and informational access costs. *Evolution and Human Behavior*, 29(5), 350–363.
- Molina, E., Carella, L., Pacheco, A., Cruces, G., & Gasparini, L. (2017). Community monitoring interventions to curb corruption and increase access and quality in service delivery: A systematic review. *Journal of Development Effectiveness*, 9(4), 462–499.
- Mungiu-Pippidi, A. (2018). Seven steps to control of corruption: The road map. *Daedalus*, 147(3), 20–34.
- Mungiu-Pippidi, A., & Dadásov, R. (2017). When do anticorruption laws matter? The evidence on public integrity enabling contexts. *Crime Law and Social Change*, 68, 387–402.
- Muthukrishna, M., Francois, P., Pourahmadi, S., & Henrich, J. (2017). Corrupting cooperation and how anti-corruption strategies may backfire. *Nature Human Behaviour*, 1(7), 0138.
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- North, D. C. (2005). *Understanding the process of economic change*. Princeton University Press.
- North, D. C., Wallis, J. J., & Weingast, B. R. (2009). *Violence and social orders*. Cambridge University Press.
- Nunn, N., & Wantchekon, L. (2011). The slave trade and the origins of mistrust in Africa. *American Economic Review*, 101(7), 3221–3252.

- Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Maja Schlüter, W., Adger, N., et al. (2016). Social norms as solutions. *Science*, 354(6308), 42–43.
- Olken, B. A., & Pande, R. (2012). Corruption in developing countries. *Annual Review of Economics*, 4, 479–509.
- Over, H. (2020). The social function of imitation in development. *Annual Review of Developmental Psychology*, 2(1), 93–109.
- Paluck, E. L., Porat, R., Clark, C. S., & Green, D. P. (2021). Prejudice reduction: Progress and challenges. *Annual Review of Psychology*, 72, 533–560.
- Parsons, T., & Shils, E. A. (Eds.). (1951). *Toward a general theory of action*. Harvard University Press.
- Perry, J., & Hondeghem, A. (2008). Building theory and empirical evidence about public service motivation. *International Public Management Journal*, 11(1), 3–12.
- Posner, E. A. (1996). Law, economics, and inefficient norms. *University of Pennsylvania Law Review*, 144(5), 1697–1744.
- Prentice, D., & Paluck, E. L. (2020). Engineering social change using social norms: Lessons from the study of collective action. *Current Opinion in Psychology*, 3, 138–142.
- Rose-Ackerman, S. (1978). *Corruption: A study in political economy*. Academic Press.
- Rosenbaum, M., Billinger, S., & Stieglitz, N. (2013). Private virtues, public vices: Social norms and corruption. *International Journal of Development*, 12(3), 192–212.
- Rotberg, R. I. (2018). Accomplishing anticorruption: Propositions & methods. *Daedalus*, 147(3), 5–18.
- Rothstein, B. (2011). Anti-corruption: The indirect ‘big bang’ approach. *Review of International Political Economy*, 18(2), 228–250. <https://doi.org/10.1080/09692291003607834>
- Rothstein, B., & Teorell, J. (2015). Getting to Sweden, part II: Breaking with corruption in the nineteenth century. *Scandinavian Political Studies*, 38(3), 238–254.
- Rotondi, V., & Stanca, L. (2015). The effect of particularism on corruption: Theory and empirical evidence. *Journal of Economic Psychology*, 51(2015), 219–235.
- Salmon, T. C., & Serra, D. (2017). Corruption, social judgment and culture: An experiment. *Journal of Economic Behavior & Organization*, 142, 64–78.
- Schram, A., Di Zheng, J., & Zhuravleva, T. (2022). Corruption: A cross-country comparison of contagion and conformism. *Journal of Economic Behavior and Organization*, 193, 497–518.
- Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5), 429–434.
- Shalvi, S. (2016). Corruption corrupts. *Nature*, 531(7595), 456–457.
- Shamir, J., & Shamir, M. (1997). Pluralistic ignorance across issues and over time: Information cues and biases. *Public Opinion Quarterly*, 61, 227–260.
- Shleifer, A., & Vishny, R. W. (1993). Corruption. *Quarterly Journal of Economics*, 108(August), 599–617.
- Simpser, A. (2020). The culture of corruption across generations: An empirical study of bribery attitudes and behavior. *Journal of Politics*, 82(4), 1373–1389.
- Skidmore, M. J. (1996). Promise and peril in combating corruption: Hong Kong’s ICAC. *Annals of the American Academy of Political and Social Science*, 547, 118–130.
- Sole-Olle, A., & Sorribas-Navarro, P. (2014). *Does corruption erode trust in government? Evidence from a recent surge of local scandals in Spain* (IEB working paper No. 2014/26). IEB.
- Spadaro, G., Molho, C., Van Prooijen, J. W., Romano, A., Mosso, C. O., & Van Lange, P. A. (2023). Corrupt third parties undermine trust and prosocial behaviour between people. *Nature Human Behaviour*, 7(1), 46–54.
- Sparkman, G., & Walton, G. M. (2017). Dynamic norms promote sustainable behavior, even if it is counternormative. *Psychological Science* 28(11), 1663–1674.
- Stephenson, M. C. (2020). Corruption as a self-reinforcing trap: Implications for reform strategy. *The World Bank Research Observer*, 35(2), 192–226.
- Sugden, R. (1982). On the economics of philanthropy. *Economic Journal*, 92(366), 341–350.

- Sulitzeanu-Kenan, R., Tepe, M., & Yair, O. (2022). Public-sector honesty and corruption: Field evidence from 40 countries. *Journal of Public Administration Research and Theory*, 32(2), 310–325.
- Sunstein, C. R. (1996). Social norms and social roles. *Columbia Law Review*, 96, 903.
- Tankard, M. E., & Paluck, E. L. (2017). The effect of a supreme court decision regarding gay marriage on social norms and personal attitudes. *Psychological Science*, 28(9), 1334–1344.
- Tanzi, V. (1998). Corruption around the world: Causes, consequences, scope, and cures. *IMF Staff Papers*, 45(4), 559–594.
- Tirole, J. (1992). *Persistence of corruption* (Working paper No. 2282-2019-4176). Institute for Policy Reform.
- Treisman, D. (2000). The causes of corruption: A cross-national study. *Journal of Public Economics*, 76, 399–457.
- Treisman, D. (2007). What have we learned about the causes of corruption from ten years of cross-national empirical research? *Annual Review of Political Science*, 10, 211–244.
- Underhill, K. (2016). When extrinsic incentives displace intrinsic motivation: Designing legal carrots and sticks to confront the challenge of motivational crowding-out. *Yale Journal on Regulation*, 33, 213.
- Urinboyev, R., & Svensson, M. (2018). Corruption, social norms and everyday life in Uzbekistan. In I. Kubbe & A. Engelbert (Eds.), *Corruption and norms: Why informal rules matter* (Political corruption and governance series). Palgrave Macmillan.
- Wang, L., & Keith Murnighan, J. (2016). How much does honesty cost? Small bonuses can motivate ethical behavior. *Management Science*, 63(9), 2903–2914.
- Weisel, O., & Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences*, 112(34), 10651–10656.
- World Bank. (1997a, August 13). *Helping countries combat corruption: The role of the World Bank*. Pamphlet, Poverty Reduction and Economic Management (PREM) Network, R97–201.
- World Bank. (1997b). *Helping countries combat corruption: The role of the World Bank* (Working paper 24687).
- World Bank. (2017). *World Development Report 2017: Governance and the law*. World Bank.
- Zhang, M., Zhang, H., Li Zhang, X., Peng, J. Z., Liu, D., & You, S. (2023). Corruption, anti-corruption, and economic development. *Humanities and Social Sciences Communications*, 10(1), 1–13.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Part VIII

Methodology

Combining Machine Learning and Econometrics to Examine the Historical Roots of Institutions and Cultures

39

Peter Grajzl and Peter Murrell

Contents

39.1	Introduction	1030
39.2	Machine Learning Applied to Institutions and Culture in Historical Settings: A Scattered Literature	1032
39.2.1	Addressing Challenges in Conventional Historical Research: Data Collection	1033
39.2.2	Addressing Challenges in Conventional Historical Research: Data Analysis	1033
39.2.3	Leveraging Text-as-Data	1035
39.3	The Corpora Utilized in Our Analysis	1035
39.3.1	The English Reports (ER)	1035
39.3.2	The Early English Books Online-Text Creation Partnership (EEBO- TCP)	1036
39.3.3	Pre-Processing the Corpora	1037
39.4	Three Example Techniques	1038
39.4.1	Topic Modeling	1038
39.4.2	Measuring Novelty Using Words	1042
39.4.3	Word Embeddings to Characterize Change in Ideas	1043
39.5	A Necessary Digression: A Different Methodological Perspective?	1045
39.6	From Machine Learning to Econometrics	1047
39.6.1	Example 1: Uncovering the Legal Legacies of Early English Caselaw	1048

P. Grajzl

Department of Economics, The Williams School of Commerce, Economics, and Politics,
Washington and Lee University, Lexington, VA, USA

CESifo, Munich, Germany

e-mail: grajzlp@wlu.edu

P. Murrell (✉)

Department of Economics, University of Maryland, College Park, MD, USA

e-mail: pmurrell@umd.edu

39.6.2	Example 2: Assessing the Effects of Caselaw on Economic Performance ...	1049
39.6.3	Example 3: Estimating the Coevolution of Ideas	1051
39.6.4	Example 4: Finding Critical Junctures	1052
39.7	Conclusion	1053
	References	1055

Abstract

Machine learning (ML) and associated computational advances have opened entirely new avenues for the processing and analysis of large data sets, especially those containing text. In this chapter, we show how ML can extend the scope of historical institutional and cultural analysis. We first provide an overview of some of the scattered existing literature using ML methods to study historical institutions and culture. We then use our own work on pre-nineteenth-century English caselaw and print culture to illustrate the possibilities and the challenges in using ML as a tool for systematic quantitative inquiry into the origins, change, and impact of institutions and culture. We highlight the power of ML for distilling core facts from large corpora and generating datasets amenable to analysis using conventional econometric analysis. We demonstrate how our work allowed us to explore the deep institutional roots of specific legal and cultural ideas, analyze the coevolution of ideas within caselaw and culture, examine the impact of caselaw on economic development both before and during the Industrial Revolution, and discern critical junctures in England’s legal and cultural development. Focusing on historical institutions and culture, the chapter illuminates the types of lessons that can be learned from the application of ML in new institutional economics. It also suggests a pathway that researchers applying ML to history can follow when trying to find a practical, implementable set of methods among the proliferation of new techniques that is usual when an area of research is in its infancy.

Keywords

Machine learning · Econometrics · England · Topic modeling · Word embeddings · VAR · Texts · Law · Culture

39.1 Introduction

We provide an overview of our own experience in combining machine learning (ML) and more traditional econometrics to investigate the historical development of institutions and culture. This experience was gained while building a portfolio of research projects on early-modern England, but the lessons we have learned could apply in any historical context. Our intended audience is NIE scholars who are currently interested in developments in this area of research, but are, like we have been, primarily focused on the application of tried-and-tested techniques. We hope to convey two sets of insights in tandem: first, what these new methodological tools can add to the more traditional literature on the origins, evolution, and effects of

institutions and culture; second, the insights into English institutional and cultural development that have arisen in our research.

An overarching element of our research follows McCloskey (2016, p. xiii): “Our riches did not come from piling brick on brick...but from piling idea on idea.” Institutional change is thoroughly linked to that process of accumulation of ideas, which are part of culture. In most contexts, the raw data on institutions and culture appears as text, not numbers. This dominance of text is multiplied if, as in our work, the primary focus is on developments before 1800. Then, it is entirely possible that a complete empirical project will have to proceed without the input of any pre-existing numerical data.

The challenges abound. One is the size of text data sets, both in terms of the number of observations and the dimensions of the core element of each observation, a legal opinion or a sermon, perhaps. For English at least, language usage became much more uniform only during the nineteenth century. In the English of the seventeenth century, one might find many different spellings of exactly the same word. Any given word in a seventeenth-century document cannot be relied upon to have the same meaning as it would have today. Untranslated Latin words appear often. The past is a foreign country where Google Translate does not reach.

Gradually over the last few decades, these challenges have receded in difficulty. Desktop computers can now handle datasets of the size likely to be available for early modern history. The bespoke software and dictionaries needed to standardize spelling and translate Latin words can be created with user-friendly programming languages. Powerful off-the-shelf ML tools can reduce the dimensionality of individual observations. These same tools can represent the text data in numerical form, making the resultant datasets analyzable using existing econometric methods. By summarizing documents as mixtures of ideas, the methods facilitate an understanding of the estimated ideas that reflects the specific context, even though those ideas are expressed in a seventeenth-century English that is very different from our own. By leveraging word-use context, individual words can be represented as multi-dimensional numerical vectors, thereby allowing one to identify synonyms by finding those word-vectors that are most similar to the vector representing a word of interest. Then, a study of a word’s close synonyms gives insights into the way that the word was used in context, for example in seventeenth-century English legal opinions. These are truly tools to produce “verbal thermometers” (McCloskey, 2016, p. 236). None of this is to say that the computer is a substitute for the close reading of texts. Rather, it complements reading as the researcher tries to figure out history’s puzzles. And in sharp contrast to human analysts, these ML tools can synthesize and summarize the content of tens of thousands of texts without importing preconceptions.

Any researcher who has discovered a treasure trove of texts and wants to apply ML will be immediately struck by the large variety of tools that are available and the lack of a clear guide as to which ones will be most useful. This is a characteristic of any powerful new methodology in its infancy. In addition to the proliferation of different tools, many papers contribute ad hoc methods of summarizing the content of the texts. There is nothing that corresponds to a core econometrics textbook that leads one easily into the methodological literature. There are not readily identifiable

schools of thought to orient one's thinking, such as the clear divide between structural and reduced-form methods in econometrics. To be sure, there are very good reviews, but the differences in perspective between these attest to the fact that this area of research has not reached a level of maturity where it is obvious where to begin a rigorous piece of applied work that is focused purely on generating new substantive insights. And these reviews are certainly not focused on the trials and tribulations of dealing with pre-nineteenth-century data. For examples of such works, see Burkov (2019), Athey and Imbens (2019), Gentzkow et al. (2019), Prüfer and Prüfer (2020), James et al. (2023), and Grimmer et al. (2021a).

We have no ambition of bringing order to the inchoate sets of methodologies. This is not a work of synthesis. Rather, by showing the reader what we have done, we hope to convey one viable route to producing new sets of quantitative insights from texts. By emphasizing that the combination of ML methods and econometrics is the cornerstone upon which a research project is built, we highlight the types of fundamental substantive results that one could hope to generate. Indeed, some of our findings recast elements of conventional wisdom that have been in place for centuries.

To set the stage, we begin in Sect. 39.2 with a brief overview of the scattered literature that uses ML to advance the understanding of the historical origins and evolution of institutions and culture. Sections 39.3, 39.4, 39.5, and 39.6 chronicle the processes and the deliberations involved in our application of ML. We describe text preprocessing and then the key ML methods relevant to institutional economists primarily interested in learning about the past. We discuss the methodological paradigm that the use of ML entails. We also clarify that the real power of ML appears when it is used to develop a comprehensive empirical overview of institutional and cultural development. We illustrate this point with examples from our own research. Section 39.7 concludes with broader reflections on the use of ML and empirical research in economics.

39.2 Machine Learning Applied to Institutions and Culture in Historical Settings: A Scattered Literature

With institutions and culture now occupying a central place in theories and narratives of comparative economic development, scholars have turned to empirically examining the historical origins and evolution of institutions and culture. In this pursuit, new ML techniques have emerged as a valuable complement to traditional econometrics. Here, we point to a sampling of studies demonstrating the power of the new methods either as a means of addressing preexisting empirical challenges or, more interestingly, charting new research avenues. We focus primarily on scholarship in the era before the nineteenth century. Because this earlier setting inherently comes with an abundance of data challenges, it facilitates discussion of the productive application of the new methodological tools as well as provides insights into their limits.

As we clarify below, the research examined in this section addresses institutions and cultures scattered across a wide range of subperiods and geographical contexts. It is the existing or potential application of ML that is the primary cross-cutting

theme. This section then provides a useful backdrop to the discussion of our research, which has endeavored to combine ML and econometrics to examine one specific historical setting, early modern England.

39.2.1 Addressing Challenges in Conventional Historical Research: Data Collection

The first challenge in quantitative historical research on institutions and culture is to create the datasets themselves. The creation of new datasets has dramatically increased at the same time as the application of ML, often using the pre-processing tools that have gained widespread use in the service of creating text-based-data for ML. For example, Pagé-Perron (2018) employs text processing and network analysis on a cuneiform corpus, providing insights into how ML converts unwieldy corpora into a more usable form that will facilitate future research into Mesopotamian culture and institutions. This is one aspect of ML that we emphasize in future sections. In old texts, tokenization, the separation of a document into lexical units, and lemmatization (removal of inflections to focus on the same basic words) will often require the creation of bespoke processing software. But after the first creation, methods and database will be available for future researchers. Many other studies have made available databases constructed using ML methods. Moreover, datasets are also being constructed for more traditional empirical exercises without the direct application of ML algorithms, but with the use of the types of computational resources for big data that have become available only in recent decades. Table 39.1, part A, lists examples of relevant contributions.

39.2.2 Addressing Challenges in Conventional Historical Research: Data Analysis

One common empirical challenge that can arise in historical research on institutions and culture involves addressing the curse of dimensionality. In pre-nineteenth-century historical settings, where cross-sectional samples tend to be small and the frequency of observations over time is low, the number of observations may barely exceed, or even fall short of, the number of variables. ML can then be productively employed in choosing predictors and allowing for flexible model specifications.

In standard settings, techniques such as ridge regression and lasso explicitly select a subset of pre-specified variables to avoid overfitting. For example, Michalopoulos and Xue (2021) use lasso to analyze age-old folklore concepts as predictors of present-day cultural values and beliefs as measured in surveys. Notably, the authors also discover that the set of models selected by ML might exhibit inherent instability across various specifications, particularly when predictors are highly correlated. Therefore, a supplementary human classification exercise was essential to increase precision of the results.

In settings where non-linearities are particularly important, decision-tree methods can be employed effectively to cast light on the determinants and importance of institutions and cultures. Such methods, often implemented in the form of random forests (ensembles of individual trees), select non-linearities and variable-interactions in an iterative fashion, with little prior model specification by the researcher. While random forests allow the researcher to ascertain the relative importance of the different variables in predicting an outcome, this information is usually uninformative of the direction of a variable's influence. Quite often, therefore, it is valuable to complement the random-forest results with structural regression

Table 39.1 Examples of articles using ML applied to institutions and culture in historical settings

Part 1: Addressing challenges in conventional historical research: Data collection
Barber et al. (2023): pre-enlightenment Europe, historical data on religiosity
Buckles et al. (2023): historical U.S. censuses, record matching
Clark (2023): England 1600 to present, data on lineage of English people
Dittmar and Seabold (2019): Europe 1456–1600, data on books
Hitchcock et al. (2023): English criminal trial records, 1674–1913
Pagé-Perron (2018): Mesopotamia, cuneiform corpus
Poulos (2019): nineteenth-century Georgia, record matching
Part 2: Addressing challenges in conventional historical research: Data analysis
Düben and Krause (2023): cities in imperial China, random forest
Ma and Li (2020): government bureaucracy in thirteenth century China, random forests
Michalopoulos and Xue (2021): folklore and present-day beliefs, lasso
Perrin (2022): demographic transition in eighteenth and nineteenth century France, clustering
Part 3: Charting new research avenues: Leveraging text as data
Almelhem et al. (2023): English print texts 1500–1900, topic modeling
Barron et al. (2018): French Revolution parliamentary assembly speeches, topic modeling
Bi and Traum (2019) U.S. newspapers 1840s, clustering and topic modeling
Blaydes et al. (2018): medieval texts, topic modeling
Funk and Mullen (2018): nineteenth-century US, similarity measures
Gavin (2016): early-modern English texts, networks
Gennaro and Ash (2022): U.S. Congress speeches, 1858 onwards
Guldi (2022): Britain, similarity measures
Law (2019): nineteenth and twentieth century national constitutions, topic modeling
McCannon and Porreca (2023): England, criminal trial records 1800's, topic modeling
Miller (2013): late imperial China administrative records, topic modeling
Newman and Block (2006): eighteenth-century US newspaper, topic modeling
Rice (2019): nineteenth and twentieth century US Supreme Court opinions, topic modeling
Rockmore et al. (2018): eighteenth to twentieth century national constitutions, topic modeling
Rule et al. (2015): US State of the Union addresses, miscellaneous
Van Vugt (2022): letters in the Republic of Letters, networks

Notes: The table provides a guide to a sampling of the scattered literature on use of ML applied to institutions and culture in historical settings. Under each part, entries are listed alphabetically. For each entry, we provide brief information on the examined historical context and illustrative use of ML

estimates, where the model facilitates easy interpretation of the sign of an effect. Table 39.1, part 3, offers examples of pertinent contributions.

39.2.3 Leveraging Text-as-Data

Several studies from widely different historical contexts use ML methods for analysis of text as data to illuminate previously unknown aspects of institutional and cultural evolution. For instance, what were the themes emphasized by newspapers in colonial and early independent America? In, to our knowledge, the first application in a historical-cultural setting of topic modeling, a technique discussed later in the paper, Newman and Block (2006) focuses on the *Pennsylvania Gazette*. Their findings indicate, for example, that the presence of ideas involving a national government experienced a surge from the 1760s to the 1790s. But perhaps even more importantly, the authors demonstrate the potential of ML to “reveal cultural. . . histories,” to enable an investigation devoid of “fallible human indexing or their own preconceived identification of topics,” and to utilize “orders-of-magnitude more documents than a person can reasonably view.”

Also drawing on information encapsulated in or pertaining to texts, studies of networks using ML methods have become common across a wide variety of fields. The idea is to take a set of interactions (articles and their citations are a modern example) and use the computational methods to find patterns in interactions when the core dataset might contain millions of potential interactions. For example, Van Vugt (2022) employs a network-based approach to investigate the exchange of letters between Antonio Magliabechi, an eminent Florentine librarian, and his fellow scholars. The analysis involves examining data on Magliabechi’s correspondents, the timing of the correspondence, and the connections of those correspondents beyond direct interactions with Magliabechi. The findings provide a nuanced depiction of the dynamic and evolving relationships among a group of interacting social agents, that is, on the spread of culture. Table 39.1, part C, lists examples of a number of recent contributions leveraging text as data.

39.3 The Corpora Utilized in Our Analysis

Our analysis utilizes two major text corpora: the English Reports (Renton, 1900–1932; hereafter ER) and the Early English Books Online-Text Creation Partnership (2022; hereafter EEBO-TCP). In the following, we describe each of the two corpora and the pre-processing challenges that they entailed.

39.3.1 The English Reports (ER)

The ER compiles the definitive set of 129,042 reports on decisions rendered in the superior English courts of law between the early thirteenth century and the

mid-nineteenth century. Coverage is sparse for the early years, but the number of reports becomes substantial by the mid-sixteenth century.

As is familiar in historical research, the ER constitutes neither the population nor a random sample of cases adjudicated in the English courts. Rather, the ER includes a subset of the cases heard by the superior courts. Reporters were especially eager to provide a record of cases that elucidated novel or unsettled aspects of law, that is, cases giving rise to legal development (Grajzl & Murrell, 2021b). However, there is no precise record detailing the criteria used to select cases for incorporation into the compilations of reports that were subsequently consolidated in the ER.

The ER became the *de facto* record of court cases that the English legal profession used as its authoritative source for legal precedent. Thus, the ER contains a record of nearly all cases that came to influence the law. In this sense, the ER provides unique insight into the nature of English caselaw development between the mid-sixteenth and late eighteenth centuries. There is no alternative machine-readable legal-historical corpus of comparable depth and breadth available for this era. Any exploration of English legal history has to use the cases in the ER.

39.3.2 The Early English Books Online-Text Creation Partnership (EEBO-TCP)

The EEBO-TCP is a machine-readable corpus of 60,331 texts, which captures core works representative of pre-1700 English print culture. (Gavin (2017) provides a history of the EEBO-TCP project, with interpretation.) The texts were prepared by the Text Creation Partnership (TCP) and included in Early English Books Online (EEBO). The starting point of the creation of the texts were catalogs that “trace the history of English thought from the first book printed in English in 1475 through to 1700” (TCP, 2022). The EEBO-TCP corpus comprises about one-half of the texts listed in a master catalog of works known to still exist.

Texts were included in the EEBO-TCP in an attempt to create a representative corpus. The underlying vision was “to key as many different works—as much different text—as possible” (TCP, 2022), that is, to build a corpus that approximated a random sample of available works. Yet, it would be inappropriate to consider the EEBO-TCP corpus as representing a random sample of English culture. The texts mirror the culture embedded in printed materials rather than representing broader popular culture. The texts included are those that have survived, thus reflecting the extent to which subsequent generations valued them. And the EEBO-TCP project prioritized first editions, making TCP oriented to the creation of culture rather than the consumption of culture.

Nevertheless, the EEBO-TCP corpus is unique as a resource for understanding pre-eighteenth-century English print culture. No alternative machine-readable corpus has been assembled that is both capable of supporting a quantitative exploration

of English culture before 1700 and has such a comprehensive range of texts. The EEBO-TCP is an invaluable collection for those interested in understanding English history from the advent of the printing press to the time when peaceful accession of new governments was assured. It is especially valuable for those who want to use computational methods.

39.3.3 Pre-Processing the Corpora

Computational analysis of historical texts requires much arduous, time-consuming work pre-processing texts. The challenges are numerous. We escaped the step of parsing the inscrutable historical fonts because we were able to access two corpora that had already been made machine-readable. But there are many steps between machine-readable and ML-friendly. English orthography was not standardized until at least the eighteenth century. A machine does not know that *thynkyng* and *thinking* are the same word. Modern natural language programs will not know archaic inflections, and therefore that *endeth* indicates the presence of the verb-stem *end*. And given that the learned in the seventeenth century and before were fond of using Latin, the computer will not understand, unless one instructs it, that *equus* can be rendered as *horse*. Nor will it know that *equo* is the dative of *equus*, a problem that we solved by programming a Latin stemmer, that is, an algorithm to convert words to their Latin roots.

Thus, the texts were replete with words that could not be found in extremely comprehensive dictionaries of modern English words. Translation was required for all these words. This entailed the construction of bespoke project-specific dictionaries to do translations by machine, since any process relying on humans would be infeasible without a large budget. At the same time, all production-inserted formatting symbols were removed, as well as non-recognizable symbols that had strayed into the files. Then we dropped documents that contained either an especially small number of words or an uncharacteristically high share of words that could not be matched to any word in the modern English dictionary even after the processing. Some early English legal cases were written in Law French, an idiosyncratic blend of French, Latin, and English. Unfortunately, we had to drop these cases, a lingering regret.

We then assigned to each document metadata variables such as the year of publication, the number of words, and, in the case of the ER, the adjudicating court and reporter name. The resultant corpora were additionally processed using the steps recommended for the variety of ML that we first used, a structural topic model (Roberts et al., 2014, 2016). We thereby converted all words to lower case and applied a standard stemming algorithm that converted all inflected words to their root form. Finally, we removed standard English stop words (e.g., *the*, *an*, etc.), numbers, words with fewer than three characters, words appearing in only one document, and punctuation.

39.4 Three Example Techniques

So, one has a data set that comprises a corpus that has been cleaned. What does one do with it? Here, we do not discuss the technicalities of the methods. Understanding the basics of what is being estimated does not require knowledge of the technicalities. Moreover, we only discuss the tools that we think will provide easy entry points for institutional economists whose primary goal is to learn about the past using tools readily accessible to the non-specialist. The selection of methods to discuss is based on our own experience.

39.4.1 Topic Modeling

In the ML literature on text-analysis, by far the most widely used tool is topic modeling. Topic modeling takes a set of documents and asks what subjects or ideas (i.e., topics) can be found in these documents. Like all data analysis it is about finding useful (aggregate, predictive, evocative, causal) patterns in the data, patterns that can be interpreted as reflecting a subset of the ideas that one could plausibly expect to find in the data.

Topic modeling is an unsupervised ML method, meaning that the dataset does not contain any labeled observations that are examples of the pattern that one is seeking. Thus, if one takes the works of Francis Bacon (English philosopher, statesman, and scientist) and Edward Coke (the most celebrated common-law lawyer), as we did in Grajzl and Murrell (2021a), one might expect that the corpus would contain ideas about epistemology and constitutional law. But these ideas appear in greater or lesser amounts in different documents, and in some documents they might be only fleetingly employed. Thus, there are no preexisting document labels, saying, for example, “this document is primarily about epistemology and not about constitutional law.” One simply has patterns in the correlations of word usage across documents. Topic modeling uses these patterns to identify ideas emphasized in the corpus: these are the topics (e.g., epistemology and constitutional law). Of course, we do not directly observe the ideas themselves and cannot test how well we have characterized the ideas.

Topic modeling then has two fundamental objectives. The more prosaic is simply a reduction of dimensionality in the data, for example, reducing to 105 topics the 142 million words in 67,455 reports on court cases heard in England’s high courts between 1765 and 1865 (Grajzl & Murrell, 2023c). In this sense, topic modeling can be viewed simply as an extension of existing latent variable techniques for numerical data, such as principal components. When pursuing dimensionality-reduction, the prime consideration is to finish up with a set of variables—often authors focus on only one—that can be used in some other statistical exercise, such as a differences-in-differences regression.

But, for us, the most exciting aspect of topic modeling follows from a second objective—using the results to interpret and use a large set of latent variables that summarize the texts completely. It is a way to uncover deeper ideas that are implicit in many documents, even ideas that are never the explicit focus of any specific

document. In Grajzl and Murrell (2019), where we focused on Bacon alone, the topic epistemology was, unsurprisingly, found to be one of the most important, but so was a less expected topic, religious law, reflecting the emphases of a man bathed in Puritanism from an early age and later to become Lord Chancellor. Even less expected, our later investigation of the outputs of the topic model showed that these two topics were deeply connected, suggesting the origins of Bacon's epistemology in his religious and legal background.

Topic-modeling algorithms can be motivated by a crude conceptualization of document generation that lends itself to formalization. The model presumes that the author of a specific document being examined begins with a fixed number of ideas, or topics, which are available for use when writing. Use of a particular topic implies that the author has greater affinity for the vocabulary more closely associated with that topic. Moreover, the author's emphasis will vary across documents. Then, even with each document viewed simply as a frequency distribution of word usage, semantic content is embedded in the corpus via correlations between word distributions across documents.

What in particular does a topic model estimate? The researcher specifies *ex ante* the number of topics (T), or sets of ideas, that are in the dataset, a corpus of documents. The algorithm estimates T separate probability distributions over vocabularies. Those distributions are the topics. This is intuitive. When authors write about transubstantiation, as in the corpus examined in Grajzl and Murrell (2023d), they will place a high probability on the use of words such as sacrament, body, and bread. In contrast, when they are producing their lusty entertainments they would tend to favor a different set of words, such as cuckold, fop, and whore. Thus, the words most used by a topic are one key to interpreting the ideas captured by the topic. A majority of papers use this information, and only this information, to give the topic a name.

We think it unfortunate that, in naming the topics, most papers rely only on the words most associated with each topic, without additionally considering which documents feature each topic most prominently. In our work, we found it equally important to read those top documents, sometimes 20, but more usually as many as 40 or more per each topic. This was crucial: topics can only be understood in relation to all other topics—nuances appear in the differences between closely related topics. This reading requires domain-specific knowledge: ML is never purely computational. For example, in Grajzl and Murrell (2021b), the topics Contract Interpretation & Validity and Identifying Contractual Breach had many characteristics in common and it was only by comparing the two that the underlying ideas behind each topic could be readily isolated. Our experience has been the opposite of Rule et al.'s (2015) who view estimated topics as being inherently hard to identify. Indeed, we have found that scrutiny of the estimated topics helps ease one into the study of a historical period whose language and context might, at first, feel very foreign.

By far the most useful output of the topic model is the document-topic matrix (DTM). With N documents in a corpus and T topics estimated, then the DTM is an $N \times T$ matrix where element ij indicates the proportion of document i that arises from the author's use of topic j . Connecting this matrix to data on the characteristics

of documents provides a wealth of opportunities to explore how ideas change over time, how they differ between authors, etc. For example, if one is interested in running a regression with documents as observations, then one might include just one column of the matrix as an explanatory variable. If one is interested in changing emphases over time and one knows the year of each document, then it is trivial to aggregate rows of this matrix to produce a year-by-topic matrix showing how the attention to each idea, or topic, changed over time.

Casting this abstract description into a concrete setting might be helpful to readers. In Grajzl and Murrell (2023d) we aim to use ML to further the understanding of English culture in the sixteenth and seventeenth centuries. After pre-processing, the final corpus comprised 57,863 documents and contained 83,337,912 words. After a series of exercises using both objective statistical measures and subjective evaluation of the topic-modeling output, we chose 110 topics. We then examined the words most highly associated with each topic and the documents prominently featuring the topic. If the topic model is being used to present an integrated picture of relevant history, then this very laborious step is necessary to confidently draw conclusions, or state new facts.

For example, the top word-stems for one topic emphasized logical connectives, such as yet, though, and thus. The top documents often contained the word “experiment” and most focused on religion. One top document stated that “If a man were but well read in the story and various passages of his life, he might be able to make an experimentall divinitie of his own. He that is observant of Gods former dealings and dispensations towards him, may be thence furnished with a rich treasury of experience. . .” We named the topic Baconian Theology, indicating theological ideas that used the epistemology that is characteristic of Bacon. Then, a timeline of the use of this topic was produced. This topic appeared in the corpus before the contributions of Bacon, suggesting that Bacon used what was present in English culture before he himself made his mark. This fundamental observation about the evolution and origins of ideas is readily, almost trivially, extracted from the relevant data, once the data have been generated.

Importantly, the topic-naming process can lead to new findings. Therefore, it is important not to place this naming exercise within the straitjacket of clearly identified categories already appearing in the literature. If such matches do occur, then it should be because the topic-modeling estimates dictated them. If they did not occur, then one enters discovery mode, which was exactly the situation arising in the detection of Baconian Theology, a specific set of ideas that none of the preexisting literature prepared us for. Therefore, it is important to identify the ideas underlying all estimated topics.¹ Thereby, one is forced to go beyond one’s presuppositions and identify the unexpected. Being able to name all topics with terms that are logically compelling is a verification of the quality of the topic modeling.

¹Sometimes topics are interpretable but say nothing about substance. For example, Grajzl and Murrell (2021b) found the topic non-translated Latin, arising because pre-processing worked imperfectly on the highly idiosyncratic Latin of sixteenth- and seventeenth-century lawyers.

Above we have suggested two different levels of analysis—use a very small number of topic-related variables in a standard econometric analysis or develop a broad picture using all topics, often 100 or more, to provide an overview of an era of institutional or cultural history. But there is an intermediate level: aggregating the topics (the columns of the DTM) into a smaller number of themes, maintaining a comprehensive overview but fewer variables, a number that might be more manageable in a traditional econometric analysis. Thus, Grajzl and Murrell (2023c) characterize caselaw development during 100 years of English industrialization (1765–1865) with 105 topics aggregated into 13 readily inferred broader themes. The topics are grouped into themes using researcher judgment. This fits with our general philosophy of using ML where necessary (creating a 67,455-by-105 DTM) and using judgment and existing knowledge where possible (aggregating 105 columns of that matrix into 13).

The above description reflects elements of the methodologies of a number of papers, covering several different subjects and historical eras. The methodologies had many common elements. The findings are disparate and cannot be combined into a simple story. But perhaps one element running through all the findings can be emphasized. We find that specific ideas appeared earlier than is generally assumed in the existing literature. Thus, our analysis of early English law suggests that major developments in financial markets were already ongoing in the first part of the seventeenth century (Grajzl & Murrell, 2021c, 2022b), earlier than mooted in most previous studies. In addition to the findings on Baconian Theology noted above, in Grajzl and Murrell (2023b) we locate the stirrings of the English scientific revolution in 1558 to 1610, earlier than is conventional (Wootton, 2015). Given the central place of the financial and scientific revolutions in any story of England's rise, these are fundamental findings.

We believe that there are two prime reasons for this earlier dating. The first is familiar in modern academia. The synthetic work that summarizes new developments in an organized and evocative way gets more attention than the often-inchoate publications that were the most innovative (McMahan & McFarland, 2021). Indeed, in our ongoing work, we show that this very process might account in no small part for an over-estimation of the novelty of the ideas of Edward Coke, often characterized as the greatest lawyer in the period 1550–1750. The second is due to the difference between ML and traditional text analysis. Fragments of ideas spread across many documents might be as important as the same ideas dominating a single work. The latter is much more likely to appear later, when ideas have coalesced, than the former, a result of incomplete ideas percolating. The latter will normally be dominant in traditional text analysis, but not in ML.

The above has only referred to the descriptive findings of our topic modeling analysis, findings that come almost transparently from the topic-model estimates. But much more can be done with the DTM, as we will discuss in Sect. 39.6 on the integration of topic modeling and standard econometrics. In the next two subsections, we turn to alternative ways to encapsulate and investigate large text datasets.

39.4.2 Measuring Novelty Using Words

Wootton (2015, p. 18) provides a capsule summary of English history that has a large effect on our work: “. . . let us take for a moment a typical well-educated European in 1600 – we will take someone from England, but it would make no significant difference if it were someone from any other European country as, in 1600, they all share the same intellectual culture. . . . But now let us jump far ahead. Let us take an educated Englishman a century and a quarter later. . . [Then] what was true of an educated Englishman in 1733 would not be true of a Frenchman, an Italian, a German or even a Dutchman.” Although science is the subject of Wootton’s book, this observation should not be restricted to science, but to all manner of English ideas. So something remarkable happened in England during 1550–1750, with many debates about what that was and when that was. Our topic modeling results, as we have mentioned, do reflect on that, but alternative methods are available.

One such method that we are pursuing in research that is still forthcoming is to implement the method of Kelly et al. (2021) on our legal and cultural databases. The essence of that method is as follows. Any text document can be characterized by a frequency distribution over the vocabulary it uses. Because of the computational demands resulting from the use of large matrices, pre-processing steps beyond those normally applied have to be used to reduce the number of items in the vocabulary, generating a computationally manageable representation of the corpus. Standard packages are available to do this. For the 1765–1865 ER corpus, for example, this distribution is summarized by a vector of length 10,791 for each document. Take any two documents and visualize their word distributions as vectors in n -dimensional space. Then, the cosine of the angle between the vectors is a measure of how similar the documents are. A document is more novel the smaller is the sum of its cosine-similarities with all the documents produced in previous years.

Thus, one can calculate a novelty score for every single document in the corpus. It is then trivial to calculate timelines of aggregate yearly novelty. And using data characterizing documents one can generate disaggregated timelines. For example, using the document-theme matrix from a previous topic modeling exercise on the development of caselaw, one can find which of procedure caselaw (subsuming topics such as procedural bills, correct pleas, and evidence gathering and admissibility) or real property caselaw (encompassing topics such as manorial tenures, common-land disputes, and uses) was most novel, and when. Our early results suggest the former after 1640 and the latter before 1640.

Or one can find out which particular courts were producing the most novel decisions and when, contrasting, perhaps, the contributions of equity (especially in Chancery) and common-law courts. (Equity was a distinct system of legal principles and remedies that supplemented the common law. Equity relied on laxer pleading rules, used an inquisitorial procedure without juries, and provided relief in the form of decrees and injunctions.) Our early results suggest the common-law was dominant in innovation until the 1660s but after that equity courts were hugely important.

Or one might be interested in, for example, which cultural outputs were emphasized to a greater or lesser degree at different periods of history. Take literature or institutions for example. Our early results suggest that, not surprisingly, 1640 was a great dividing line, with institutions emphasized much more thereafter, as the production of literature became relatively less important.

Both of the methods we have discussed above disregard word order when analyzing a text. But methods that leverage order, albeit in a limited way, also exist. The next subsection describes one that was first proposed in 2013.

39.4.3 Word Embeddings to Characterize Change in Ideas

Word embeddings focus on capturing the semantic relationships between individual words. Any specific word is characterized as a vector in a low-dimensional space (e.g., 300 dimensions has been a common choice). The position of the word in the vector space reflects the word's meaning. Estimates of the numerical vector for each word are obtained by leveraging the words that in a text appear close to the target word. Utilizing these word-embeddings estimates, one can gain insights into the usage of the words by examining the relationships between the vectors for the words. For example, if one found in a given corpus that the difference between the vectors for man and woman was very similar to the difference between the vectors for career and family, then this would be an indication of stereotyping of gender roles (see, e.g., Ash et al., 2024).

One can readily see how word-embeddings models can be productively employed to cast light on the historical evolution of institutions and culture. For instance, Gennaro and Ash (2022) investigate the text of speeches given in the U.S. Congress from 1858 onwards in order to characterize the relative emphasis of reason and emotionality in political discourse. The authors first estimate a word-embedding vector for each word in the corpus, and then, by averaging the vectors for each word in a document, obtain a vector for each document. At the same time, they use standard external dictionaries such as those included in the LIWC (2022) to define reason- and emotionality-related word lists. A vector for each word list can be obtained as the average of vectors for individual words on the list. Gennaro and Ash (2022) then characterize every speech along an emotion-reason continuum by computing measures of vector similarity between the document vector and vectors for the emotionality and reason word lists. An examination of the results indicates, for instance, that emotional expression in the U.S. Congress remained comparatively low and stable until the mid-twentieth century but experienced a notable increase from the late 1970s onwards.

In our ongoing work, we are using a similar approach to offer the very first quantitative assessment of one of the most famous hypotheses on socio-legal evolution—Henry Maine's 1861 dictum that the evolution of progressive societies follows a pattern "from Status to Contract." In premodern societies, legal rights

and obligations, according to Maine, were rooted in “status.” This status was defined by intricate relational and hierarchical networks within the family and broader societal structures. A person’s ascribed position in society determined the range of choices available to them in, for example, commerce and marriage. In contrast, in what Maine perceived as progressive societies, such as Victorian England, there emerged a shift toward relationships based on “contract.” In this context, the law explicitly acknowledges the capacity of individuals to freely assume powers, responsibilities, and authority.

To assess the validity of Maine’s conjecture, we are using a word-embeddings approach to generate document-level measures of the relative emphasis on status and contract concepts in early-modern English caselaw and culture. One could, of course, simply follow the approach by Gennaro and Ash (2022) and rely on predetermined external dictionaries to find the words related to these concepts. However, the use of word-embeddings ultimately requires that the dictionaries be specified sufficiently reliably, a task that is especially challenging in historical contexts such as ours.

For instance, in our setting, it would be difficult even for a well-trained modern-day historian to specify *ex ante* an encompassing set of terms that adequately reflect the notion of tenure, the essence of status-based relations in a feudal society. By inspecting the set of words that each have a word-embedding highly correlated with that for tenure, one generates a list of synonyms that are highly pertinent for the seventeenth century. This reveals the importance of, e.g., knight-service, heriot-service, and frankalmoign (tenure in free alms), which can be added to the word list of tenure-related terms. The procedure might also reveal words that must be dropped from an initial word list because their early modern usage was very different from usage today. Thus, our procedure uses the word embeddings not only to characterize concepts and documents, as in Gennaro and Ash (2022), but also to finetune the external dictionary word lists by identifying a whole set of relevant synonyms for any particular word. Importantly, the synonyms implied by word embeddings reflect word usage in exactly the historical period and context that one is studying.

The corresponding iterative, word-embeddings-aided approach to dictionary construction (see, e.g., Rice & Zorn, 2021) is therefore especially productive in the context of specialized corpora, such as legal documents, and generally for corpora from times when word usage was very different from our own. In our setting, the word embeddings effectively teach the researcher seventeenth-century English. For example, in early-modern caselaw, drawer is most closely associated with endorser and acceptor. In early-modern print culture, fellow is most similar to rascal and rogue.

Our preliminary results indicate that the sociolegal evolution of English society was not nearly as linear as Maine would have predicted. Moreover, there were considerable differences between caselaw and print culture in the temporal evolution of the emphasis on status versus contract.

39.5 A Necessary Digression: A Different Methodological Perspective?

The previous sections have focused primarily on the new methodologies. The following section will discuss our combining of these methodologies with standard econometrics and the substantive findings that resulted. But before we move to that discussion, it is necessary to consider the overarching methodological paradigm that has been the natural outcome of work.

The reader will notice that, so far, we have not talked about standard errors, hypothesis testing, and suchlike. We have been more concerned with how to expand the set of reliable and, hopefully, broadly accepted, historical facts. There are several reasons for this.

First, the processing of texts brings a whole new range of information into easy view. After all, where, for example, could one ever hope to obtain a timeline depicting the way in which lusty entertainments replaced historical romance within the production of English culture (Grajzl & Murrell, 2023d)? The questions of what and when come before the question of why.

Second, the hypothetic-deductive methodological straitjacket—theory, hypothesis, and one statistical test—fits awkwardly with an exercise in which data is being analyzed for the first time, where the most pressing task is to find out what could be in the data. At that stage, it is not clear what one should be theorizing upon. The isolation of interpretable facts from raw text data is itself a creative enterprise in which discovery leads to new ideas that in turn could be fashioned into a new theory. Induction seems much more apposite in providing a methodological underpinning to this exercise than the hypothetico-deductive method.

Third, the very methods of ML have not been shaped by those who prioritize the hypothetic-deductive methodology. Instead, the primary objectives for developers of ML tools have been prediction and summarization. Therefore, the set of easily implementable routines—for those of us who are not computer scientists, with limited computer power, and without an army of research assistants—does not usually focus on standard errors of output. Perhaps this will change, but we do not think that will be very soon. Instead, for the type of work that we do, the original output of an ML exercise—say, the document-topic matrix—is treated as if it were a set of solid facts. Notably, this is exactly the way national income numbers are used in an enormous number of econometric exercises.²

Fourth, the hypothetic-deductive methodology is impossible to implement when one is trying to make sense of new data types. We confronted this in our study of English law. A particularly clear example was the presence of the idea of *assumpsit* in English law, essentially the circumstances under which a debt could be assumed to arise after the non-fulfillment of a contractual promise. The usual narrative has

²And indeed, it is exactly in this way that researchers treat a variable they have derived from a ML process and fitted into a regression. There is little acknowledged recognition that a value of this variable would be best represented by a distribution, rather than a single number.

centered on a long struggle between courts, then a decisive case very early in the seventeenth century, and acceptance thereafter. But we found a timeline of *assumpsit*'s appearance in English law as an inverted-U with a peak at 1635, indicating a significant rise before 1600, followed by a fall to insignificance as the seventeenth century proceeded. Yet we knew that the idea of *assumpsit* was universally accepted as the seventeenth century drew to a close. We then had to build a model to understand why the topic-model output could look this way. The theory followed from insights provided by the data, rather than preceding the empirical analysis. This is the very essence of induction. Our model incorporated the evident notion that ideas are most discussed when they are in contention or being spread to new situations. When they are broadly accepted, they do not appear in discussion.

Thus, to us at the moment, the most productive use of ML in historical institutional and cultural research lies in adding to, or modifying, the set of facts that are readily accepted as characterizing a historical period. So just as historians are no longer prone to debating whether Henry VIII had five or seven wives, with the accepted answer even being the cornerstone of a Broadway musical, we would hope that new implementations of ML establish similar facts with regard to institutional and cultural history—facts that, eventually, do not need further questioning. For example, in our work, we show a change towards a less combative tone in culture during the sixteenth and seventeenth centuries: English religious and political discourse became less antagonistic and more scholarly in tone, while authority relationships became less important, literature more playful, and economic topics more prominent. Similarly, we establish that the chief legal ingredients of a financial revolution were present in caselaw well before the Glorious Revolution, the political event in 1688 that institutional scholars previously long perceived as critical to subsequent financial development.

Is this a radically new way of thinking about historical research? We think not. Rather, it just describes candidly the way things have always been done. We believe that much research that is framed in hypothetico-deductive terms is really inductivist in disguise. Consider digging only slightly below the surface of virtually any applied econometrics exercise demonstrating the presence of a hypothesized effect and produced by any researcher that you truly admire and whom you believe approaches research with the highest levels of integrity. You would invariably encounter a process in which the hypothesis has been shaped by perusals of the data, in which various approaches have been followed and discarded, and in which the results that have been chosen for presentation are those viewed most representative of the many produced. It is induction, or perhaps more properly abduction, under the cloak of hypothetico-deductive.

But ardent hypothetico-deductivists might object that their method is always about testing general rules, ones that will inform about something more than the events in the particular time and place characterized by the data. But this is not usually the sphere of historical research, where data are scarce and have to be constructed using myriad assumptions. Historical research is primarily about time and place where omitted variables and data-selection issues are legion. In that case,

the hypothetico-deductivist method in history boils down to isolating specific facts, rather than discovering general broadly applicable rules.

There is another sense in which we believe that ML historical research is very similar to the ways things have always been done. Humans and machines do the same thing: topic modeling. Informal topic-modeling exercises have been carried out for centuries by learned scholars reading many texts, finding commonalities between those texts, and interpreting the underlying ideas. That is after all how McCloskey (2016) discovered a Bourgeois Civilization in Western Europe; how Skinner (1965) found the idea of a Norman Non-Conquest in sixteenth- to eighteenth-century England; how Hirschman (1977) saw the Interests taming the Passions at the same time; and how Zweigert and Kötz (1992) were able to find empiricism in common law and abstraction in continental law.

We have stated these points bluntly, because we feel they need to be stated so. There seem to be vast terrains of economics where these points are not recognized, at least openly. But as ML has become more common in the social sciences there is more acknowledgment of these points. Kahneman (2019, p. 609) notes that when applying ML to large datasets, “you will find out much more than your theory is designed to explain,” which in turn allows for the possibility that “machine learning can be a source of hypotheses.” In political science, “. . . The introduction of machine learning methods also invites us to reevaluate the typical model of social science. . . the current abundance of data allows us to break free from the deductive mindset that was so previously necessitated by data scarcity” (Grimmer et al., 2021a, p. 396). In sociology, “Engagement with computational text analysis entails more than adapting new methods to social science research questions. It also requires social scientists to relax some of their own disciplinary biases, such as a preoccupation with causality. . .” (DiMaggio, 2015, p. 4). In the digital humanities “. . . the mathematical assumptions of machine learning—both unsupervised and supervised approaches—are. . . better equipped for use in the type of inductive, exploratory, and contextual research traditionally conducted using qualitative methods” (Nelson, 2021, p. 2). And even in economics, “In many applications of topic models the goal is to provide an intuitive description of text rather than inference. . . Interpretation or story building. . . tends to be a major focus for topic models and other unsupervised generative models” (Gentzkow et al., 2019, p. 556).

39.6 From Machine Learning to Econometrics

As we have commented above, ML analysis of texts reduces data dimensionality. For example, our application of topic modeling to a pre-1765 ER corpus summarizes 31 million words with a 52,949-by-100 document-topic matrix. The resulting estimates can then be productively employed in subsequent econometric analysis.

The existing literature using ML for text-as-data in historical-institutional contexts has, by and large, focused on extracting from the ML estimates a single variable to solve a specific data-econometric problem. For example, Bi and Traum (2019) explore how U.S. government bond prices in the 1840s were affected by newspaper

reporting. Upon applying clustering and topic modeling to the corpus of newspaper articles for each state, the authors construct a state-specific, time-varying index of state legislative activities and fiscal actions. The resulting variable is then used as a key determinant of bond prices. Similarly, McCannon and Porreca (2023) apply topic modeling to the Old Bailey records on criminal trials in nineteenth-century London to estimate the attention paid to a topic that is interpreted as reflecting the professionalization of the court. They then use the attention to the pertinent topic as an outcome variable in a difference-in-differences framework where the treatment of interest is a change in law that introduced the right to counsel.

The focus on a single variable, as in the above studies, is usually driven by the hypothetico-deductive method. It has a laser-like focus on constructing a new variable rather than leveraging the whole set of dimensionality-reduced data produced via ML. In these types of studies, the ML methods are handmaids to existing approaches: they do not form the bedrock upon which the projects are built.

In contrast, when the goal is to paint a broader empirical picture of the historical roots, flow, and repercussions of ideas about institutions and culture, the use of ML tools will, by necessity, form the very foundation on which a project is built. In our view, it is this conception that offers the opportunity for an especially productive combination of ML and statistical-econometric methods. In what follows, we offer examples from our own research.

39.6.1 Example 1: Uncovering the Legal Legacies of Early English Caselaw

The path-dependence of caselaw development is broadly acknowledged by legal historians. But which early-modern legal ideas were especially relevant for later caselaw development during the Industrial Revolution? We addressed this core question about the evolution of English law in Grajzl and Murrell (2022a). In conducting the analysis, our unit of observation was a pre-1765 document (a case). Our dependent variable was post-1764 citations to pre-1765 cases (Murrell, 2021). A citation reflects the precedential relevance of the cited case to the citing case, and thus citations to case reports are a quantitative measure of the influence of those case reports on the development of the law.

We used a standard econometric model for count data, explaining post-1764 citations to pre-1765 cases with the characteristics of the pre-1765 cases. Under the reasonable assumption that all aspects of a prior case could be an input into future deliberations, and thus decisions, the number of explanatory variables was determined by our ML exercise on the corpus of pre-1765 cases, as many as 100 given that exercise. We therefore encountered a methodological challenge for which the use of ML has been strongly advocated: perhaps we, too, should apply lasso to reduce the number of predictors (see, e.g., Michalopoulos & Xue, 2021). But, in fact, our number of observations (documents in a particular post-1764 period) was large enough that we could easily estimate all of the desired parameters. Using lasso would have unduly restricted the substantive scope of our analysis, thereby failing to

provide a comprehensive overview of the early-modern legal origins of later legal development.

In substantive terms, we found that the ideas (i.e., topics) having the strongest effects were precedent-based thought and legal reporting in the style of Edward Coke, one of the preeminent makers of English law. The primary legacy of early English caselaw, therefore, lay in bestowing modes of reasoning. The other side of the coin was that a readily identified subset of preexisting legal ideas played no discernible role in caselaw development during the Industrial Revolution. For example, questions of jurisdiction—which courts had which authority—exhibited the weakest effect of all the sets of legal ideas.

We then investigated why a subset of preexisting legal ideas played no discernible role in subsequent caselaw development during the Industrial Revolution. Is it that those legal ideas became so widely accepted within the legal profession by the late eighteenth century that, even though they were relevant, legal professionals using them in the later era no longer felt the need to cite specific cases (hypothesis A)? Or is it that those legal ideas were generally less applicable in the late eighteenth and nineteenth centuries (hypothesis B)? The essence of our approach in distinguishing between these two possibilities rests on using a measure of the degree to which the corresponding legal ideas had been settled in pre-1765 caselaw. If hypothesis A is true, we would expect this pre-1765 measure to be negatively related to whether the ideas exert a detectable effect post-1764. In contrast, if hypothesis B is true, we would not expect to observe a systematic relationship between this pre-1765 measure and an indicator for whether the applicable ideas exert a detectable effect post-1764. Our evidence is consistent with hypothesis B. Thus, the reason why certain preexisting legal ideas do not exert a detectable effect is that those ideas were generally no longer key to post-1764 legal disputes.

As we emphasize in Grajzl and Murrell (2022a), our estimates of the later effects of earlier legal ideas provide a picture of the relevance of early English law for caselaw development in the Industrial Revolution that has never been provided before in such a systematic quantitative, comprehensive fashion. Our approach to investigating legal development could be applied in many other contexts.

39.6.2 Example 2: Assessing the Effects of Caselaw on Economic Performance

England's Industrial Revolution was an epoch-defining event. But just how important was the law emanating from English courts in facilitating industrialization? We tackled this pivotal, and previously unanswered, question in Grajzl and Murrell (2023c). Methodologically, we combined ML with conventional time-series econometrics. In particular, given the inherently intertwined nature of legal and economic development, we used vector autoregression (VAR). In modern macroeconomics, VARs are used extensively to characterize the dynamics of multiple time series and to conduct inference in settings that are fraught with endogeneity but lack natural experiments of the kind usually exploited by empirical microeconomists. As such,

VARs have often been used in economic history to examine the interaction among socioeconomic variables.

We do not view the application of VAR as in any way superior to analyses employing natural experiments. Natural experiments provide incisive lessons when a suitable truly exogenous variable is available. But, as we argue below, in earlier historical periods, and particularly when examining broad institutional and cultural data, the required truly exogenous variable to proxy an appropriate shock to the system is usually unavailable. In such cases, VARs can still provide valuable insights.

In one such exercise, examining the ER cases from 1765 to 1865, we first estimated a 105-topic model. With our intended unit of analysis being England in a particular year, it would have not been feasible to employ as many as 105 variables when estimating a VAR. Therefore, we used the intermediate level of analysis discussed in Sect. 39.4. Using the estimated document-topic prevalence matrix and the year of each case, we aggregated the topics to produce annual time series of attention to 13 distinct themes. Examples of such themes are procedure and reasoning, contract, debt and finance, and inheritance. Thus, a 67,455-by-105 document-topic proportions matrix was replaced in the analysis by a 67,455-by-13 document-theme-proportions matrix.

We then aggregated the document-level theme-proportions at the yearly level, generating 13 caselaw time series of per-capita annual attention of the ER to the pertinent themes. Crucially, and as indicated in Sect. 39.4, a simple model of cultural diffusion allows one to conclude that the attention paid to any theme in a given year reflects the amount of change in adherence to the corresponding legal ideas in that year. Therefore, our time series also measure the intensity of development in the pertinent areas of law. Finally, we augmented the 13 caselaw series with a real per-capita GDP series.

We were focused on examining the interaction over time of our caselaw variables and real per-capita GDP. These variables coevolve, each potentially affecting each other, most likely with a lag, but possibly contemporaneously. Those conditioned by standard microeconometrics will immediately think of unaccounted-for factors. These would be of three broad types: proximate causes of per-capita GDP, secular trends, and exogenous shocks. Proximate causes, for example capital accumulation, can be omitted if one is clear that the analysis then focuses on caselaw as a fundamental legal-institutional driver of economic performance. Secular trends are included via a linear time trend. Idiosyncratic (exogenous) shocks, unrelated to both past developments in caselaw and secular trends, are subsumed in error terms. Given this view of the world, one has a structure that exactly corresponds to a standard VAR.

However, as is well known, extracting the maximum amount of information from such a structure requires incorporating information to aid identification. Our goal was to offer a comparative account of the importance of different caselaw domains—13 in total—for England's economic performance. There was no possibility of finding a source of exogenous variation in each of these areas of law. Ours is therefore a very different empirical context than those encountered in a subset of

recent studies that also employ historical time-series data but are interested in identifying the effects of one specific exogenous variable (e.g., Jordà et al., 2022; Palma, 2022). We thereby followed the standard approach employed by macro-economists, using a recursive identification scheme naturally implied by our legal-institutional setting. This identification method effectively implies making assumptions on which variables cannot affect which other variables simultaneously.

Our estimates reveal that caselaw developments were a crucial determinant of economic fluctuations. Caselaw shocks together accounted for more of the variability in per-capita GDP around its long-term trend than did shocks directly to per-capita GDP. The direction of the response of per-capita GDP to caselaw innovations, however, critically depended on the legal domain. Developments in caselaw on intellectual property, organizations, debt and finance, and inheritance boosted economic performance. For instance, during this era, court decisions were critical in stimulating patenting, facilitating incorporation, resolving insolvency, and disambiguating wills, all developments that would have stimulated the supply and reallocation of capital. In contrast, developments in property and ecclesiastical caselaw exerted a negative effect on per-capita GDP. For example, eighteenth-century developments in rules on tithes, which remained in the domain of ecclesiastical courts, disincentivized capital accumulation.

Interestingly, our analysis identified an era when the legal system evidenced reduced attention to output-fostering areas of law and increased attention to output-hindering. We named this the “bleak-law era,” depicted in Dickens’ *Bleak House*. This was also the time when Bentham was making his thoroughgoing critiques of English legal processes.

Similarly, in Grajzl and Murrell (2023a) we investigated if caselaw developments impacted English economic development in the preindustrial era. Our dataset used variables standard in the Malthusian framework (real per-capita income and vital rates) alongside three yearly time-series of legal themes, that is, aggregates of legal topics that were estimated using the pre-1765 ER corpus. These three naturally fit into the Malthusian context: caselaw on land, inheritance, and families. Using a similar methodology to that in Grajzl and Murrell (2023c) described above, we estimated a VAR on the resulting dataset. The results show that preindustrial economic development was profoundly shaped by caselaw developments, with caselaw on families and inheritance being especially impactful. In the preindustrial era, developments in inheritance and family caselaw, for example, facilitated greater access to capital and enhanced financial security for wives, widows, younger sons, and daughters, reduced transaction costs within the family, and facilitated productive matching in the marriage market.

39.6.3 Example 3: Estimating the Coevolution of Ideas

Our exploration of the ER and EEBO-TCP corpora using the combination of ML and econometrics has allowed us to generate a rich set of entirely novel insights about the historical evolution of institutions and cultures. Among these, we have especially

focused on eliciting patterns in the coevolution of ideas. In the legal context, for example, scholars have long been pondering the question of the degree of autonomy of legal processes (Tomlins, 2007). Similarly, cultural historians have been debating the influence of religion on scientific thought (see, e.g., Merton, 1938; Wootton, 2015). Our application of topic modeling and time-series methods casts systematic quantitative light on these issues in the context of early-modern England.

In Grajzl and Murrell (2022c), we explored the coevolution of legal ideas on property, contract, and procedure, the three core elements of English caselaw. Once more we used topic-modeling on the pre-industrial part of the ER corpus (Grajzl & Murrell, 2021b, c) to construct yearly measures of attention to the three legal domains. Our VAR approach then directly implies two distinct conceptualizations of legal autonomy: vis-à-vis the rest of the legal system and vis-à-vis societal factors fully external to the legal system. The main insight generated by our empirical analysis is that the internal dynamics of the legal system explain much less of the development of caselaw on contract than that on property and procedure. Contract caselaw was more buffeted by unusual events occurring outside the normal functioning of the system of litigation (e.g., civil conflict, passage of important legislation, ascent of an especially innovative judge). Thus, in comparison with property and procedure, contract development was relatively more autonomous from the internal dynamics of the legal system, but relatively less autonomous from society.

In Grajzl and Murrell (2023d) we turned our attention to culture. Upon estimating a 110-topic model using the TCP corpus, we used multivariate time-series methods analogous to those in Grajzl and Murrell (2022c, 2023a, c), casting light on the coevolution of ideas within three broad themes: religion, science, and institutions. The question of how developments in these three areas affected each other is a staple of historical inquiry. Our analysis reveals, for example, that innovations in religious ideas induced strong responses in both science and institutions, especially at times when Puritanism was prominent in religious discourse. Thus, the religion-to-science link was strong already in the second half of the sixteenth century, much earlier than stressed by previous scholars (Merton, 1938, pp. 414–416; Mokyr, 2016). Again, as emphasized in Sect. 39.4, our analysis seems to identify significant events taking place earlier than the prevailing consensus in the literature.

Finally, two episodes that have received great attention from historians and social scientists interested in institutional change, the Civil War and the Glorious Revolution, did not lead to an unusually large amount of attention in print culture to matters pertaining to institutional development. Rather, cultural debates about institutions preceded, and thus perhaps helped spur, each of those revolutions.

39.6.4 Example 4: Finding Critical Junctures

As we have commented in Sect. 39.4, something remarkable happened in England during 1550–1750, and one of the enduring puzzles of history is what that was and when it happened. Grajzl and Murrell (2023b) address this puzzle directly. In the age of two major political revolutions, this paper looks for quiet revolutions—periods of

fundamental institutional and cultural shifts that do not occur via force nor have clear beginning and ending dates. To identify quiet revolutions, we apply the econometrics of unknown structural breaks to each of the time-series of 100 caselaw topics and 110 culture topics discussed above. The insight that drove this particular choice of econometric tool is that attention to a topic rises steeply (i.e., features a structural break) at the beginning of a conceptual revolution, as proponents of new ideas preach their gospels to the unconverted. But one such upturn does not identify a quiet revolution; rather, a quiet revolution is characterized by a cluster of upturns for similar topics in a short time period.

The results can be conveyed in two evocative figures, showing when England experienced change and in which areas of ideas. The visual exposition of the results is a crucial part of the analysis; without it, the reader would be left at the half-way stage of verbal summary of 210 facts isolated from each other. The results show that early-modern England featured several profound conceptual revolutions. The financial revolution began by 1660. Significant changes in land law began during the Interregnum. Caselaw relating to families began to change significantly in the late seventeenth century. Elizabethan times saw dramatic change in the emphasis on basic skills. At the same time, but perhaps not coincidentally, a Puritan revolution affected both theology and political ideas. An upturn in dissent preceded the mid-seventeenth-century revolution and Civil War.

39.7 Conclusion

The machine learning of history is not normal science in the sense articulated by Kuhn. When implementing the core methods, the beginning researcher cannot fall back on decades of scholarship, which usually guides decisions on the choice of many assumptions in contemporaneous research. There are no standard tests produced by tried and true methods, which usually dictate the acceptability of a piece of research in the relevant scientific community. The machine learning of history will therefore involve invoking new types of assumptions or combining old ones in unusual recombinations. Inevitably, these new ways of approaching research will collide with the normal-science paradigm expected of contributors to a particular discipline. We ourselves have received bemusing criticisms of the type of research that we are currently undertaking, some with phrasing that seemed to indicate we had breached some unwritten code of conduct. Our answers to these critics appear in the previous pages, an integrated body of findings, developed using rigorous, transparent, and replicable methods, adding fundamental insights about English legal and cultural history.

When powerful new tools drive radical changes in methods, the research community enters the early stages of a multi-strategy, multi-period, coordination game, moving hesitantly toward an equilibrium set of procedures that are widely spread and broadly understood. Eventually, a workable equilibrium will be found, establishing a whole new set of standard assumptions and practices that will guide

a large part of the research. The equilibrium is Kuhn's normal science. But costs are inevitable before that coordination equilibrium is found.

The process of reaching an equilibrium of a coordination game involves competitive elements: researchers have an incentive to develop approaches that they hope will 1 day become standard. In our reading of the literature, we have found that this seems to provide too many incentives to do things differently from other researchers. When there is no set of broadly accepted norms it is too tempting to construct ad hoc embellishments of techniques. These make it harder to interpret results across studies on similar subjects. These ad hoc elements of new studies also make it more difficult for a beginning researcher to understand which are the most productive ways of proceeding in a new research effort.

These observations have guided the way that we have pursued the research summarized in the previous pages. We are not interested in technique, *per se*, but have been motivated by two simple facts: texts provide a large proportion of historical data; and computational power and algorithmic development has proceeded far enough to make the processing and summarization of those texts possible for all researchers. There is now no need for the researcher interested primarily in the substance of history to engage in ad hoc development of new methods. This straightforward approach is exemplified in our own work on novelty, where we rely on standard measures of the similarity of texts.

However, the text processing and summarization does present distinctive challenges. A beginning one is that the production of text data suitable for machine learning will rest on bespoke databases that depend highly on the historical context. In our case, this involved understanding the nuances of seventeenth-century English. Nevertheless, as we showed in discussing our use of word embeddings, machine learning itself can provide tools to refine the accuracy of those databases.

Next is the problem of the sheer number of variables produced by the summarization of a large corpus of texts. After all, a useful cataloging of the fundamental ideas in English seventeenth-century law, for example, can be expected to comprise many dozens of variables. This, above all, suggests that descriptive analyses will provide a core output of any machine-learning exercise. Those analyses will require a thorough understanding of context, as in our discussion of developments in English law and culture. Moreover, in order to evocatively convey the lessons of, perhaps, hundreds of stark facts, the researcher will have to develop new types of figures (as we did in studying quiet revolutions) or find ways of conveying the results of many estimated parameters (as we did in examining the legal legacies of early caselaw).

The fact that standard econometric methods are most useful when the focus is on a limited set of variables presents another problem. This means that the researcher will usually need to further aggregate the results of the summarization (as when we produced themes from topics for both our legal and cultural databases). But then a further challenge often arises: the text summarization might produce a set of variables that are all endogenous, with no obvious exogenous explanatory variables to aid identification. This will require the use of techniques such as VAR together with an approach to identification that relies on broad theorizing rather than the adding of

data that provides exogenous information (as when we explored the role of law in the Malthusian framework or studied the interaction of religion and science). But this use of macro empirical tools that does not rely on natural experiments leveraging exogenous variation can fit uncomfortably with the perspectives of those scholars who are only willing to embrace the prevailing treatment-effect paradigm that nowadays dominates microeconomic empirical research.

In conclusion, machine learning applied to historical texts provides a whole new analytical toolkit facilitating insight into the historical roots of institutions (e.g., caselaw) and culture (e.g., theology), the study of which has previously relied predominantly on the reading of a limited number of texts. As we have made clear, the potential is enormous. But the beginning researcher considering using these tools needs to reflect on the fact that they will be working outside the framework of Kuhn's normal science. Then, pre-conceived ideas about what constitutes publishable research can affect evaluations in ways that are beyond the examination of rigor and quality. We have provided one methodological model for such research by describing the processes and products of our own work on English history. It is work that combines new machine learning methods with econometric techniques that have already been tried and tested, with the primary objective being the production of new substantive insights. We believe that following this model offers the opportunity to expand knowledge dramatically in a broad swathe of economic, legal, cultural, and intellectual history.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Almelhem, A., Iyigun, M., Kennedy, A., & Rubin, J. (2023). *Enlightenment ideals and belief in science in the run-up to the industrial revolution: A textual analysis* (IZA Discussion Paper No. 16674). IZA.
- Ash, E., Chen, D. L., & Ornaghi, A. (2024). Gender attitudes in the judiciary: Evidence from US circuit courts. *American Economic Journal: Applied Economics*, 16(1), 314–350.
- Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685–725.
- Barber, L., Jetter, M., & Krieger, T. (2023). *Foreshadowing Mars: Religiosity and pre-enlightenment warfare* (CESifo Working Paper No. 10806). CESifo.
- Barron, A. T. J., Huang, J., Spang, R. L., & DeDeo, S. (2018). Individuals, institutions, and innovation in the debates of the French revolution. *Proceedings of the National Academy of Sciences*, 115(18), 4607–4612.
- Bi, H., & Traum, N. (2019). *Sovereign risk and fiscal information: A look at the U.S. state default of the 1840s* (Federal Reserve Bank of Kansas City Working Paper No. 19-04). Federal Reserve Bank.
- Blaydes, L., Grimmer, J., & McQueen, A. (2018). Mirrors for princes and sultans: Advice on the art of governance in the medieval Christian and Islamic worlds. *Journal of Politics*, 80(4), 1150–1167.

- Buckles, K., Haws, A., Price, J., & Wilbert, H. E. B. (2023). *Breakthroughs in historical record linking using genealogy data: The census tree project* (NBER Working Paper No. 31671). NBER.
- Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov.
- Clark, G. (2023). The inheritance of social status: England, 1600 to 2022. *Proceedings of the National Academy of Sciences*, 120(27), e2300926120.
- DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society*, 2(2), 1–5.
- Dittmar, J., & Seabold, S. (2019). *New media and competition: Printing and Europe's transformation after Gutenberg* (CEP Discussion Paper No 1600). CEP.
- Düben, C., & Krause, M. (2023). The Emperor's geography – City locations, nature, and institutional optimisation. *Economic Journal*, 133(651), 1067–1105.
- Funk, K., & Mullen, L. A. (2018). The spine of American Law: Digital text analysis and U.S. legal practice. *The American Historical Review*, 123(1), 132–164.
- Gavin, M. (2016). Historical text networks: The sociology of early English criticism. *Eighteenth-Century Studies*, 50(1), 53–80.
- Gavin, M. (2017). How to think about EEBO. *Textual Cultures*, 11(1/2), 70–105.
- Gennaro, G., & Ash, E. (2022). Emotion and reason in political language. *Economic Journal*, 132(643), 1037–1059.
- Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3), 535–574.
- Grajzl, P., & Murrell, P. (2019). Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas. *Journal of Comparative Economics*, 47(1), 111–135.
- Grajzl, P., & Murrell, P. (2021a). Characterizing a legal-intellectual culture: Bacon, coke, and seventeenth-century England. *Cliometrica*, 15(1), 43–88.
- Grajzl, P., & Murrell, P. (2021b). A machine-learning history of English Caselaw and legal ideas prior to the industrial revolution I: Generating and interpreting the estimates. *Journal of Institutional Economics*, 17(1), 1–19.
- Grajzl, P., & Murrell, P. (2021c). A machine-learning history of English Caselaw and legal ideas prior to the industrial revolution II: Applications. *Journal of Institutional Economics*, 17(2), 201–216.
- Grajzl, P., & Murrell, P. (2022a). Lasting legal legacies: Early English legal ideas and later Caselaw development during the industrial revolution. *Review of Law & Economics*, 18(1), 85–141.
- Grajzl, P., & Murrell, P. (2022b). Using topic-modeling in legal history, with an application to pre-industrial English Caselaw on finance. *Law and History Review*, 40(2), 189–228.
- Grajzl, P., & Murrell, P. (2022c). A macrohistory of legal evolution and coevolution: Property, procedure, and contract in early-modern English Caselaw. *International Review of Law and Economics*, 73, 106113.
- Grajzl, P., & Murrell, P. (2023a). Of families and inheritance: Law and development in England before the industrial revolution. *Cliometrica*, 17(3), 387–432.
- Grajzl, P., & Murrell, P. (2023b). Quiet revolutions in early-modern England. *Public Choice*, 200, 357–381.
- Grajzl, P., & Murrell, P. (2023c). Caselaw and England's economic performance during the industrial revolution: Data and evidence. *Journal of Comparative Economics*, 52, 145–165.
- Grajzl, P., & Murrell, P. (2023d). *A macroscope of English print culture, 1530–1700, applied to the coevolution of ideas on religion, science, and institutions* (SSRN Working Paper). SSRN.
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021a). Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24, 395–419.
- Grimmer, J., Roberts, M. E., & Stewart, B. M. (2021b). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.
- Guldi, J. (2022). The algorithm: Mapping long-term trends and short-term change at multiple scales of time get access arrow. *The American Historical Review*, 127(2), 895–911.

- Hirschman, A. O. (1977). *The passions and the interests: Political arguments for capitalism before its triumph*. Princeton University Press.
- Hitchcock, T., Shoemaker, R., Emsley, C., Howard, S., McLaughlin, J., et al. (2023). The Old Bailey proceedings online, 1674–1913. <http://www.oldbaileyonline.org>, version 9.0, Autumn 2023.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). *An introduction to statistical learning with applications in R* (2nd ed.). Springer.
- Jordà, Ò., Singh, S. R., & Taylor, A. M. (2022). Longer-run economic consequences of pandemics. *Review of Economics and Statistics*, 104(1), 166–175.
- Kahneman, D. (2019). Comment on artificial intelligence and behavioral economics. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *The economics of artificial intelligence: An agenda* (pp. 608–610). The University of Chicago Press.
- Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3), 303–320.
- Law, D. S. (2019). Constitutional dialects: The language of transnational legal orders. In G. Shaffer, T. Ginsburg, & T. C. Halliday (Eds.), *Constitution-making and transnational legal order*. Cambridge University Press.
- LIWC. (2022). <https://www.liwc.app/>
- Ma, L., & Li, M. (2020). *What helped officials of song dynasty in climbing the greasy pole: An empirical study* (SSRN Working Paper). SSRN.
- McCannon, B. C., & Porreca, Z. (2023). *The right to counsel: Criminal prosecution in 19th century London* (SSRN Working Paper). SSRN.
- McCloskey, D. N. (2016). *Bourgeois equality: How ideas, not capital or institutions, enriched the world*. University of Chicago Press.
- McMahan, P., & McFarland, D. A. (2021). Creative destruction: The structural consequences of scientific curation. *American Sociological Review*, 86(2), 341–376.
- Merton, R. K. (1938). Science, technology and society in seventeenth century England. *Osiris*, 4, 360–632.
- Michalopoulos, S., & Xue, M. M. (2021). Folklore. *Quarterly Journal of Economics*, 136(4), 1993–2046.
- Miller, I. M. (2013). Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach. *Poetics*, 41(6), 626–649.
- Mokyr, J. (2016). *A culture of growth: The origins of the modern economy*. Princeton University Press.
- Murrell, P. (2021). Did the Independence of judges reduce legal development in England, 1600–1800? *Journal of Law and Economics*, 64(3), 539–565.
- Nelson, L. K. (2021). Leveraging the alignment between machine learning and intersectionality: Using word Embeddings to measure intersectional experiences of the nineteenth century U.S. South. *Poetics*, 88, 101539.
- Newman, D. J., & Block, S. (2006). Probabilistic topic decomposition of an eighteenth-century American newspaper. *Journal of the American Society for Information Science and Technology*, 57(6), 753–767.
- Pagé-Perron, É. (2018). Network analysis for reproducible research on large administrative cuneiform corpora. In B. Juloux, A. R. G. Vanessa, & A. di Ludovico (Eds.), *CyberResearch on the ancient near east and neighboring regions: Case studies on archaeological data, objects, texts, and digital archiving* (pp. 194–223). Brill.
- Palma, N. (2022). The real effects of monetary expansions: Evidence from a large-scale historical experiment. *Review of Economic Studies*, 89(3), 1593–1627.
- Perrin, F. (2022). On the origins of the demographic transition: Rethinking the European marriage pattern. *Cliometrica*, 16(3), 431–475.
- Poulos, J. (2019). Land lotteries, long-term wealth, and political selection. *Public Choice*, 178(1), 217–230.

- Prüfer, J., & Prüfer, P. (2020). Data science for entrepreneurship research: Studying demand dynamics for entrepreneurial skills in The Netherlands. *Small Business Economics*, 55(3), 651–672.
- Renton, A. W. (1900–1932). *The English reports. Great Britain. Parliament. House of lords.* W. Green & Sons.
- Rice, D. (2019). Measuring the issue content of supreme court opinions. *Journal of Law and Courts*, 7(1), 107–127.
- Rice, D. R., & Zorn, C. (2021). Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Political Science Research and Methods*, 9(1), 20–35.
- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural topic models for open ended survey responses. *American Journal of Political Science*, 58(4), 1064–1082.
- Roberts, M. E., Stewart, B. M., & Airolidi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515), 988–1003.
- Rockmore, D. N., Fang, C., Foti, N. J., Ginsburg, T., & Krakauer, D. C. (2018). The cultural evolution of national constitutions. *Journal of the Association for Information Science and Technology*, 69(3), 483–494.
- Rule, A., Cointet, J.-P., & Bearman, P. S. (2015). Lexical shifts, substantive changes, and continuity in state of the union discourse, 1790–2014. *Proceedings of the National Academy of Sciences*, 112(35), 10837–10844.
- Skinner, Q. (1965). History and ideology in the English revolution. *The Historical Journal*, 8(2), 151–178.
- The Text Creation Partnership (TCP). (2022). <https://textcreationpartnership.org>
- Tomlins, C. (2007). How autonomous is law? *Annual Review of Law and Social Science*, 3, 45–68.
- van Vugt, I. (2022). Networking in the republic of letters: Magliabechi and the Dutch Republic. *Journal of Interdisciplinary History*, 53(1), 117–141.
- Wootton, D. (2015). *The invention of science: A new history of the scientific revolution.* Harper. Kindle Edition.
- Zweigert, K., & Kötz, H. (1992). *Introduction to comparative law* (Second revised edition, T. Weir. Trans.). Clarendon Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



New Developments in Case Study Methods for Institutional Economics 40

Andrew Bennett 

Contents

40.1	Introduction	1060
40.2	Comparative Advantages of Case Studies in Research on Institutions	1061
40.3	New Developments in Case Study Methods: Typological Theorizing	1065
40.4	A Typological Theory Example: <i>Pawned States</i>	1070
40.5	New Developments in Case Study Methods: Formal Bayesian Process Tracing ...	1074
40.6	Example: Formal Bayesian Analysis of Evidence in <i>Pawned States</i>	1079
40.7	Conclusions	1084
	References	1085

Abstract

The present chapter reviews the advantages of case studies in the study of economic institutions and outlines two new case study methods: typological theorizing and formal Bayesian process tracing. Case studies have many comparative advantages relative to quantitative and experimental methods, including strengths in analyzing path dependency, equifinality, high-order interactions, anomalies, substantively important cases, complex concepts that are hard to measure, and causal mechanisms and processes. These advantages are a good fit for institutional economists' interest in big picture questions, critical junctures, and complex concepts. The chapter introduces and illustrates typological theorizing and formal Bayesian process tracing using the example of Didac Queralt's *Pawned States* (2022), a study of state building and international finance in the global south in the 1900s. The chapter concludes that wider and more rigorous use of case study methods can contribute greatly to new institutional economics and the field of economics more generally.

A. Bennett (✉)
Georgetown University, Washington, DC, USA
e-mail: bennetta@georgetown.edu

Keywords

Case study · Methodology · Process tracing · Typology · Bayes · Bayesian · Institutional economics

40.1 Introduction

Case study methods have gained renewed respect in the social sciences as methodologists have clarified their comparative advantages relative to quantitative and experimental methods. These comparative advantages include the study of path dependency, equifinality, high-order interactions, anomalies, substantively important cases, complex concepts that are hard to measure, and causal mechanisms and processes. Case studies can also help address potential endogeneity, and contrary to the views of some critics, case studies can at times lead to generalizations about large populations of cases. While quantitative methods have some advantages in assessing causal effects at the population level, and experiments can be strong at causal identification, case studies can also at times allow strong causal identification in individual cases and contribute to the study of causal effects.

These comparative advantages make case studies particularly useful in the study of institutions, and case study research is common in institutional economics. Yet case studies can also contribute to other areas of economics as well, and economists studying institutions can benefit from new developments in case study methods. The present chapter reviews the comparative advantages of case study methods and their relevance for the study of institutions. It then outlines two new developments in case study methods: typological theorizing and formal Bayesian process tracing. It illustrates these methods using the example of Didac Queralt's *Pawned States: State Building in the Era of International Finance* (Queralt, 2022), which argues that easy access to international finance in the 1800s inhibited state building in the global south. *Pawned States* provides a particularly good example of how typological theorizing and formal Bayesian process tracing could have made an award-winning study even clearer and stronger.¹ The chapter concludes that wider and more rigorous use of case study methods can contribute greatly to new institutional economics and the field of economics more generally.

¹*Pawned States* has won four prizes: the 2023 Georgetown University Mortara Center Leppgold Book Prize for the best book on International Relations (for which I was an award committee member), the 2023 William H. Riker Book Award (ex aequo) of the APSA *Political Economy Section*, the 2023 Best Book Award of the APSA *International Collaboration Section*, and the 2023 Gaddis Smith International Book Prize for best first book by a Yale ladder faculty member.

40.2 Comparative Advantages of Case Studies in Research on Institutions

While research on institutions uses many quantitative and experimental methods (Ménard & Shirley, 2022, p. 15), case studies are more common in the study of institutions than in other subfields in economics. Beckmann and Padmanabhan (2009, p. 344) list more than a dozen examples, including famous case studies by North and Thomas (1973) on institutional changes in European countries; Ronald Coase's study of the British Lighthouse system (1974); and Elinor Ostrom's studies of common pool resource management (1990). These scholars intuitively understood that case studies were suited to the research questions they raised and the inferential challenges they faced: big picture questions, critical junctures, path dependency, potential endogeneity, high-order interaction effects, and complex concepts that are hard to measure quantitatively (Ménard & Shirley, 2022).² Institutionalists are also interested in anomalous cases that do not fit experts' theoretical expectations, and as these are by definition few in number, case studies are the best method for examining them. At the most general level, case studies are stronger than quantitative methods in examining causal mechanisms (George & Bennett, 2005). While quantitative methods are strong at assessing average causal effects in populations, and experimental methods are good at causal identification, case studies also have some capacity for exploring causal effects, generalizing beyond the cases studied, and achieving causal identification. I consider below each of these comparative advantages of case studies, and their relevance for the study of institutions.

Institutional economics often focuses on fundamental legal institutions; critical junctures in the evolution of these institutions; and substantively important cases (Ménard & Shirley, 2022, p. 131). This often translates into an interest in specific individual cases, but frequentist statistical analysis is unable to assess single cases (except when these "cases" are not actually single units but include many data points across space, time, or substantive issue areas). Even if there are a few dozen cases of particular types of critical junctures, frequentist statistical analyses are not powerful. In contrast, because they use Bayesian rather than frequentist logic, case studies can use process tracing, a form of within-case analysis detailed below in Sect. 40.4, to make strong inferences on the causes of outcomes in single cases if the evidence is more likely under some explanations than under others. Because case studies are the only observational method that can examine one or a few cases, this has sometimes created a misimpression that case studies are useful only when a small number of cases of a phenomenon exist. In fact, due to their ability to assess causal mechanisms in greater detail, case studies are useful even when many cases make statistical analysis possible. Many multimethod studies identify correlations in a population

²On the comparative advantages of case studies relative to quantitative methods, see Brady & Collier, 2010, George & Bennett, 2005, pp. 3–64, and Mahoney & Goertz, 2012. For a similar view of the comparative advantages of case studies in the study of economic institutions, see Alston, 2008).

and then use case studies to assess whether any hypothesized causal mechanisms are actually evident in individual cases.

Frequentist statistical studies also often exhibit limited interest in the most substantively important cases. Studies of voting in the U.S. Congress, for example, often use datasets where important votes are lumped together with trivial ones. Statistical studies also tend to give little attention to cases that are “influential” in the statistical sense, or the cases whose removal from a dataset would have big effects on the parameter estimates of a model. At times statistical claims depend on a relatively small number of influential cases, but as statistical methods cannot investigate individual cases, statistical studies often overlook these cases, even though measurement errors or omitted variables in these cases could profoundly affect the interpretation of statistical results.³

Similarly, case study researchers are often interested in studying “deviant” cases, or cases that do not fit theoretical expectations or that have a large error term in statistical models. Such cases can be a fruitful venue for discovering variables omitted from a model. Statistical researchers, on the other hand, are often uncurious about deviant cases, as they expect such cases will occasionally arise due to measurement error or the cumulative small effects of the many weak variables left out of a model. Of course a large error term could arise for many reasons—one or a few important omitted variables, a large number of weak omitted variables, or measurement error—but it is not possible to assess which of these is true without studying these cases.

The emphasis that institutional economics places on substantively important cases and critical junctures extends to an interest in the study of path dependency (Ménard & Shirley, 2022, pp. 133–135). In its most general form path dependency is a claim that “history matters,” or that “current and future states, actions, or decisions depend on the path of previous states, actions, or decisions” (Page, 2006, p. 88). Within this general framework, many different causal mechanisms can lead to path dependency. Scott Page identifies four path-dependent processes: increasing returns (when the more a choice is made, the greater its benefits), self-reinforcement (often through creation of complementary institutions), positive feedbacks (arising from positive externalities), and lock-in (when a choice becomes better than the alternatives once sufficiently numerous or powerful actors make that choice; Page, 2006, p. 88; on increasing returns, see also Pierson, 2000, 2004). James Mahoney adds that there can also be negative externalities and decreasing returns, which can lead to either cycling between two equilibria or sequentially determined but non-repeating events (Mahoney, (2000) terms the latter “reactive sequences”).

It is possible to study diverse mechanisms and kinds of path dependency with statistical methods when enough cases exist, or to model them formally with game

³ A related issue is the “ecological inference problem:” statistical correlations at the population level do not necessarily hold true for or explain any particular individual case in the population. Among other possible reasons for this (selection effects, measurement error, etc.) the outcome of any particular case could be driven by the values of variables omitted from a model.

theory or agent-based modeling (for example, see Antonelli, 1997). Because critical junctures on fundamental institutions are rare, however, and also because case study methods have comparative advantages in studying causal mechanisms, nearly all of the macrohistorical research on path dependence in political science, economics, and sociology uses case study methods (including the examples cited by Page, Mahoney, and Pierson). Many studies of smaller-scale institutional path dependence use case studies as well, for similar reasons: case studies can give careful attention to the sequences of events in individual cases, they can study rare events, they can include inductive searches for the variables that influence contingent choices at critical junctures, and they can examine interaction effects within one or a few cases (Bennett & Elman, 2006).

Another comparative advantage of case studies is that they can help address potential endogeneity, which is a common challenge to research on institutions (Ménard & Shirley, 2008, p. 15).⁴ Reciprocal feedback and causal direction can be difficult to sort out through statistical analysis, as when A and B are correlated this does not tell us whether A causes B, B causes A, or both (or neither, if some omitted variable is driving both A and B so the A-B correlation is spurious). Statistical analysts sometimes use instrumental variables to address potential endogeneity, but it is famously difficult to find instruments that influence the independent variable of interest, have no independent effect on the dependent variable, and are not themselves correlated with the error term. Case studies, in contrast, can address potential endogeneity by focusing in detail on sequences of events, examining questions like “who said or did what to whom when and with what information and expectations?” Put another way, case studies treat time differently from statistical studies—they use the sequencing of multiple events *within individual cases* to provide leverage for causal inference on the outcomes of those cases. This is quite different from statistical studies that use time series cross-sectional analysis, hazard models, or other measures of time as a variable for populations of cases.⁵

Case studies have advantages in exploring high-order interaction effects as well. High-order interaction effects arise when the processes and effects of an independent variable depend on the values of several other independent variables, which is a common challenge in the study of institutions (Ménard & Shirley, 2008, p. 2; Ostrom, 2008, p. 825). Statistical methods can model only 2-variable interaction effects, and they can only address the possible interactions that a researcher has identified in advance, as calculating all of the possible interactions among even a small number of independent variables essentially creates a large number of models

⁴In statistics, endogeneity arises when an independent variable is correlated with the error term of a statistical model, and it can arise from several causes: omitted variables, measurement error, selection effects, and simultaneity or feedback loops between independent and dependent variables (Hill et al., 2021). For present purposes, I focus only on the issue of simultaneity, which is how the term “endogeneity” is most frequently used in political science.

⁵As Beckmann and Padmanabhan note, “The detailed analysis of processes of institutional choice and change that may encompass short as well as long-term causes and effects may only be possible within a case study design” (Beckmann & Padmanabhan, 2009, pp. 356–7).

and makes the interpretation of the statistical results problematic. Interaction effects pose challenges for case studies as well, but case study researchers can identify high-order interactions inductively and they can model them with typological theories, discussed below, which treat cases as particular *configurations* of interacting variables (Ragin, 1987; Ostrom, 2008, pp. 825–826). Studying individual cases allows access to the diverse and often hard-to-quantify evidence of interaction effects, which can be uncovered inductively in documents and interviews. This is no guarantee that a researcher will identify and understand any or all of the relevant interaction effects—they may still not recognize them in the available evidence—but it provides an inductive avenue to the discovery of high-order interaction effects that is missing in statistical analysis.

A related notion that is central to case study methods and neglected in statistical analyses is the concept of equifinality, or the existence of multiple paths to the same outcome. There are many ways to become rich, for example: success in business, winning the lottery, corruption, marrying into a wealthy family, etc. For institutionalists, there can be different paths into, and out of, democracy, capitalism, empire, constitutionalism, state-building, and so on. Equifinality can be relatively simple, when only a few paths to an outcome exist and they have much in common, or complex, when there are many paths that have little or nothing in common. Equifinality implicitly lurks in the background of statistical analyses—behind statistical estimates of the “average effects” of variables there can in fact be many different combinations that produce the outcome of interest—and it receives very little attention in these analyses. Case study researchers have foregrounded the issue, often starting with a default assumption that there are different paths or configurations that can produce an outcome, and then charting out these different paths case by case in a “building block approach” (George & Bennett, 2005; see also the discussion of typological theorizing below).

Case studies are also suited to working with the kind of complex and difficult-to-quantify concepts that are common in institutional economics. Case study researchers measure concepts in depth in a few cases, rather than trying to code them quantitatively across a large number of cases. They also frequently use typologies to capture multi-dimensional concepts. The complex concepts in institutional economics that have been the subject of case studies include rules, norms, practices, electoral laws, property rights, transaction costs, asset specificity, organizations, and the rights and enforcement of contracts (Ménard & Shirley, 2008, pp. 14–15). More recently, institutionalists have focused on such complex concepts as mental models, ideas, and cognitive processes (Ménard & Shirley, 2008, p. 1).

Three common critiques—that case studies are weak at developing generalizable knowledge, achieving causal identification, and examining causal effects—are overstated at best. Sometimes case studies develop highly generalizable theories and sometimes they do not. Whether a case study’s findings are generalizable depends on the mechanisms it uncovers, as the nature of those mechanisms informs whether, how, and to what conditions or contexts they will generalize. Some mechanisms uncovered by case studies are very generalizable—Ostrom’s, 1990 book has over 54,000 citations on Google Scholar because her findings on the dynamics of

common pool resources are broadly applicable (Ostrom, 1990). At other times, case studies may uncover mechanisms that operate only in very narrow and unusual contexts (on generalizing from cases, see Bennett, 2022).

Causal identification from case studies is also contingent, depending both on the research design and the probative power of the evidence gathered. Strong instances of most-similar case comparisons, where only the value of one independent variable and that of the dependent variable differ between two cases, are conducive to causal identification. Pure examples of most-similar cases are rare, as usually several potential causal factors, in addition to the independent variable of interest, differ between cases.⁶ Even when most-similar comparisons are imperfect, however, process tracing evidence can help rule out residual differences between the cases as possible causes of their differing outcomes. Process tracing can also assess whether the independent variable that was present in one case but not the other might account for their differing outcomes. Even in single cases, using the Bayesian logic of process tracing outlined below, evidence that is highly likely under one explanation and unlikely under the alternatives can establish causal identification with high confidence.

Finally, while case studies can get close to the action of causal mechanisms, this does not mean they have no capacity for examining causal effects. They just go about this task very differently from quantitative studies. Rather than seeking average causal effects for a population, a concept that case study researchers often find problematic due to equifinality or causal heterogeneity, case study researchers are often interested in why the same change in a variable can have very different effects in different contexts (a phenomenon known as multifinality). A one percent change in interest rates by a central bank, for example, can have very different effects on inflation in different countries, with different recent trajectories, leaders, mixes of different sectors, and so on. Case studies can assess these effects and the paths that led to them.

40.3 New Developments in Case Study Methods: Typological Theorizing

A recent innovation in case study methods relevant to the study of institutions is typological theorizing, which can address high-order interaction effects, equifinality, and path dependence.⁷ Typological theories treat cases as configurations, or

⁶The strongest candidates for most-similar case comparisons are often comparisons of the same unit before and after a rapid and discontinuous change in one of the independent variables of interest, as this controls for many potential differences between units. This is why institutional economists often compare the same institution before and after some kind of exogenous shock.

⁷One form of interactions effect that typological theories can model consists of combinations of variables that jointly involve elements of necessity and sufficiency. For example, a variable X can be “INUS”: Insufficient by itself to cause an outcome Y, but a Necessary part of a *combination* of variables XABC that is itself Unnecessary but Sufficient for Y. In other words, ABC does not cause

combinations of variables, also termed “types.” When portrayed in tabular form, typically the rows of typological theories are the combinations of independent variables that constitute the types, which are instantiated in particular cases, and the columns are the independent variables; the dependent variable is then the final column at the right side of the table, as in the example below.

A typological theory can be defined as a set of hypotheses that delineates all of the logically possible combinations of the independent variables and predicts and explains the expected outcome for each combination. This is an idealized definition, as we often do not have a strong theoretical prior for every possible combination of the independent variables, nor even confidence that we have fully identified the complete list of independent variables relevant to the outcome of interest. Yet often we do have or can develop strong theoretical expectations for at least some of the possible combinations of the independent variables we have identified, which allows developing partial typological theories. Note that typological theories are different from taxonomies: taxonomies classify entities into categories that may or may not relate to causal phenomena, whereas typological theories explain and predict outcomes via implicit or explicit reference to causal mechanisms.

Typological theories are already common in institutional economics, but economists typically do not use the term, nor do they fully take advantage of the methodological potential of typological theorizing. Examples of typologies in the previous edition of the Handbook include those on modes of governance (Williamson, 2008, p. 49), constitutional regime types (Carey, 2008, p. 93), features of agricultural contracts (Allen & Lueck, 2008, p. 467), responses to regulations (Benham, p. 592), levels of institutions (Benham, 2008, p. 601, reproducing a table from Williamson, 1999; Ostrom, 2008, p. 843 has a similar typology), laws governing the franchise and extent of voting (Engerman & Sokoloff, 2008, pp. 657–658), and supply and demand for public order contract enforcing institutions (Greif, 2008, p. 764). Some of these examples are taxonomical, and some, like that by Greif, are the kind of typological theory that I focus upon herein (for numerous examples of typologies in the social sciences, see Collier et al., 2012). The section below outlines six steps for building typological theories and discusses how such theories can help researchers decide which cases to select from a population for close study via case comparisons and process tracing, illustrating the approach with the example of Queralt’s study of international finance and state formation.

The first step is to specify the dependent variable of interest and the possible outcomes on that variable. The dependent variable can be measured qualitatively or quantitatively, and it can be a nominal or ordinal variable, or even a typological concept itself, such as types of states or types of other institutions. Whether they use quantitative or qualitative concepts and measures of variables, or some combination

Y in the absence of X, every time there is XABC this causes Y, but there are other combinations such as DEF that can also cause Y, so the combination XABC is not necessary for Y. There are many possible kinds of conjunctive combinations related to necessity and sufficiency (Goertz & Mahoney, 2005), and typological theories can model any of them.

of these, typological theories typically use dichotomous or trichotomous variables to keep the number of possible combinations from getting too large to understand or convey.⁸ The number of possible combinations increases exponentially with each independent variable and each additional categorical measurement of that variable: five dichotomous independent variables have $2 \times 2 \times 2 \times 2 \times 2$ or 32 possible combinations of variables, for example, while five trichotomous independent variables have 243 possible combinations.

The next step is to identify the individual theories that relate to the dependent variable, to specify their associated independent variables and the nominal or ordinal values they might assume, and to create a typological theory table. A typological table includes a column for each of the independent variables, a column at the right side of the table for the dependent variable, and one row for each unique combination of the values of the independent variables. One way to identify the individual theories relevant to the phenomenon is to consider both agent-based and structural theories, and to consider within each of these categories theories that focus on ideas/identities/social relations, material or institutional balances of power, and transaction costs. This encompasses twelve kinds of theories that are common in the social sciences.⁹ This is the step at which the typological space can become unduly complex, with many rows of combinations or types; ways to simplify the typological space are identified below.

The third step is to theorize the expected outcome for each combination or row, and to write that expected outcome into the dependent variable column at the right side of the table. For some combinations or rows the expected outcome may be obvious or over-determined, with all of the independent variable values pointing together toward a particular value on the dependent variable. For other combinations the theoretically expected outcome is not obvious, and it does not depend on the number of variables that by themselves might push toward a positive versus a negative outcome. Rather, the researcher has to think through how the functional

⁸When variables range along a quantitative spectrum, researchers can use either theoretical criteria or empirical observation to decide where to draw the partitions among categories like high, medium, and low. Also, the simplification of continuous variables is meant to facilitate building a typological table that is simple enough to enable thinking about possible interactions among the variables, but the case studies can measure quantitative variables more continuously.

⁹More precisely, one can consider theories on agent→structure, agent→agent, structure→agent, and structure→structure mechanisms in the three categories of identity, power, and transactions costs, which yields twelve families of theories. This covers a wide range of social science theories—for example, the study of organizations as both enablers and constraints on the choices of agents is often divided into organizational cultures (identity), bureaucratic politics (power), and transactions costs. Researchers seldom find a need to include theories from all twelve categories to study the phenomenon of interest, but considering each of the twelve categories for candidate theories helps prevent omission of viable potential explanations. On this taxonomy of theories and its usefulness in typological theorizing see Bennett, 2013; a later elaboration adds intra-agent theories such as motivational, cognitive, and personality models of behavior (Bennett & Mishkin, 2023). Beckmann and Padmanabhan (2009, p. 351) provide an analogous taxonomy of theories, focusing on the effects, causes, and processes of social embeddedness, institutional environments, governance, and resource allocation.

interactions of the independent variables, at the values they occupy in the row or combination in question, would work together to produce one kind of outcome or another. If the answer for a row is still highly uncertain, at this step the expected dependent variable for that row can be left blank or filled with a question mark.

Up to this point the construction of the typological table relies wholly on prior theories and the decisions the researcher makes in modeling their interactions and expected outcomes. The fourth step starts to integrate this initial theoretical framework with the researcher's preliminary knowledge of the values of the independent variables in each of the cases that might be chosen for full case studies. This proceeds through placing each case in the row that fits the values of the case's independent variables.¹⁰ Cases in the same row are "typologically similar," that is, they have the same categorical values on their independent variables. Typically at this point the cases start to cluster together, with one or several rows having multiple cases, while others have a few cases and some have no cases. The classification of a case might change quickly at this stage with some initial research to establish the values of the independent variables. If the population of interest is small, the researcher might code all the cases in it and place them in their respective rows, but it is not necessary to code every case if the population is large. In that instance, the researcher might assess for each row whether there is at least one case in the population that fits, and if so, choose one or more representative cases to illustrate that row, as well as coding the substantively important cases and the cases that stand out due to particularly high or low values on one or more independent variables. It is also important to try to include one or several cases that represent each of the possible outcomes on the dependent variable.

The fifth step is to compare the theoretically expected outcomes from step three to the actual outcomes for the cases placed in each row. There are two kinds of anomalies of interest at this point: cases that did not have the expected outcome when the researcher had a strong theoretical prior for the row in question, and cases in the same row that have different outcomes whether or not the researcher had a strong theoretical prior for that row. Thinking through and doing some quick research on these anomalies and their possible causes can lead the researcher to revise a concept or its associated measurement, either generally or for a particular case. It could also lead the researcher to add an independent variable that resolves an anomaly. If no re-conceptualization, re-measurement, or additional variable comes to mind that might resolve the anomaly, the researcher might choose the anomalous case for a full case study that uses process tracing. Another task for this fifth step is to consider whether the empty rows—those lacking a case known to the researcher—are indeed empirically empty, or might include an overlooked case. If no such case

¹⁰Note that "cases" can be different from "units"; the same country can be separate cases in separate time periods when one or more of the independent variables changes in its value over time, which requires placing them in different rows. For example, the institutional features relating to democracy or its absence might lead to classifying Tsarist Russia up to 1917 in one row, communist Russia 1917–1990 in another row, Yeltsin era Russia 1991–1999 in a third row, and Putin's authoritarian era from 1999 onward in a fourth row.

can be identified, the researcher can record whether they think the row is empty for a reason—namely, that the particular combination of variable values that it represents is unlikely because of the causal relationships among the independent variables—or if it is a theoretically possible kind of case that to the researcher's knowledge has simply not yet arisen.

The sixth step, and the final one prior to carrying out process tracing on a select set of cases, is to work through any changes to the typological table introduced in the fifth step, and to re-work some of the earlier steps as needed. Introducing an additional independent variable to explain an anomalous outcome, for example, might offer a potential resolution to that anomaly, but it can create new anomalies in other rows once the new variable is coded for the other cases. Once the researcher is satisfied that they have integrated their theoretical knowledge and their preliminary empirical knowledge of the cases as best they can without considerably more research, the final task at his sixth step is to use the typological table to help select cases for full case studies, process tracing, and case comparisons. The typological table at this stage readily identifies the cases and comparisons that fit different research designs. The anomalous (or deviant, or outlier) cases are now clearly evident, as noted in step five, and by definition these are cases from which the researcher is likely to learn something. The cases that could be used in most-similar or least-similar case comparisons are obvious from identifying, respectively, rows that differ in one independent variable and in the value on the dependent variable, and rows that have only one independent variable in common and the same value on the dependent variable. The researcher can also choose to pursue a building block approach, selecting for study one case from each of several rows, perhaps including cases that represent each of the possible outcomes on the dependent variable, to learn about different paths to each outcome. Choosing two cases in the same row with the same outcome can strengthen the understanding of that path. It is possible to pursue more than one of these research designs in a single study, and individual case studies can do double-duty if they fit more than one design or more than one comparison. Pragmatic criteria for case selection can be employed as well, such as a focus on substantively important cases, under-studied cases, cases for which evidence is newly available, cases for which the researcher has the necessary language or other skills, cases resembling current policy challenges, or cases for which field work is not too expensive or dangerous.

As noted above, typological tables can become very complex, but they can be simplified. One approach is to have columns only for the component theories, rather than for each of the variables in these theories, and to create rows based on whether each theory considered by itself pushes toward a particular value on the dependent variable. Then the individual variables of each theory, and the possible values that they can assume and the ways that each combination pushes toward an outcome, can be backgrounded as separate typological theories. Elman (2005) suggests additional ways to simplify typological spaces, including re-scaling the variables at higher levels of aggregation, setting aside the rows that have no extant cases, and collapsing contiguous cells if there are no relevant differences between them. One might also set aside over-determined cases unless they have surprising outcomes or allow most- or least-similar comparisons. If the typological space is still too complex after such simplifications, the researcher can narrow their focus to one or a few sub-types, or rows, of greatest theoretical or policy relevance, leaving the others for future study.

40.4 A Typological Theory Example: *Pawned States*

Didac Queralt's *Pawned States* provides an excellent illustration of typological theorizing. Queralt addresses the puzzle of why European states pursued paths of development that resulted by the 1800s in strong states, or states with effective tax collection bureaucracies capable of raising funds for major expenditures such as wars, while most of the countries in the global south have taken much longer to develop such states and in some cases still lack strong fiscal capabilities. While colonialism is of course a strong explanation for underdevelopment in the global south, Queralt is interested in the role that external financing played in intensifying the effects of colonial repression (Queralt, 2022, p. 21). He also considers factor endowments/resource curses and social/ethnic divisions.

Queralt's theory is that the way in which a state finances fiscal shocks powerfully influences that state's development (Queralt focuses in particular on the fiscal shocks created by wars). There is path dependency here: European states, which developed before 1815, had no good alternative to building strong state bureaucracies that could extract domestic resources to pay for wars, despite the unpopularity of taxing citizens. Capital surpluses and international financing for the most part were extremely limited prior to the end of the Napoleonic Wars in 1815, and domestic actors had limited willingness to loan money to their governments without credible state commitments to repayment from future tax revenues. So European states developed strong institutions for taxing citizens to finance wars, and in exchange yielded to citizens elements of representative governance.

After 1815, first British and later French and German investors loaned their growing capital to countries in the global south in the form of sovereign loans with surprisingly low interest rates (Queralt, 2022, pp. 9–10). Many states chose to finance their wars through foreign borrowing, as this generated little domestic opposition and did not require sharing power with citizens or building effective institutions for tax collection. A few countries did choose, at least in some time periods, to build a strong tax apparatus to pay for wars, and these countries made progress in state building in these periods. Most countries in the global south, however, chose to rely on foreign borrowing to pay for their wars. The few that managed to either repay or settle their debts by building tax collection capacities were able to build strong states, but many were forced by the terms of their loans to allow foreign creditors to take control of their sources of revenue, such as tariffs, when they defaulted on their loans, as many did. These countries failed to build strong state capacities.¹¹

¹¹For the period after WWII, which is not the focus of Queralt's book, he argues that "extreme conditionality," or the direct foreign seizure and control of domestic revenue sources to repay debts, disappeared for three reasons: states were no longer in imperial competition and thus were not automatically willing to intervene to enforce extreme conditionality, private lending was largely replaced by official government lending, and new institutions arose for resolving and enforcing private international debts so harsh receiverships were not necessary (Queralt, 2022, pp. 296–297).

Queralt summarizes his theory in a flow diagram (Queralt, 2022, p. 14), but as is typical, this diagram can be translated into a typological table, whereby the different paths through the flow diagram become rows in the typological table. The result is the typological table below. With one trichotomous independent variable (states choose domestic taxation, domestic debt, or external debt to address the fiscal shock of a war) and two dichotomous independent variables (the choice between post-war taxation and debt repayment versus default on and settlement of foreign debts, and the choice between a post-default domestic tax or a debt-equity swap and receivership) there are $3 \times 2 \times 2$ or twelve rows, or combinations of independent variables, with a strong, clear prediction on the dependent variable for each row.

In this tabular form of Queralt's theory, the coding entries in bold indicate choices that were rendered unnecessary or impossible by earlier choices—for example, the choice to fund a war through taxes makes it unnecessary to fund it through later foreign borrowing. As Queralt's focus is on the global south after 1815 rather than on the early modern European monarchies (EMEM), specific country examples of the EMEM countries are not included in the table (Table 40.1).

Queralt's flow chart is arguably a better way to represent his theory, as states proceed sequentially through the choices in the columns from left to right. The advantage of rendering his theory as a typological table, however, is that it clarifies his case selection and case comparisons, and it draws attention to a type of case that is theoretically possible but is not among the cases he chose to study.

First, Queralt chose to study and contrast the cases of Japan (Type 9 in the table) and Argentina (Type 12), which highlights the role of domestic credit markets. Japan in the Meiji period beginning in 1868 had a strong domestic credit market, and the government developed a professional centralized bureaucracy that collected a variety of taxes. While Japan did borrow foreign funds to pay for military modernization, it used domestic tax revenues from its rapidly growing economy to repay its foreign creditors without ever defaulting, thereby maintaining its strong state and its independence from foreign rule (Queralt, 2022, pp. 251–258).

In contrast, Argentina had a robust economy by the late 1800s, but following a series of civil and interstate wars, it became addicted to rapidly growing foreign loans in the 1880s. By 1890 it was a net capital exporter as it sought to repay these loans. Weak domestic credit markets and a lack of fiscal discipline led to a doubling of Argentina's foreign debt between 1920 and 1930, and Argentina's lack of a strong domestic tax bureaucracy contributed to a chronic cycle of debt restructuring for decades (Queralt, 2022, pp. 258–262).

Queralt also chose to carry out before/after case studies of Siam and Ethiopia, which offer most similar case comparisons of Type 9 (the earlier period in both cases) versus Type 12.¹² In both Siam and Ethiopia, fiscal discipline and reliance on domestic capital in the earlier period was followed by heavy reliance on foreign

¹²It is not immediately obvious that these are most-similar comparisons, since Type 9 and Type 12 differ on more than one independent variable coding. The key difference, however, is that in the

Table 40.1 Queralt's theory on fiscal shocks and state building rendered as a typological theory table

Type number, country category, and/or Country examples	Financing before or during War: Domestic Tax (Dom. Tax), Domestic Debt (Dom. Debt), or External Debt (Ext. Debt)	Immediate post-war financing: War Tax (War Tax) or Default and Settlement of Foreign Debts (Default-Settle)	Post-War financing after default: Tax (Default-Tax) or Debt-Equity Swap and Receivership (Debt Eq. Swap)	Predicted outcome: State building (State Bldg) or Weak State (Weak State)
Type 1 Early Modern European Monarchies (EMEM)	Dom. Tax	War Tax	Default-Tax	State Bldg, due to Dom. Tax
Type 2 EMEM	Dom. Tax	War Tax	Debt Eq. Swap	State Bldg, due to Dom. Tax alone
Type 3 EMEM	Dom. Tax	Default-Settle	Default-Tax	State Bldg, due to Dom. Tax alone
Type 4 EMEM	Dom. Tax	Default-Settle	Debt Eq. Swap	State Bldg, due to Dom. Tax alone
Type 5 EMEM	Dom. Debt	War Tax	Default-Tax	State Bldg, due to War Tax
Type 6 EMEM	Dom. Debt	War Tax	Debt Eq. Swap	State Bldg, due to War Tax
Type 7	Dom. Debt	Default-Settle	Default-Tax	Not a possible combination
Type 8	Dom. Debt	Default-Settle	Debt Eq. Swap	Not a possible combination
Type 9 Japan Siam before 1932 Ethiopia until WWII Chile 1879–1883	Ext. Debt	War Tax	Default-Tax	State Bldg, due to War Tax
Type 10	Ext. Debt	War Tax	Debt Eq. Swap	Not a possible combination
Type 11	Ext. Debt	Default-Settle	Default-Tax	State Bldg, due to Default-Tax
Type 12 Most countries in the nineteenth century outside of Europe, including: Siam after 1932 Ethiopia after WWII Late Ching China Ottoman Empire	Ext. Debt	Default-Settle	Debt Eq. Swap	Weak State, due to Debt Eq. Swap

loans in a later period, with a corresponding slow down or reversal in building a professional tax apparatus. In Siam, leaders avoided relying on foreign credit, which they viewed as a form of colonialism, up until 1932, and used domestic funds to achieve considerable state building by 1914. After a revolution in 1932, however, reliance on external credit led to stagnation in state building in Siam (known as Thailand after 1939; Queralt, 2022, pp. 262–269).

In the case of Ethiopia, after first accepting a loan from Italy in 1889 with extreme conditionality in the event of default, Emperor Menilek II paid the loan back within 4 years to preserve Ethiopia's independence (Queralt, 2022, pp. 269–273). He relied instead on domestic taxation, and quickly succeeded in building a state powerful enough to defeat Italy in the Italian-Ethiopian war of 1895–1896. Following World War II, however, Haile Selassie played upon Cold War rivalries to extract massive aid and loans from the United States, and Ethiopia's state building suffered as its professional bureaucracy devolved into a patronage system. Subsequent governments continued to suffer from weak capacity for tax collection and to rely heavily on foreign funds.

As a final case study, Queralt chose the case of Chile during the 1879–1883 War of the Pacific, a Type 9 case. In this case, Chile's default on its debts just as the war started and its lack of a strong reliable revenue stream meant Chile could not restructure its existing foreign debts or take on new foreign loans, so Chile had to rely on state building to generate domestic revenues to pay for the war.

Thus, while the Type 9 cases were fairly rare and the Type 12 cases were very common, Queralt wisely over-sampled the Type 9 cases to highlight exceptions to the general trend, and the Type 12 cases he chose for process tracing provided strong most-similar comparisons to his Type 9 case studies. Queralt also chose not to study Types 1–6 as these are the already-well-studied cases of early modern European monarchs, and as Types 7, 8, and 10 are not possible combinations they are of course not included as case studies.

The typological table also draws attention on Type 11, a type of case that is theoretically possible but not among the cases Queralt chose to study.¹³ If Type 11 did not occur in the time period under study, then the Table and its associated theory can be simplified. The whole fourth column labeled “Tax after Default (TD) or Debt-Equity Swap (DES) and Receivership” could be dropped from the Table, as there would be no variance among the ED→DS cases and all the other rows would have an impossible entry in that column.¹⁴ Indeed, the Table could then be further simplified by collapsing Types 1–4 into a single type, combining Types

earlier period both countries instituted taxes to pay foreign debts, and in the later period, they did not, and their resulting defaults led fairly directly, perhaps even unavoidably, to debt-equity swaps.

¹³(In personal correspondence, Queralt noted to me that “This is the path followed in modern-day default settlements, in which the IMF bails out the country that defaulted in return of fiscal austerity plus debt renegotiation (e.g., Greece 2007–8).” He references it briefly in his book on pp. 159 and 296–7 (Queralt, 2022).

¹⁴Note that if the fourth column is dropped, then it is more obvious that the Type 9 and Type 12 cases can be compared as most-similar cases, as per the previous footnote. This is especially true for before-

5–6 into another single type, and dropping the impossible types 7, 8, and 10 as well as 11, leaving only four types or paths that are actually possible and clarifying that Queralt studied the only two non-European combinations that were possible.

40.5 New Developments in Case Study Methods: Formal Bayesian Process Tracing

Process tracing is a within-case method of analysis that uses evidence from a single case to adjudicate among alternative explanations for the outcome of that case. Much like detective work and medical diagnosis, process tracing involves both “suspects” (alternative possible causes) and “clues” (evidence from the case and its context that the investigator intentionally looks for or stumbles across). The main advantage of process tracing is that it gets close to the action of causal mechanisms, using the sequencing of events, as well as the events themselves, as a source of causal inference. This can help untangle potential endogeneity, assess whether conjunctions of events are spurious rather than causal, inductively identify omitted variables and interaction effects, and strengthen comparisons between cases. Process tracing has costs and limitations as well: it can require lots of information and time, extant theories are not always clear on their observable implications in particular contexts, and the results may or may not generalize. Moreover, the conclusions are only as strong as the evidence allows—if the evidence is equally likely under the alternative explanations, then examining it does not clarify our understanding of what caused the outcome of a case, even if the explanations are logically mutually exclusive.

Process tracing has gone by various names and it has been around as long as people have been trying to explain historical events, but a significant recent innovation is the method of formal Bayesian process tracing. For present purposes, I very briefly review traditional process tracing, as practiced in political science since the 1970s (and informally for centuries before that), and then focus in more detail on formal Bayesian process tracing.

Process tracing has always relied on Bayesian logic, and many of the misunderstandings of case study methods have arisen from the inappropriate application of frequentist statistical thinking to case studies (George & Bennett, 2005). Until recently, however, traditional process tracing used Bayesian inference only informally. In the informal approach, the researcher develops a comprehensive set of alternative explanations for the outcome, drawing on prior theoretical knowledge and the explanations of analysts familiar with the case. Next, the researcher thinks deductively about what observable implications of the alternative explanations would likely be true if each explanation is true, paying particular attention to implications that are more likely under one explanation than the others. The researcher then gathers evidence on these implications, while remaining alert as

after comparisons of the same country, which controls for many possible variables not included in the theory that vary across countries but do not vary over time within the same country.

well to additional and previously un-theorized evidence from the case that may discriminate among the explanations. Finally, the researcher weighs the totality of the evidence and considers which of the explanations is most likely in light of that evidence.¹⁵

Formal Bayesian process tracing gives added precision to each of these steps, and it clarifies the logic that underlies traditional process tracing.¹⁶ The essential difference between frequentist and Bayesian logic is that frequentism treats probability as the limit of an event's frequency in a large number of repeated trials, while Bayesianism treats probability as our degree of belief in a proposition (which could be a hypothesis or it could be evidence) given the information we already know. In other words, for frequentists probability relates to the process of sampling from a population, while for Bayesians probability represents degrees of belief. Bayesians use evidence to update their initial degrees of belief in alternative explanations (called the "priors"), together with the relative likelihood of the evidence under the explanations (called the "likelihood ratio"), to update their degrees of belief in light of the evidence (called the "posteriors"). I outline each step in turn; for reasons explained below, it is convenient to express the priors and the likelihood ratios in the form of log odds rather than percentage probabilities.

Like traditional process tracing, formal Bayesian process tracing begins with developing alternative explanations for a case, drawing on extant theories as well as adding any new ones the researcher devises. The difference is that Bayesian process tracing gives even greater emphasis to stating these explanations precisely, and it clarifies that they must be logically mutually exclusive and, to the extent possible, exhaustive (Fairfield & Charman, 2022).¹⁷ This does not mean that the explanations need to be single-variable explanations, or that they must draw on non-overlapping sets of variables. They can even have the same exact set of variables, as long as they posit that these variables interact in different functional ways to produce the outcome, and/or associate different relative weights to the importance of these variables. For example, an internal combustion engine needs fuel, air,

¹⁵For recommendations on best practices in process tracing, see Bennett and Checkel 2015. For overviews of the process tracing literature and practical advice on the method in the context of political economy, see Trampusch & Palier, 2016, and Skarbek, 2020).

¹⁶This general overview of formal Bayesian process tracing is not meant to provide adequate training for using the method—for a detailed presentation and worked examples see Fairfield and Charman (2022).

¹⁷As Fairfield and Charman note, the exhaustivity assumption requires that one of the hypotheses considered is correct, so the assumption must be considered tentative as we can never be sure we have included a correct explanation among our hypotheses. But they suggest that methodologically the hypotheses should be treated as exhaustive, even if we must acknowledge that they may not be. As they put it, "we cannot ask for more than inference to the best *existing* explanation (Fairfield & Charman, 2022, p. 84). From a pragmatic viewpoint, they suggest leaving aside hypotheses that seem extremely unlikely under the researcher's background knowledge, and including any hypotheses that the researcher or experts in the relevant literature deem plausible. As they note, additional explanations can be added at any point in the research if inductive analysis of the evidence suggests this would be useful (Fairfield & Charman, 2022, pp. 84–86).

compression, and a spark to work. If it is not working, it could be a combination of small problems with each element, or it could be small problems in most elements but a bigger problem in one. These draw on the same variables but they are logically mutually exclusive and they posit different counterfactuals on what needs to be fixed for the engine to run.

The next step in formal Bayesian process departs more clearly from the traditional approach. In Bayesian process tracing the researcher explicitly assigns priors—the relative degrees of belief they attach to the alternative explanations of a case before they look at evidence from the case. These priors reflect the researcher’s existing theoretical and empirical knowledge. Priors need not be point estimates—they could be ranges between high and low estimates for each explanation, and the researcher can work through the Bayesian mathematics and end up with updated high- and low-end probabilities for the hypotheses.¹⁸

Next, the researcher identifies the observable implications of the alternative explanations, paying particular attention, as in traditional process tracing, to those that are more likely under one explanation than under the others.¹⁹ Bayesian logic clarifies why the relative likelihood of the evidence under the different explanations (the likelihood ratio) is critical, as Bayesian logic uses the laws of probability to move from the relative likelihood of the evidence under the different explanations to the relative likelihood of the explanations given the evidence. If a piece of evidence is equally likely under two explanations, it is uninformative, as it does not change the degrees of belief that we should invest in the explanations. The researcher estimates the likelihood of the evidence under each explanation, and its relative likelihood among all the explanations, by mentally “inhabiting the world” in which an

¹⁸Ideally, estimates of priors and likelihood ratios (discussed below) are derived from plentiful high quality evidence, such as insurance company actuarial tables. In social science practice, these estimates are based on the researcher’s limited background information, and experts with different background information might give different estimates. This “subjectivity” of likelihood ratios is a common concern in Bayesian analysis, but it is not necessarily a fatal flaw, as: (1) researchers can share the background knowledge and reasoning on which their estimates are based, and often this will narrow the differences in their estimates; (2) researchers might be able to agree that pieces of evidence are more likely under one explanation than another even if they don’t agree on how much, and they can assess whether their remaining differences matter in the end; (3) researchers’ explanations of how they made their estimates of likelihood ratios makes transparent the kinds of analysis researchers have to do anyway in informal process tracing. One question for future research is whether crowd-sourcing or expert elicitation of estimates of priors and likelihood ratios in some sense makes them more accurate. If a researcher has limited background information, they can start with “uninformed” priors, or assign equal prior likelihoods to all the alternative explanations. Assigning equal likelihood ratios would make no sense, however, as it would prevent any updating on evidence (and also the logic of alternative explanations is usually sufficiently different that most evidence has a higher likelihood under one than under another).

¹⁹An important point here is that the observable implications are never “on” or “for” only one explanation: the evidence has some probability (even if only near zero in some instances) under every alternative explanation, and it is the *ratio* of these probabilities that matters. Thinking terms of the logic of each explanation can help identify what evidence to look for, but the researcher still has to consider if that evidence is more, less, or equally likely under the other explanations as well.

explanation is true and assessing how likely a piece of evidence would be in that world (Fairfield & Charman, 2022).

Finally, the researcher gathers the evidence, again including evidence that they had not initially identified as an observable implication but that, once stumbled upon, appears to have different likelihoods under the alternative explanations. It is perfectly acceptable at this stage to introduce new explanations from inductive encounters with the evidence, as long as the researcher then reassesses the priors and the likelihood ratios with the new full set of alternative explanations.²⁰

The final step is to use Bayes' Theorem to derive the updated relative likelihoods of the explanations in light of the evidence. This is the step at which the researcher uses the laws of probability to move from the relative likelihood of the evidence under the different explanations to the relative likelihood of the explanations given the evidence. There are various forms in which to express Bayes Theorem—for present purposes the relative odds form is the most useful (Fairfield & Charman, 2022, pp. 75–78):

$$\text{posterior odds} = \text{prior odds} \times \text{likelihood ratio}$$

This is the formal Bayesian approach in a nutshell, but there are several complications and caveats (for details see Fairfield & Charman, 2022). First, it is convenient not just to use the odds form of Bayes' Theorem, but to express priors and relative likelihoods in log odds form, analogous to the decibel (dB) scale. Using the logarithmic scale and the decibel analogy takes some practice, but it greatly simplifies the mathematics that arise when there are multiple explanations and multiple pieces of evidence. Using regular odds, analysis of multiple pieces of evidence and multiple explanations requires many multiplicative terms, but using log odds makes these terms into simple operations of addition and subtraction. One can then “weigh” the evidence as if placing the pieces that favor explanation A over explanation B on one side of a weight scale and putting the pieces that favor B over A on the other side. The new posterior log odds is then simply the net difference (still in log odds) generated by the evidence, added to the log odds priors.²¹

²⁰For Bayesians, it does not matter if the explanations were deductively or inductively derived—all that matters is the relative likelihood of the evidence under the alternative explanations, which does not *logically* depend on whether a theory led to the evidence or the evidence led to the theory. It is possible that there is a difference *psychologically*, however, as one form of confirmation bias is to give different likelihood estimates depending on whether evidence is already known versus as-yet unknown. Fairfield & Charman, 2022 suggest consistency checks to root out such biases.

²¹Students have some difficulty learning and understanding this practice, but once they do understand it, Bayesian updating becomes much easier. An additional advantage is that the human perceptual apparatus, whether for sound, sight, or touch, works on a logarithmic scale (Fairfield & Charman, 2022, pp. 129–137). Students are also initially tempted to convert the posterior log odds back into percentages. This can certainly be done, and there is nothing wrong with it, but once one is comfortable with log odds and the analogous decibel scale, it is fine to keep the posterior relative odds in log form. For a table that converts among dB, odds, and probabilities, see Quick Bayes Table.

To provide perspective on the correspondence among the dB scale, relative odds, and percentages, here are some samples of equivalent quantities:

Decibels/dB	Odds	Percentages
0 dB	1:1	50%
3 dB	2:1	67%
6 dB	4:1	80%
10 dB	10:1	91%
13 dB	20:1	95%
20 dB	100:1	99%
30 dB	1000:1	99.9%

Thus, estimating as a prior or a posterior that a hypothesis is 6 or more dB more likely than an alternative explanation, or that a piece of evidence is 6 or more dB more likely under one explanation than under another, is a fairly strong claim, and estimates of 10 dB or more are extremely strong claims. Note that 13 dB corresponds to the significance level of $p = 0.05$ in frequentist studies, a level that was for a long time used (too mechanically) as the standard for when a study was worthy of being published in a journal.

As an example of how using the logarithmic dB scale allows easy summation of the weights of different pieces of evidence, assume our prior for a case favors explanation A over explanation B by 5 dB. Assume further that we have four pieces of evidence E1-E4, and we think E1 favors B by 7 dB (note that we record evidence that favors B as negative decibels for A), E2 favors A by 6 dB, E3 favors A by 8 dB, and E4 is equally likely under A and B (0 dB). Then, starting with the prior and weighing all four pieces of evidence together, our posterior assessment favors A over B by 12 dB, or 16:1 odds:

$$\begin{aligned}
 &5 \text{ dB (the prior)} - 7 \text{ dB (E1)} + 6 \text{ dB (E2)} + 8 \text{ dB (E3)} + 0 \text{ dB (E4)} \\
 &= 12 \text{ dB in favor of A over B}
 \end{aligned}$$

Second, when there are multiple alternative explanations, as there usually are, using the log odds approach the researcher arbitrarily picks one explanation, calls it H1, and compares the log odds likelihood of each piece of evidence between H1 and H2, H1 and H3, H1 and H4, and so on. After doing this for each piece of evidence, the results are then added together to get the sum log odds relative likelihood of each pair of hypotheses (H1:H2, H1:H3, H1:H4, etc.). Because these are all relative odds, by comparing H1 to all the other explanations one at a time the researcher has already implicitly compared the log odds of every explanation relative to every other explanation vis-à-vis the evidence through their log odds relative to H1. There is no combinatorial explosion of the number of comparisons one must make as the number of explanations or pieces of evidence increases.

Third, while this clarifies the logic of inference from evidence in individual cases, even proponents of the method do not advise writing up this whole analytic process for readers, as this would be exceptionally tedious for all concerned. Instead, if there

are a few pieces of particularly probative evidence, or evidence that is much more likely under one explanation than the others, the researcher might write up the formal analysis of only these pieces of evidence.

Finally, the method also requires considerable training to do well, and even then the results of formal Bayesian process tracing, like those of traditional process tracing, can only be as strong as the evidence and our background information allow. In some instance the evidence may not allow significant updating between priors and posteriors.

Despite these complications and limitations, formal Bayesian process tracing has distinct advantages (Fairfield & Charman, 2022). It clarifies and formalizes the logic that has always been behind traditional process tracing, and it can make more transparent the ways in which we are using both background knowledge and new evidence to make inferences in case studies. Even if researchers choose not to use the formal approach at all, or not to do so explicitly, by understanding Bayesian logic they can approximate it more faithfully in their informal analyses. Far too many process tracing accounts give insufficient attention to alternative explanations or fail to consider the likelihood of the evidence under all of the explanations. Articulating the Bayesian logic of process tracing also makes case study methods more intelligible and legitimate to researchers who primarily use quantitative methods. In the end, formal Bayesian process tracing makes explicit, and hopefully more systematic, the kind of analyses researchers already have to be carrying out to make causal claims on the basis of case study evidence.

40.6 Example: Formal Bayesian Analysis of Evidence in *Pawned States*

While Queralt's *Pawned States* uses traditional process tracing, applying formal Bayesian process tracing to his evidence illustrates this method, indicates how Queralt implicitly used Bayesian logic very well, and identifies a few ways in which he might have used it even more clearly and powerfully. I focus here on Queralt's case study of state building in Chile during the 1879–1883 War of the Pacific and its aftermath. Queralt notes that leading into the war, the 1870s global recession pushed Chile to brink of default, and in April 1879 it suspended its debt payments to foreign creditors just days after the War of the Pacific began, so it could not finance the war through additional foreign loans (Queralt, 2022, p. 276–288).²²

First, we begin with Queralt's alternative theories that might explain state-building in Chile in this period.²³

²² Queralt essentially has 3 Chile case studies, one for financing after each of the wars in 1836–1839, 1864–1866, and 1879–1883 (Queralt, 2022, pp. 273–288).

²³ I set aside some explanations Queralt discusses early in the book that are either complementary to his debt financing explanation or not very relevant to the case (colonialism, developmental finance, and other forms of finance such as expropriation of church funds; Queralt, 2022, pp. 19–24).

- H1 Foreign Finance undermines state building, Domestic Finance advances it
- H2 There is a “resource curse” regarding state building: high natural resource rents undermine state building, as “the availability of nontax revenue weakens incentives to initiate tax bargaining with taxpayers and to invest in the bureaucratic apparatus of the state” (Queralt, p. 19)
- H3 Social divisions such as ethnic heterogeneity undermine state building by inhibiting investment in public goods, including state bureaucracies; social unity contributes to state building (Queralt, p. 20)

Second, we need to assess a prior, in log odds form, of these explanations. The prior is driven by the researcher’s background knowledge before they look in detail at evidence on the case. If Queralt knew nothing at all about the case of Chile when he selected it for study, including what form of finance it used, whether it did or did not have natural resource rents or social divisions, and whether it did or did not engage in state building in this time period, then his prior would reflect his general theoretical expectations for state building in states like Chile in this time period. In that instance, he might expect Chile would have used foreign financing and engaged in limited state-building, as happened with most states outside of Europe in this period. But it appears that Queralt chose to study Chile in this period, after coding Chile and other cases, *precisely because* he already knew it was cut off from foreign financing, so it had to rely on domestic financing, and it seems he already knew as well that it did engage in state building in this period. If he also already knew that Chile had substantial natural resource rents and social divisions (how to measure social divisions, and what their level was in Chile, are rather ambiguous in his account), then he might put a low prior on their associated explanations, as they are inconsistent with the outcome of significant state building, and a high prior on the finance explanation, as it is consistent with the outcome in Chile.

So for the purposes of this exercise, let us treat this information—that Chile went to war for 4 years, that it was in default and could not get foreign financing during the war and for years afterward, that it engaged in considerable state building in this period, that resource rents were an important part of its economy, and that there were some social divisions—as background information that informs the prior odds among the hypotheses. Because the known outcome of the case—strong state building in Chile—is unlikely under H2 and H3 given these facts, but likely under H1, we might estimate a prior of 9 dB favoring H1 over H2, and 7 dB favoring H1 over H3 (the ambiguity of the social divisions hypothesis and of the background information on such divisions in Chile makes it hard to estimate H1:H3).

Queralt’s account then introduces four key pieces of evidence:

- E1 The war was over control of the Atacama desert, an area rich in nitrogen deposits used for fertilizer. The desert was within Bolivia’s borders, but 90 percent of its population was Chilean. Chile went to war in early April 1879 when Bolivia, contrary to an 1874 agreement, unilaterally raised export taxes on nitrogen fertilizer. Peru, a treaty ally of Bolivia, joined Bolivia in fighting against Chile (p. 276)

- E2 Cut off from foreign funding, in May 1879 the Chilean Congress passed the *mobilaria*, a tax on capital gains, securities, and high incomes—essentially a tax on the rich and the nitrogen that was the source of the wealth of many elites. The war led to substantial bureaucratic and political reform that strengthened the state beyond the post-war nitrate boom (p. 281)
- E3 After successfully seizing the nitrate fields, Chile quadrupled the export tax rate despite the strong political ties of the Antofagasta Nitrate and Railway Company to legislators and the cabinet (p. 278)
- E4 Jose Manuel Balmaceda, who became president in 1886 (after the war ended in 1883), pursued substantial state-building investments in infrastructure, the military, and education, even though the expansion of education threatened the oligarchs (who lacked technocratic skills), and he antagonized the nitrate industry by considering nationalizing it to pay for government programs. Balmaceda shut down the Congress in 1889 when it contested his decree to cancel the Nitrate Railways Company, and within 2 years tensions between Balmaceda and the Congress led to a civil war in which the Congress prevailed, leading to Balmaceda's suicide. Despite the ouster and death of Balmaceda, the Chilean tax apparatus, investments in public goods, and the number of tax personnel continued to grow after the civil war and into the early 1900s (pp. 281–288).

A Bayesian analysis might assess each piece of evidence in turn as follows.²⁴ First, we can arbitrarily choose H1 as the hypothesis to be compared to the other two (thereby also implicitly comparing H2 and H3). Then we assess each piece of evidence vis-à-vis two hypotheses at a time.

Evidence E1 seems likely when we step into the world of the resource curse explanation, as it appears that actors eager to grab nitrogen, even at the cost of a war, are running the show in Chile. Yet E1 is not unlikely under the finance explanation, as grabbing valuable land is a common motive for war. On balance E1 seems slightly less likely under the finance explanation than it is under the resource curse view, as we might expect a state to be reluctant to start a war when it is already having trouble paying foreign debts and will be cut off from further foreign financing, forcing it to raise unpopular taxes at home. So we might estimate that this evidence favors H2 over H1 by 3 dB. As for comparing H1 to H3 vis-à-vis E1, leaders might have hoped to reduce social divisions by starting a war with weaker adversaries and seizing valuable land, achieving a “rally around the flag.” On the other hand, wars can exacerbate social divisions in the longer term, so they create risks for leaders. So on balance we might estimate that E1 very slightly favors H3 over H1, by an estimated 2 dB. So:

²⁴Fairfield and Charman (2022, pp. 186–204) also provide a formal Bayesian analysis of state building in Chile, assessing the work of Kurtz (2009) and his hypotheses focused on labor repressive agriculture, the resource curse (both theorized as barriers to state building), and war (hypothesized as a stimulus to state building).

H1:H2 -3 dB

H1:H3 -2 dB

As for E2, the Congress's approval of a tax on the rich appears to strongly favor H1 over H2. A tax on the rich, including many members of the Congress itself, is highly surprising if the state were dominated by the beneficiaries of resource rents. The finance explanation might find a tax on the rich, rather than the middle class, mildly surprising, but it would expect some tax increase to finance a war when foreign financing was not available. So we might estimate that E2 favors H1 over H2 by 8 dB. As for H1 versus H3, a tax on the rich is not so surprising under H3, as the alternative of a tax on the middle classes and the poor to pay for a war that could benefit the rich could lead to social unrest. So we might estimate that E2 favors H3 over H1 by 3 dB:

H1:H2 8 dB

H1:H3 -3 dB

The third piece of evidence—the quadrupling of the export tax on nitrate fertilizer and other goods—is highly likely under the finance explanation. Lacking access to foreign credit, we would expect the state to tax what it can to pay for the war, despite domestic opposition. In contrast, this evidence is extremely unlikely under the resource curse hypothesis, as it cuts directly against the logic of this hypothesis. So we might estimate that it favors H1 over H2 by 10 dB. Perhaps because this evidence so strongly discriminates between these hypotheses, Queralt gives it considerable attention, quoting historian William Sater, the author of two books on the war, at length (Sater, 1985, p. 140). A key excerpt (p. 278):

The passage of the nitrate export tax surprised many . . . [as the *Chilean Times* noted] “Large sums of money and the influence of many of the most important men in the country have failed to prevent the bill from passing [by] a very large majority.”

While Queralt is intuitively aware that this is important evidence, he does not explain to the reader the Bayesian reason for this: *because it is far more likely under one explanation than under another*. He also does not explicitly consider whether and how this evidence discriminates between H1 and H3. In fact, a tax on nitrate exports is not very surprising if the government fears unrest if it were to tax citizens more broadly. Still, such a tax seems more likely under H1 than H3, perhaps by about 4 dB:

H1:H2 11 dB

H1:H3 4 dB

The fourth piece of evidence is even more critical, as it discriminates sharply between H1 one and *both* of the rival explanations. Our background information already stipulated that Chile engaged in state building in this period, but E4 adds key

details on the timing of this effort and on the social divisions it exacerbated. President Balmaceda took on the nitrate oligarchs even more directly than indicated by E3, challenging their property rights and pushing conflict on this issue to the point that a civil war broke out between Balmaceda and the Congress. Yet even after prevailing in the civil war, the Congress picked up where Balmaceda left off in the state building project, continuing to expand the state’s taxation capacities and the number of its tax personnel into the early 1900s. This is shockingly unlikely under the resource curse explanation, and sharply contrary to both its expected dynamics and its expected outcome.²⁵ It is also highly unlikely under the social divisions explanation, as we have the ultimate social division—civil war—and yet the state-building project continues. Yet this outcome is highly likely under the finance explanation and its path-dependent consequences: having found its way, out of necessity, to domestic rather than foreign financing, the Chilean state further expanded its domestic taxation capacities. To the extent that it is surprising under this explanation, it is because the explanation is over-fulfilled—ex ante, we might not have guessed that the state- and taxation-building momentum would be so strong that it would survive a civil war over taxation, especially when the anti-tax side won. So we might estimate:

H1:H2 16 dB
H1:H3 16 dB

The final step in formal Bayesian analysis is simply to add up the weights of the evidence to get the posterior log odds ratios among the hypotheses:

	Priors	E1	E2	E3	E4	Total/Posterior
H1:H2	9 dB	−3 dB	8 dB	11 dB	16 dB	41 dB
H1:H3	7 dB	−2 dB	−3 dB	4 dB	16 dB	22 dB

This analysis clarifies several points. First, Queralt’s traditional process tracing tracks the more formal Bayesian approach fairly well, and the Bayesian results affirm Queralt’s conclusion, extremely strongly favoring the finance explanation over the resource curse and very strongly favoring the finance explanation over the social divisions explanation for the case. Second, one need not agree exactly with the proposed estimates of the priors and likelihood ratios, nor is it necessary to agree with every detail of the reasoning, to get to this result. If the reasoning and the likelihood estimates are even approximately right and in the correct direction, a strong conclusion in favor of the finance explanation is warranted. Third, Queralt could have made his reasoning more transparent, and perhaps he could have done an even better job of process tracing, by approximating Bayesian reasoning more

²⁵This is especially so if part of our background information from the resource curse literature is that resource rich elites will focus any state-building efforts on a security apparatus to keep themselves in power, and will limit funding to public goods like education so they can keep as much of their resource rents as possible.

closely, even if he did not use the formal apparatus of assigning numerical estimates to his priors and likelihood ratios. His social divisions explanation, and the evidence relating to it, for example, could be more clearly specified, and he could give an even clearer and stronger rationale for why E3 and especially E4 are critically important to his conclusions.

Assigning numerical priors and likelihood ratios, especially in the logarithmic dB scale, can seem a bit arbitrary. What matters most, however, is the background information that informs the prior and the reasoning relating the logic of the hypotheses to the relative likelihood of each piece of evidence under the hypotheses. The Bayesian formalization forces us to ask exactly the right questions, and clarifies where an author and their readers might disagree. In this example, a reader would have to disagree substantially with the assessment of Queralt's evidence to reach a different conclusion, and it would be incumbent on the reader to explain where and why their estimates differ. The strongest argument for learning formal Bayesian analysis of evidence is that it clarifies and adds logical consistency to the kind of assessments researchers have to be making informally anyway in order to make explanatory claims about cases based on process tracing evidence.

40.7 Conclusions

The comparative advantages of case study methods, relative to quantitative methods, closely fit the intellectual interests of institutional economists. Institutional economists are interested in big questions, critical junctures, historically important cases, anomalous cases, path dependency, and high-order interaction effects, and they often face the methodological challenges of addressing potential endogeneity and measuring complex concepts. Case study methods are strong at each of these types of processes and research challenges, and in general they are better able than observational quantitative methods to get close to the action of causal mechanisms in individual cases.

Accordingly, institutional economists have been making excellent use of case study methods for decades. Yet many of the extant case studies in institutional economics (and in the social sciences more generally) apply case study methods informally, and few economists have formal training in case study methods. Institutional economists have intuitively understood the value of process tracing and of conceptual and theoretical typologies, and using their more recent variants of formal Bayesian process tracing and typological theorizing is not likely to radically alter the main findings in the field. These new methods might add transparency in how scholars are making inferences from case study evidence, however, and help narrow theoretical and empirical disagreements within institutional economics. More rigorous use of case study methods will also help the quantitative scholars who dominate many other subfields in economics understand and appreciate the qualitative research carried out by institutional economists. Typological theorizing clarifies the process of case selection, and it is a useful approach for theorizing about high-order interactions. Formal Bayesian processing tracing can be useful in analyzing the

most critical pieces of evidence in a case study, even if the proponents of this method do not advise using it and writing it up fully on every piece of evidence. Even if one does not explicitly deploy formal Bayesian process tracing at all, simply understanding the logic of this method can improve process tracing inferences and clarify to both quantitative and qualitative scholars how these inferences are made.

Acknowledgments I would like to thank Tasha Fairfield and Didac Queralt for comments on an earlier draft of this chapter.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Allen, D. W., & Lueck, D. (2008). Agricultural contracts. In C. Ménard & M. Shirley (Eds.), *An advanced introduction to new institutional economics*. Elgar Publishing.
- Alston, L. (2008). The 'case' for case studies in new institutional economics. In E. Brousseau & J.-M. Glachant (Eds.), *New institutional economics: A guidebook* (pp. 103–121). Cambridge University Press.
- Antonelli, C. (1997). The economics of path-dependence in industrial organization. *International Journal of Industrial Organization*, 15(6), 643–675.
- Beckmann, V., & Padmanabhan, M. (2009). Analysing institutions: What method to apply? In V. Beckmann & M. Padmanabhan (Eds.), *Institutions and sustainability*. Springer.
- Benham, L. (2008). Licit and illicit responses to regulation. In C. Ménard & M. Shirley (Eds.), *An advanced introduction to new institutional economics*. Elgar Publishing.
- Bennett, A. (2013). The mother of all isms: Causal mechanisms and structured pluralism in International Relations theory. *European Journal of International Relations*. <https://doi.org/10.1177/1354066113495484>
- Bennett, A. (2022). Drawing contingent generalizations from case studies. In D. O. Nieto, J. Widner, & M. Woolcock (Eds.), *The case for case studies methods and applications in international development*. Cambridge University Press.
- Bennett, A., & Checkel, J. (Eds.). (2015). *Process tracing: From metaphor to analytic tool*. Cambridge University Press.
- Bennett, A., & Elman, C. (2006). Complex causal mechanisms and case study methods: The case of path dependence. *Political Analysis*, 14, 250–267.
- Bennett, A., & Mishkin, B. (2023). Nineteen kinds of theories about mechanisms that every social science graduate student should know. In H. Kincaid & J. van Bouwel (Eds.), *The Oxford handbook of philosophy of political science*. Oxford University Press.
- Brady, H., & Collier, D. (2010). *Rethinking social inquiry* (2nd ed.). Rowman and Littlefield.
- Carey, J. M. (2008). Presidential versus parliamentary government. In C. Ménard & M. Shirley (Eds.), *An advanced introduction to new institutional economics*. Elgar Publishing.
- Coase, R. H. (1974). The lighthouse in economics. *The Journal of Law and Economics*, 17, 2.
- Collier, D., Laporte, J., & Seawright, J. (2012). Putting typologies to work: Concept formation, measurement, and analytic rigor. *Political Research Quarterly*, 65(1), 217.
- Elman, C. (2005). Explanatory typologies in qualitative studies of international politics. *International Organization*, 59(2), 293–326.
- Engerman, S. L., & Sokoloff, K. L. (2008). Institutional and non-institutional explanations of economic differences. In C. Ménard & M. Shirley (Eds.), *An advanced introduction to new institutional economics*. Elgar Publishing.

- Fairfield, T., & Charman, A. (2022). *Social inquiry and Bayesian inference: Rethinking qualitative research*. Cambridge University Press.
- George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. MIT Press.
- Goertz, G., & Mahoney, J. (2005). Two-level theories and fuzzy-set analysis. *Sociological Methods & Research*, 33(4), 497–538.
- Greif, A. (2008). Commitment, coercion, and markets: The nature and dynamics of institutions supporting exchange. In C. Ménard & M. Shirley (Eds.), *An advanced introduction to new institutional economics*. Elgar Publishing.
- Hill, A. D., Johnson, S. G., Greco, L. M., O’Boyle, E. H., & Walter, S. L. (2021). Endogeneity: A review and agenda for the methodology-practice divide affecting micro and macro research. *Journal of Management*, 47(1), 105–143.
- Kurtz, M. (2009). The social foundations of institutional order: Reconsidering War and the ‘Resource Curse’ in third world state building. *Politics and Society*, 37(4), 479–520.
- Mahoney, J. (2000). Path dependence in historical sociology. *Theory and Society*, 29(4), 507–548.
- Mahoney, J., & Goertz, G. (2012). *A tale of two cultures: Qualitative and quantitative research in the social sciences*. Princeton University Press.
- Ménard, C., & Shirley, M. (2008). *An advanced introduction to new institutional economics*. Elgar Publishing.
- Ménard, C., & Shirley, M. (Eds.). (2022). *Handbook of new institutional economics*. Springer-Verlag.
- North, D. C., & Thomas, R. P. (1973). *The rise of the Western world: A new economic history*. Cambridge University Press.
- Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
- Ostrom, E. (2008). Doing institutional analysis: Digging deeper than markets and hierarchies. In C. Ménard & M. Shirley (Eds.), *An advanced introduction to new institutional economics*. Elgar Publishing.
- Page, S. (2006). Path dependence. *Quarterly Journal of Political Science*, 1(1), 87–115.
- Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American Political Science Review*, 94, 251.
- Pierson, P. (2004). *Politics in time: Politics in time: History, institutions, and social analysis*. Princeton University Press.
- Queralt, D. (2022). *Pawned states: State building in the era of international finance*. Princeton University Press.
- Ragin, C. (1987; second edition 2014). *The comparative method: Moving beyond qualitative and quantitative strategies*. University of California Press.
- Sater, W. (1985). *Chile and the war of the Pacific*. University of Nebraska Press.
- Skarbek, D. (2020). Qualitative methods for institutional analysis. *Journal of Institutional Economics*, 16(4), 409–422.
- Trampusch, C., & Palier, B. (2016). Between X and Y: How process tracing contributes to opening the black box of causality. *New Political Economy*, 21(5), 437–454.
- Williamson, O. (1999). The new institutional economics: Taking stock/looking ahead. *Newsletter of the International Society for New Institutional Economics*, 2, 2.
- Williamson, O. (2008). Transactions costs economics. In C. Ménard & M. Shirley (Eds.), *An advanced introduction to new institutional economics*. Elgar Publishing.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Analyzing the Deep Determinants of Institutions: Methodological Choices and Challenges

41

Robbert Maseland

Contents

41.1	Introduction	1090
41.2	Determinants of Institutions: An Overview	1091
41.2.1	Natural Environment	1091
41.2.2	Historical Determinants	1092
41.2.3	Cultural Determinants	1093
41.2.4	Mechanisms of Persistence	1095
41.3	Developments in Empirical Approaches	1097
41.3.1	From Institutions as Cause to Institutions as Outcome	1097
41.3.2	From Countries to Grids	1099
41.3.3	From Cross-Sections to Case Studies	1101
41.4	Critical Challenges	1103
41.4.1	Operationalizing Institutional Diversity	1103
41.4.2	Independence of Observations	1104
41.4.3	Data Quality	1105
41.4.4	Compression of History	1106
41.4.5	Reading History Backwards	1107
41.5	Concluding Thoughts	1109
	References	1111

Abstract

This chapter discusses the literature on the determinants of institutions from a methodological perspective. It classifies determinants into three categories—the natural environment, historic events, and culture—and critically discusses the transmission mechanisms via which these variables are theorized to affect contemporary outcomes. Against this background, the chapter offers a critical discussion of the dominant methodological approaches that have been used to identify the impact of determinants. It argues that the literature's emphasis on causal inference has led to a relative neglect of questions about the validity and

R. Maseland (✉)
 Radboud University, Nijmegen, Netherlands
 e-mail: robbert.maseland@ru.nl

interpretation of the data going into analyses and questions about the evolutionary mechanisms linking past determinants to present outcomes. While a turn to historical narratives may remedy some of these concerns, a fuller appreciation of the disciplinary contexts from which the literature borrows (history, and ethnography in particular) is imperative if the literature seeks to translate its intriguing findings into fuller insights on institutional development.

Keywords

Deep determinants · Persistence · Methodology · Institutions · Historical analysis · Culture

41.1 Introduction

What drives the vast differences in prosperity between societies that we observe around the world? Explanations focusing on “proximate determinants” of growth, such as differences in (human) capital and overall efficiency, are ultimately unsatisfactory. That is why economists have shifted their attention to the underlying “deep determinants” (Rodrik, 2003) driving the differences in capital and efficiency around the world.

In this search for fundamental drivers of prosperity, several strands of literature come together. On the empirical side, the introduction of comparable cross-country income data (Summers & Heston, 1991) spawned an expansive literature on the drivers of economic growth (Barro, 1991). Advances in econometric design fueled this development, allowing economists to plausibly infer causality (Angrist & Pischke, 2010). On the theoretical side, the work of institutional economists like North (1990) and Greif (1993) increasingly directed economists to institutions and governance as key ingredients for economic development. The question of the origins of wealth disparities evolved into the question of the origins of growth-supporting institutions.

While others before them had sought to identify determinants of institutional quality as root causes of differences in economic development (e.g., Engerman & Sokoloff, 1997; LaPorta et al., 1999), Acemoglu et al. (2001) were arguably the first in which all the above-mentioned elements—the theoretical focus on institutional quality, large-scale quantitative analysis, and a research design allowing causal inference—came together. Others followed suit, resulting in a wide collection of historical, geographical, and cultural variables identified as determinants of institutional quality (Abad & Maurer, 2021; Maseland, 2021; Spolaore & Wacziarg, 2013).

This chapter provides an assessment of this literature’s main contributions, focusing on its methodology. I begin by sketching the theoretical background of the literature, organizing the field according to the types of mechanism and determinant identified (Sect. 41.1). After laying this necessary groundwork, I discuss the most common methodological strategies to assess the root causes of institutions, and review trends in them (Sect. 41.2). Subsequently, I discuss a number of

methodological challenges in Sect. 41.3. In this discussion, I largely skip over details of econometric design to focus on the more fundamental methodological questions underlying the literature as a whole. Concluding remarks follow.

41.2 Determinants of Institutions: An Overview

Institutional explanations of economic divergence require reasons for institutional variation. While authors differ on these reasons, the common argument underneath all of them is that institutions are somehow *embedded* in a society’s wider context (Williamson, 2000). Such embeddedness directs societies toward one of multiple steady states. Different authors have focused on different elements of the context.

41.2.1 Natural Environment

The first group of papers focuses on the natural environment. In its simplest form, that idea is that institutions are adapted to their geographical, climatic, and biological context. Since natural conditions are fairly stable, resulting institutional differences are likely to be persistent as well. As a result, there are multiple steady states, each corresponding to a different set of stable natural conditions. The reasoning is captured in Fig. 41.1a.

More commonly, natural environments have been argued to have shaped institutions in the past, which are persistent (Fig. 41.1b). In this variant, the distinction between natural determinants and historical drivers (Sect. 41.2.2) becomes blurred.

Various aspects of the environment have been addressed. Going back to Ibn Khaldun and de Montesquieu, a long list of authors has focused on climate as a driver of cultural and institutional differences (Easterly & Levine, 2003; Mayoral & Olsson, 2022; Oumer et al., 2020). Related to climate, suitability for agriculture also matters. Early adoption of agriculture has been argued to spur state formation and more advanced institutional development (Hibbs & Olsson, 2004), while also being associated with more collectivist attitudes and less democratic institutions (Olsson & Paik, 2020). Similarly, areas suitable for plow use have been shown to develop less

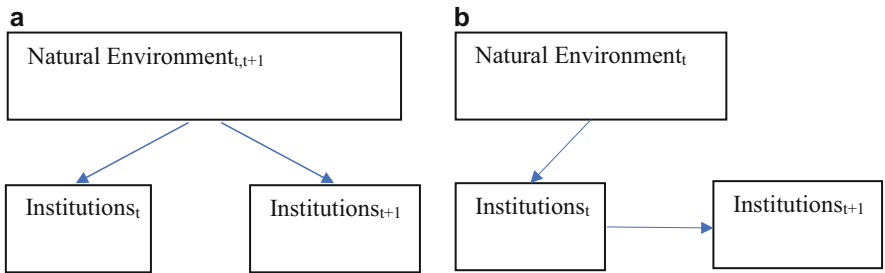


Fig. 41.1 Pathways of natural determinants of institutions (a) Direct effects (b) Indirect effects

egalitarian gender norms (Alesina et al., 2013), a factor associated with lower propensity for democracy (Norris & Inglehart, 2002).

Another strand of literature has studied the cultural and institutional impact of pathogen prevalence. Environments with more infectious diseases are argued to induce individuals to close off their communities and ward off potentially infected outsiders. This breeds collectivist attitudes, introversion, and a lack of openness to new experiences, affecting democratization negatively (Thornhill et al., 2009). Maseland (2013) shows a more contemporaneous link between pathogens, culture, and institutions, arguing that certain relatively common parasites induce small changes in personality that via interpersonal interaction result in systematic cultural differences at society level. Focusing on formal institutions, Alsan (2015) links the prevalence of parasites carried by the TseTse fly to lower use of domesticated animals, which obstructs state formation (Link, 2023).

Geography may also affect institutional outcomes via its influence on factor endowments. Engerman and Sokoloff (1997) discuss how factor endowments shape distributional conflicts, with long-term consequences on political-economic institutions. Others have pointed to the presence of mineral deposits (Tsui, 2011), distance to the coast (Acemoglu et al., 2005a; Nunn & Wantchekon, 2011), and terrain ruggedness (Nunn & Puga, 2012). In each of these cases, geography is thought to affect the pay-off of different institutional outcomes. Thus, the necessity of democratic accountability is lower in oil-rich countries where states do not rely on taxes for their budgets, inducing elites to opt for autocracy. Trust and institutional quality are lower in places where the lower costs of shipping slaves to coastal trading posts induced agents to engage in slave raids (Nunn & Puga, 2012; Nunn & Wantchekon, 2011). Relative geographical variables, such as the distance to European powers via sailing routes and ocean currents (Feyrer & Sacerdote, 2009), or the length of land routes from East Africa (Ashraf & Galor, 2013) have been shown to matter in the same way, changing the relative costs of adopting various qualities of institutions.

41.2.2 Historical Determinants

A second group of papers focuses on historical determinants outright. In this reasoning, certain historical events have caused societies to develop particular sets of institutions. Once in place, these institutions are transmitted over generations because institutional development is path-dependent (North, 1990). The argument is summarized in Fig. 41.2.

The most prominent example of a historical determinant is legal origins (LaPorta et al., 1999). The legal origins-argument states that legal traditions, typically inherited from colonial powers, shape a broad range of laws and regulations, which influence current institutional and economic outcomes. Similar arguments have been made about the persistent impact of other aspects of colonial rule. Various authors have claimed that British imperialism produced better outcomes for its colonies than other colonizers, while others have emphasized the benefits of direct versus indirect rule or stressed the duration of colonialism as important variable

Fig. 41.2 Pathways of historical determinants of institutions

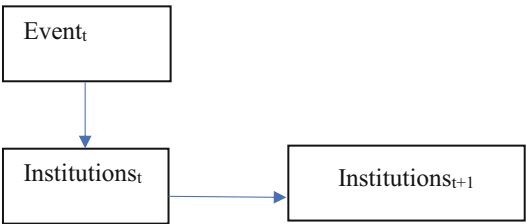
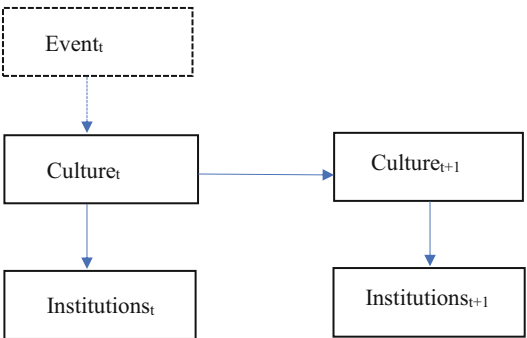


Fig. 41.3 Pathways of cultural determinants of institutions



(Maseland, 2018). Historical arguments are not confined to colonies. Becker et al. (2016) show that the former territories of the Habsburg Empire continue to have higher trust levels and less corruption today than adjacent regions outside the empire’s historical borders. Guiso et al. (2016) report that Italian cities that achieved self-governance in medieval times enjoy higher levels of civic capital today. In all these arguments, the effects of historical conditions are transmitted between generations, driving long-run differences in development.

41.2.3 Cultural Determinants

A final class of papers focuses on embeddedness in culture. Here, the idea is that formal institutions are rooted in an underlying informal set of norms, values, and beliefs. Since culture is stable over long periods of time (Hofstede, 2001), it sets constraints for institutional development that are similar from one period to the next (Williamson, 2000). Figure 41.3 visualizes the reasoning. Culture informs the formal institutional setting. Since culture at time t (potentially triggered by an exogenous contemporary event) is reproduced over time, the accompanying institutions are persistent as well. Note that the argument crucially differs from the ideas discussed in the previous section in that it is underlying culture that is reproduced over time, rather than institutions themselves. Institutional persistence is a side-effect of cultural persistence.

While Fig. 41.3 depicts a clean separation, the conceptual distinction between culture and institutions is ambiguous at best. To North (1990), culture is the informal aspect of institutions; Hofstede (2001) sees institutions as part of culture. Others maintain a distinction between the two, although the boundary is sometimes blurry (Mokyr, 2014). Even with the conceptual distinction there, empirically separating the two remains challenging (Maseland, 2013).

Cultural arguments fit in a long tradition of work relating differences in development to norms, values, or beliefs (see Beugelsdijk & Maseland, 2010). Classic contributions by de Montesquieu and Weber have linked variability in economic success to differences in propensity to hard work (Oumer et al., 2020). What makes these contributions controversial is that they start out from a distribution of power and income in the world, and then seek to trace that distribution back to cultural motives of people. Especially where authors come from the more privileged segments of the political-economic distribution, such arguments can easily appear as justifications for inequality. Variations of de Montesquieu's argument have thus been used to legitimize colonial rule (Alatas, 1977). More recently, cultural narratives have surfaced in analyses of the Greek debt crisis, attributing the crisis to moral failings of "Mediterranean" cultures (Rathbun et al., 2019), or in explanations of persistent poverty in Latin America (Harrison, 2000).

A challenging aspect of studies into cultural determinants is the measurement of culture. An analogy with the distinction between revealed and stated preferences approaches in consumer theory is instructive here.

Revealed preferences approaches infer underlying dispositions from the behavioral choices of individuals. This approach has widely been used to elicit cultural values and norms. Cross-cultural experiments capture differences in cultural orientations by comparing behaviors in experiments such as ultimatum games played in various settings (e.g., Henrich et al., 2005). Any structural differences in behavior between populations are then attributed to culture. A similar reasoning underlies the epidemiological approach, which elicits cultural dispositions by assessing differences in behavior between migrants coming from various settings. An early example is Fisman and Miguel's (2007) analysis of parking tickets picked up by diplomats in New York. Their finding that diplomats, all operating in the same setting, exhibit the behavior prevalent in their respective home countries suggests that at least part of this behavior is a cultural norm. Fernández and Fogli (2009) build on this, using differences in fertility rates among second-generation migrants in the United States as a proxy for cultural differences in fertility norms. While useful in identifying cultural differences, revealed preferences approaches have only limited value when culture is used as determinant of institutions. The reason is that the attribution of behavioral variation between groups to culture is only plausible when we can rule out all alternative drivers of behavioral differences. If cultural variation is related to institutional variation, that is not the case.

Stated preferences approaches capture cultural differences through surveys such as the World Values Survey (Inglehart et al., 2020) or Hofstede (2001). A prominent example is Gorodnichenko and Roland's (2017) investigation in the economic effects of individualism and collectivism. Capturing differences in individualism

via Hofstede's (2001) measures, Gorodnichenko and Roland (2017) argue that individualist cultures have a dynamic advantage over collectivist societies because they award originality, out of out-of-the-box thinking, and individual success, which fosters innovation. In addition, emphasis on individual rights and accountability results in higher quality formal institutions, whereas high trust and a focus on the collective good act as a substitute for the development of rule of law in collectivist societies. The latter argument has been further explored in Gorodnichenko and Roland (2021), linking individualism to democratization.

41.2.4 Mechanisms of Persistence

Studies of institutional determinants all trace contemporary differences in institutional and economic outcomes back to events or conditions in the (distant) past, arguing that differences in outcomes are *persistent*. A key question is what drives such persistence. Various mechanisms of historical transmission can be distinguished.

The first distinction is between **outcome** and **equilibrium dependence** (Page, 2006). In persistence due to **outcome dependence**, a single equilibrium exists, but convergence to this equilibrium is slow. Institutional settings are to a large extent transmitted from one generation to the next. As a result, shocks to institutions in the past continue to have an impact long after the initial event. Spolaore and Wacziarg (2013) identify two variants of historical transmission. In *biological transmission*, members of a society inherit certain biological traits that impact their attitudes and behavior directly from their ancestors. The resulting attitudes and behaviors subsequently have an impact on institutional and economic development. *Cultural transmission* implies that older generations transmit ideas, norms, values, and behaviors to the newer ones through socialization (Bisin & Verdier, 2023). Institutional persistence occurs because the rules of the game are passed on through socialization at home or in schools, or via legal texts.

Transmission is never perfect in either variant. Texts are interpreted differently from one generation to the next. Parents have a preference for transmitting the ideas they hold dear, but also for setting up their offspring to succeed in tomorrow's economy and society. Where these two objectives conflict, parents face a trade-off resulting in less than complete transmission. Similarly, biological transmission is prone to mutations. For this reason, the effects of historical determinants should wear off over time, even if only very gradually. Also, in the same way that initial conditions trigger divergent outcomes, this resulting distribution of outcomes is sensitive to new shocks. In the end, where it is only due to outcome dependence, persistence is limited.

Equilibrium dependence refers to persistence due to multiple equilibria. In this type of argument, historical shocks have pushed societies onto different trajectories of institutional development, culminating in different steady states. Overcoming divergence is not just a matter of time, but requires new shocks large enough to switch equilibrium paths.

Equilibrium dependence requires the presence of some form of negative externalities, such as feedback loops or self-reinforcement (Page, 2006). *Self-reinforcement* occurs when institutional arrangements come with complementarities that render the choice of alternative institutional arrangements less likely. Acemoglu et al.'s (2005b) argument about the symbiosis between economic and political institutions taps into such mechanisms (see also Acemoglu et al., 2021; Acemoglu & Robinson, 2023). Once certain economic institutions are in place, they distribute economic resources. The economic winners of this institutional arrangement employ these economic resources to acquire political power, and subsequently use this power to prevent any changes to the institutional status quo. In this way, political and economic inequality persists as elites become entrenched. Hall and Soskice's (2001) explanation for the existence of varieties of capitalism also revolves around self-reinforcement. Institutions, in their view, are dependent for their functioning on other institutions. Such institutional complementarities foster stability of multiple institutional constellations.

Feedback loops refer to situations in which the pay-off of any individual behavior (as well as the pay-offs of alternative options) depends on the choices of others. Such feedback is a fundamental aspect of institutions, stemming from the fact that institutions are public goods. Rules are rules only if they are recognized and adhered to by the collective. An individually determined speed limit is not a speed limit but an opinion. Since the functioning of institutions depends on them being shared and recognized by others, the collective adherence to one set of rules reduces the pay-off of adopting one's own, alternative rules (Searle, 2005). For this reason, institutional change is a collective action problem. To the individual, institutions are a given.

Two things ought to be noted about these arguments. The first is that, while economists loosely refer to path-dependence when describing the mechanisms connecting historical events to current institutional outcomes, this term is strictly incorrect. Path-dependence refers to processes in which choices today are dependent on the path of previous outcomes. In contrast, the stability of institutions in the arguments presented above derives from the contemporary presence of feedback loops or complementary institutions, irrespective of the historical trajectory that has led to this situation. Individuals reproduce the existing institutional setting because other individuals do the same, not because their great-grandparents did the same. The fact that they tend to make the same calculation in each period implies stasis, but not path-dependence.

Second, the degree of stability or divergence in these arguments is often overstated as a result. If—as in Acemoglu et al. (2005b)—economic institutions are the product of political institutions set in the period prior and vice versa, dependence for each goes back only one period. Yet this mechanism has been proposed to connect historical characteristics (settler mortality in colonies) and economic outcomes (institutional and economic performance) centuries apart. Empirically, persistence studies have by and large focused on early periods determining the selection of paths. Theoretically, the mechanisms proposed to explain persistence focus only on recent conditions.

Fundamentally, the stability of an institutional equilibrium appears not so much a function of its historical depth, but of its degree of embeddedness in a wider

institutional structure. The higher the number of contemporary institutional interdependencies, the harder it is to break out of an equilibrium and change institutional systems. Self-reinforcement in the Acemoglu et al. (2005b), Acemoglu et al. (2021), and Acemoglu and Robinson (2023) case leads to relatively modest amounts of persistence, because any new shock to the political or economic distribution of power immediately sets society onto new paths. Institutional persistence in Hall and Soskice (2001) is much stronger because resistance to change stems not from interdependencies between two sets of institutions, but from embeddedness in a whole constellation of interdependent, complementary institutions. Inertia due to feedback loops is rooted in massive amounts of interdependencies between the actions of all individuals in society, making change even more difficult.

41.3 Developments in Empirical Approaches

41.3.1 From Institutions as Cause to Institutions as Outcome

First attempts to assess the economic relevance of institutions ran into the difficulty that institutions are not exogenous. They are likely to be as much influenced by economic outcomes as the other way around, and both growth and institutions may very well be jointly driven by third variables such as geography.

The first empirical papers addressing determinants of institutions did so as attempts to remedy this endogeneity. Acemoglu et al. (2001) introduced settler mortality as an instrument for institutional quality that allowed them to capture the impact of institutions on economic performance. In these types of studies, “deep” determinants are being used as identification strategy. Drivers of institutions are identified that are exogenous to economic outcomes, using these to estimate the impact of institutions on economic performance. The basic logic of these instrumental variable (IV) designs runs as follows. Suppose we have:

$$Y_i = \mu + \beta_o \text{INST}_i + X_i' \gamma + \varepsilon_i \quad (41.1)$$

describing the effect of institutions INST on economic performance Y for society i , controlled for a vector of covariates $X_i' \gamma$. Since INST is likely to be endogenous, it may be modeled as

$$\text{INST}_i = \mu_{\text{INST}} + \beta_{\text{INST}} DD_i + X_i' \gamma_{\text{INST}} + v_{\text{INST}i} \quad (41.2)$$

where DD_i is an exogenous “deep determinant” of institutions in society i . The first step is to estimate Eq. 41.2. In the second step, researchers estimate Eq. 41.1, replacing the actual values of INST_i by the predicted values of INST_i emerging out of the estimation of Eq. 41.2. This two-stage procedure allows researchers to assess the causal effect of institutions on economic performance. In practice, both regressions are estimated simultaneously.



Fig. 41.4 Deep determinants, institutions, and economic outcomes

A characteristic of the empirical literature on institutions is that most of the instruments used to predict institutions are situated in history, as we have discussed in Sect. 41.2. In other words, the variables in Eqs. (41.1) and (41.2) are measured at different points in time. Theoretically, a third relation is therefore implied in between Eqs. (41.1) and (41.2), stating that institutions at time t (the period in which the effects on economic performance are estimated) are the product of institutions at time $t-h$ (the period in which the deep determinants have shaped institutions). The full logic is depicted in Fig. 41.4. Deep determinants affect institutions at time $t-h$, which in turn influence institutions at time t , which drive economic outcomes. Institutions at time $t-h$ may not always be directly observed, and a proxy is used instead.

An important consideration in interpreting IV designs is that the effects estimated are always local average treatment effects (LATE) only (Bisin & Moro, 2021). That is, they only reflect the impact of institutions on outcomes for the part of the population that responded to the exogenous deep determinant in the past. Were the rest of the population to differ systematically from this sub-section, persistence may be under- or overestimated relative to average treatment effects. Bisin and Moro (2021) recommend modeling the response to the treatment in order to facilitate interpretation. If such behavioral models suggest that the part of the population complying to the instrument is likely to portray greater (smaller) persistence and treatment effects than the overall population, the LATE is over- (under-)estimating persistence and treatment effects. Although reflecting upon the potential bias present in LATE effects explicitly is important, it should be noted that any behavioral models of compliance are just formal expositions of assumptions by the author. In the historical context, these assumptions are typically not testable.

Early on, almost all of the literature into determinants of institutions took the form depicted in Fig. 41.4. A switch from institutions as a cause to institutions as the outcome of interest occurred gradually. Rodrik et al. (2004) went a step beyond the standard instrumental variable framework, looking into determinants of institutions to argue that the effects of deep determinants of growth such as geography largely run via institutions. In this model, institutions are the mediator rather than the cause. While in this argument the interest in determinants of institutions is no longer purely instrumental, the ultimate variable of interest remained economic growth rather than institutions themselves.

The increasing consensus about the relevance of institutions for economic performance caused the literature to shift its emphasis to finding novel correlates of institutional and economic outcomes, rather than reproducing the result that institutions matter. Sometimes, the reasoning linking determinants to economic outcomes

via their effect on institutions was still present, but no longer the main contribution. Ashraf and Galor (2013), for instance, estimate a channel from genetic diversity (proxied by migratory distance to Africa) to innovation and social capital and finally to economic performance. The paper has raised interest mostly because of the argument linking genetic diversity to these outcomes, rather than because of its argument about the effects of innovation and social capital on income. Easterly and Levine (2016) go a step further, theorizing about but skipping direct tests of an institutional channel in the relation between European colonial settlers and economic performance today. Other work omits the economic performance link, focusing instead on the explanation of differences in institutional quality. Guiso et al. (2016) argue that social capital today is linked to historical social capital, which is in turn driven by medieval city-state independence instrumented by the presence of a bishop's seat. Alesina et al. (2013) report an effect of historical plow use on gender disparities today, using soil characteristics as exogenous instrument for the adoption of plow technology. Such "pure persistence studies" (Bisin & Moro, 2021) do not include the last block in Fig. 41.4, but focus on estimating the first part in its own right.

41.3.2 From Countries to Grids

A second trend in the literature has been a gradual move to lower levels of analysis. Early studies focused on cross-sectional research designs taking countries as units of analysis. Research designs were based on the assumption that country borders represent institutional boundaries. This approach was mainly informed by data availability. Yet, the assumption is also theoretically plausible, since states are primarily responsible for designing regulations and enforcing rules.

There are limitations to the strategy, however. First, states are limited in number and not equally distributed over the world, resulting in relatively low sample sizes and potential overrepresentation of regions such as Europe. Second, even if states represent institutionally homogenous units, the variables explaining institutions do not necessarily cluster along the same lines. Tsetse fly prevalence, for example, is not likely to be strongly correlated with country borders (Alsan, 2015). The historical explanatory variables used in many studies of institutional determinants capture mechanisms occurring before current borders took shape.

The link between institutions and the state is even more tenuous for informal institutions and culture. Even among European nation-states, different regions hold different cultural communities, within and across country borders. In Sub-Saharan Africa, the relative weakness of states implies that both the mismatch between cultural and state borders is stronger, and the relevance of cultural units is larger. Mismatches between cultural and state borders have even been proposed as an explanation for institutional development in the African context (Michalopoulos & Papaioannou, 2016).

For these reasons, research has increasingly moved away from cross-country comparisons to smaller units. First, contributions did so to be able to isolate cultural, informal forces. Tabellini (2010) estimates the effects of cultural norms and values on economic development across regions of Europe, instrumenting culture by

historical events. The attraction of this design is that it allows a clear distinction between formal (varying between countries) and informal institutions (varying by region). Michalopoulos and Papaioannou (2014) use this logic to highlight the relative irrelevance of formal institutions compared to informal institutions in Sub-Saharan Africa, where ethnic partitioning has been the norm. Within-country comparisons of individuals also allow researchers to investigate the persistence of cultural norms, values, and cognitions. Fernández and Fogli (2009) introduce an epidemiological approach to show that second-generation female migrants in the United States portray fertility patterns that correspond to the attitudes in their country of parental origin.

The attraction of lower level units of analysis is that it allows researchers to trace the impacts of geographical, cultural, or historical determinants that do not align along state borders. The question arising, however, is what appropriate units of analysis are, if not countries. Institutions and culture are about structural patterns of behavior. Although they might be observed in individuals, they by definition exist only at the level of groups or communities (Hofstede, 2001; Maseland & van Hoorn, 2017). The tension between the borders of the relevant institutional communities and the geography of its determinants needs to be addressed in any research into the determinants of institutions. Tracing Tsetse fly prevalence to ethnic communities, for example, requires the essential step of mapping geographical variation in prevalence onto ethnic communities (Alsan, 2015).

The identification of relevant institutional units is crucial. What is the historical relevance of the ethnic communities used by Alsan (2015) or Michalopoulos and Papaioannou (2014) over time? Have boundaries between them been stable? Have alternative demarcations been dominant at different points in time, or even simultaneously? Why is this particular delineation of ethnic communities singled out as the relevant one? These questions require careful, historically informed consideration before deciding upon a research design.

Even when having selected a relevant institutional unit of analysis, the challenge is to measure institutional and economic outcomes at this level. Institutional data at below state levels is scarce. Among advanced economies, the European Quality of Governance Index provides information about corruption and the quality of public administration at the regional level for European countries (Charron et al., 2022). Outside Europe, researchers typically draw upon survey data (Mitton, 2016) or anthropological sources (Murdock, 1967). Measuring economic outcomes at sub-country level is equally challenging. Regional statistics are often absent or unreliable, in particular outside advanced economies. Sub-national region-level data on human development, gender disparity, and wealth do exist (Global Data Lab, 2023), but have not been used frequently by economists. Problems of data limitations are especially pertinent when the institutionally relevant sub-unit is not an existing administrative one. Hence, economists have been interested in indicators of economic activity that can be mapped onto any institutional unit. A popular choice is using satellite data capturing nighttime lights intensity as a proxy for economic activity (Chen & Nordhaus, 2011). Although a relatively crude proxy, especially when considering changes over time, this measure allows researchers to track differences in economic performance in locations where reliable data do not exist.

41.3.3 From Cross-Sections to Case Studies

A third development in the empirical strategies of research into institutional determinants is the shift from cross-sectional designs to historical case studies. The use of analytical narratives has a long history in new institutional economics (Greif, 1993; Morgan & Stapleford, 2023; North, 1990). In this approach, researchers apply theoretical principles to a specific case in order to illustrate and evaluate the strength of parsimonious causal mechanisms, and to be able to bring down complex cases to a small number of key variables (Levi & Weingast, 2016). Often, this involves the development of formal theoretical models. Greif (1993), for instance, develops a relatively simple game-theoretic model to explain how commitment problems were solved in long-distance trade, focusing on the case of medieval Maghribi traders. A related use of cases is the application of institutional economic principles to historical events in order to illustrate a general argument, as in North (1990). Such an approach does not necessarily involve formal theorizing nor systematic analysis of data, but involves a theoretical interpretation of a case that supports the argument.

While early new institutional economics made generous use of cases in these ways, the approach gave way to more systematic econometric approaches with the merger between the new institutionalist and the empirics of economic growth literatures. Econometric identification of causal mechanisms became the norm. Apart from the instrumental variable approaches discussed above, case studies of natural experiments gained popularity. The idea in such natural experiments is that a causal variable varies between observations because of some political, natural, or otherwise plausibly exogenous event. Due to the event's exogeneity, it can be argued that the treatment is assigned as-if randomly over cases, allowing causal inference. This assumption of exogeneity can be assessed in a number of ways. The first way is to check whether treatment and control groups are balanced in terms of potentially relevant characteristics. Alternatively, one can run placebo tests to see whether the used treatment variable offers a better explanation than alternative assignments over groups. Tests to check the sensitivity of results to unobserved confounders are also available (Cinelli & Hazlett, 2020).

An advanced approach to natural experiments is the use of *regression discontinuity designs*. In the field of institutional economics, Dell (2010) pioneered the approach in her study of the persistent impact of the *mita* (a system of forced labor for the indigenous population) in colonial Peru and Bolivia. In order to show that the *mita* system has left an imprint on economic outcomes today, Dell (2010) maps the historical geographic boundaries of the *mita* catchment area onto current populations. She then zooms in on those areas within a certain range of the former *mita* border. While populations in and outside the *mita* catchment area may differ in all kinds of aspects, segments just in or outside the area are likely to be similar in characteristics, allowing causal inference. In this context, Dell (2010) proceeds to estimate outcomes in household consumption and prevalence of stunted growth as a function of geographic location and a binary variable reflecting on which side of the historical *mita* border the observation lies. Doing so, she finds that there is a jump in household consumption and stunted growth prevalence at the former border.

Dell's (2010) approach is illustrative of the general logic in RDD designs. Starting point is the identification of a potential historical treatment that had clear geographic boundaries. The researcher then models an outcome of interest c_{id} as a smooth function of geographic location $_d$ (with id denoting observation i located in district d). The idea is to observe whether, controlled for this locational polynomial, there is a jump in outcomes at the border of the historical treatment area. To test this, a binary variable T_d is added indicating whether an observation is above or below the critical threshold of the location polynomial, i.e., the historical border. The resulting basic regression is:

$$c_{id} = \alpha + \gamma T_d + f(\text{geographic location}_d) + \varepsilon_{id} \quad (41.3)$$

This regression is estimated only for those segments of the population close to the former border, under the assumption that individuals located just outside of the historical treatment area are an appropriate counterfactual for those located just inside it. To make this more plausible, controls are typically added to the basic set-up in Eq. 41.3.

Modeling the polynomial convincingly is crucial. The simplest way to do so is using a polynomial capturing distance to the border. Dell (2010) adopts a two-dimensional RDD that employs polynomials for both latitude and longitude. Also, polynomials may be estimated parametrically or non-parametrically. Parametric approaches come with the downside that assuming a specific functional form may cause the researcher to incorrectly interpret any non-linearity unaccounted for as a discontinuity. Non-parametric designs, on the other hand, lay a heavy burden on the data needed, especially in geographic RDD designs. Experimenting with different forms is typically recommended, as well as varying the bandwidth around the border.

Like instrumental variable approaches, case studies of natural experiments provide credible claims about causality. In both approaches, however, a risk is that the strong focus on identification comes at the expense of a deeper understanding of the historical processes involved. Perhaps it is for this reason that qualitative, historical narratives have made something of a comeback in recent times (Weingast et al., 2020). In various recent publications, Acemoglu and Robinson (2019, 2021, 2023) make use of historical case studies to illustrate and support arguments about institutional and cultural evolution. Bisin et al. (2023) develop an analytical narrative to explain how the joint evolution of culture and political institutions resulted in divergence between the Middle East and the West since the early modern era. The attraction of such qualitative approaches is that they are better able to capture the full complexity of interdependent historical processes than empirical models focusing on single explanations (Skarbek, 2020).

Narrative case-study approaches do come with challenges of their own. Unsystematically applied, they may easily result in cherry-picking; presenting selective historical examples that support a pre-existing interpretation. A teleological reading of history, in which current outcomes end up as the inevitable product of unidirectional historical processes, is an equally serious risk. Remedies for these problems have been discussed outside the economic discipline (Flyvbjerg, 2011; Piekkari & Welch, 2018). More awareness of such debates and more training in the use case studies to develop theory is recommended.

41.4 Critical Challenges

The previous section has discussed the main empirical approaches used in the study of institutional determinants. This section will discuss what are the most important methodological challenges for this literature.

41.4.1 Operationalizing Institutional Diversity

First, comparative research requires equivalent measures. In studies of institutional diversity, this is not straightforward. If societies vary in their “rules of the game” (North, 1990), developing a common yardstick to evaluate diverse rules is a challenge. How to reduce the wide array of potential property rights regimes to a single measure?

The common answer in the literature is not to use institutions themselves as outcomes, but the economic function institutions are theorized to have, which is facilitating functioning markets (North, 1990). Measures of institutional quality have been developed to capture the degree to which various institutional arrangements deliver on this (Kaufmann & Kraay, 2008). For instance, qualitative differences in property rights regimes are brought down to a quantitative measure of the degree of security of individual property rights (Acemoglu et al., 2001).

A downside of this approach is that describing whole institutional systems in terms of the extent to which they enable market exchange reduces societies to what Harootunian (2005) calls a *comparative existence*. Societies are described solely in terms of the extent to which they deviate from the functioning market economy benchmark. While facilitating comparison, such a negative empirical approach prevents us from understanding what is actually driving the institutional setting in many contexts. Various authors have commented on the fact that the resulting concept of institutional quality is biased toward Western market economies (Chang, 2011; Gledhill, 2018).

Another issue is the practice of using society-level scores of institutional quality. Using single scores for entire societies neglects the contested nature of institutions. An important body of institutional research in organizational studies has argued that societies are inter-institutional systems, in which individual agents draw from and negotiate between a multitude of conflicting institutional logics (Thornton & Ocasio, 2008). There is never one, unambiguous and uncontested institutional system in any society.

Aligning the methodological demands of rigorous comparative research with the complexity of qualitative institutional diversity is a major challenge that perhaps cannot be met in its entirety. Nevertheless, a few lessons can be derived from the discussion here. First, empirical approaches studying deep determinants should operationalize institutional functioning at the level of specific outcomes rather than use broad measures of institutional quality. Alesina et al. (2013), focusing on a gender inequality in labor markets, is in that respect superior to studies trying to capture some kind of general institutional performance at the societal level. A more

specific approach goes some way to remedy both the problems of unit of analysis-conflation and comparative existence. Second, operationalizations of institutional performance should aim to consider the dynamic and contested nature of institutions. A promising starting point for this is Acemoglu and Robinson (2021), who, echoing longstanding anthropological arguments (e.g., Wolf, 2001), theorize culture as a “repertoire” agents can draw from, allowing multiple cultural and institutional configurations.

41.4.2 Independence of Observations

Inferences from comparative research require sufficient numbers of independent observations. While this requirement is not unique to studies of institutional determinants, there are reasons why this requirement is particularly challenging in this area.

Kelly (2019) argues that in cross-sectional studies of institutional persistence, spatial dependence is a neglected concern. Places that are closer together are typically more alike (Tobler, 1970). Hence, the spatial distribution of any variable likely portrays a similar pattern. As a result, variables that are not connected through any intertemporal causal mechanism may still end up as highly correlated. To remedy such problems, Kelly (2019) recommends reporting Moran’s I Statistic and running horse races with spatial noise replacing actual observations. While recognizing the relevance of spatial dependence, others have accused Kelly (2019) of overstating the problem (Voth, 2021). They argue that spatial dependence is a well-known problem not unique to persistence studies, for which commonly used procedures such as geographic fixed effects or Conley standard errors are sufficient (Voth, 2021; Cirone & Pepinsky, 2022). Furthermore, it has to be recognized that spatial dependence is not so much about geography itself as about social-cultural and economic connections between locations for which geography is a proxy at a given point in time. For spatial dependence to play a role in persistence studies, the structure of these connections between the observed data-points itself needs to be persistent over time (e.g., it requires that Madrid is to Paris now what Madrid was to Paris in 1500). Such continuity of spatial socio-economic structure is precisely what the persistence argument is about. Hence, the presence of spatial autocorrelation does not necessarily undermine claims of persistence.

The issue of independence of observations is not confined to cross-sectional regressions, however. It is a perhaps even more serious concern in case studies exploiting natural experiments. While methodological questions have concentrated on the assumption of exogeneity of events inducing treatment in this approach, an equally crucial element of the argument is the link between historical event and the causal variable for which it proxies. Exogenous group membership, in and of itself, only allows us to identify a causal effect of group membership on outcomes. It does not allow us tell what it is about group membership that causes these effects. Only in cases where historical events unambiguously proxy for the treatment variable but for nothing else is causal inference beyond dispute.

To illustrate this point, consider Lowes et al. (2017) paper about the institutional legacy of the Kuba Kingdom. The paper argues that in the seventeenth century a ruler set up a kingdom in part of an African region that was previously homogenous. This Kuba Kingdom was characterized by relatively good institutions. As corollary, the people outside the kingdom's borders had worse institutions. Assuming that the establishment of the Kuba Kingdom acted as a random shock producing exogenous variation, Lowes et al. (2017) show that exposure to the Kuba kingdom has a persistent effect on culture by revealing consistent differences in norms between people from Kuba and non-Kuba descent.

The paper's sample of 499 individuals suffices to show significant differences between Kuba and non-Kuba descendants. However, the claim in the paper goes beyond that. Lowes et al. (2017) argue that the difference between Kuba and non-Kuba behavior is a result of differences in institutional quality in the seventeenth century. This second claim is not derived from 499 observations, but is based on a comparison between two groups (Kuba and non-Kuba). Two independent observations at the group level are not sufficient to show that the Kuba/non-Kuba difference is attributable to historical institutional quality. This part of the argument depends on deep historical knowledge, showing how historical institutions have evolved to affect current outcomes. In the end, it is not an econometric issue, but one based on qualitative arguments (Cirone & Pepinsky, 2022).

41.4.3 Data Quality

The search for determinants of institutional quality has led economists toward historical, anthropological, sociological, or even genetic data not originally created for the purpose of comparative economic research. Due to academic specialization, economists typically lack the training required to fully understand this data within its research context, and recognize the debates and limitations around it. Data is often simply taken as given. For instance, Ashraf and Galor's (2013) paper on the role of genetic diversity in development does not include any discussion of the controversies and limitations of the genetic diversity measure they exploit to support the "Out of Africa" hypothesis. Obviously, that does not mean that the thesis is necessarily incorrect. However, providing the reader with more information about these debates is essential for the interpretation of results.

A particularly prominent example is the widespread use of Murdock's (1967) *Ethnographic Atlas* in economics. Many authors have used this source without recognizing its controversies (Alesina et al., 2013; Alsan, 2015; Gennaioli & Rainer, 2007; Giuliano & Nunn, 2021; Michalopoulos & Papaioannou, 2014, 2016; Nunn & Wantchekon, 2011). Giuliano and Matranga's (2021) user guide for using historical data is telling. The entire section on ethnographic data is devoted to the *Ethnographic Atlas* alone, without containing any discussion of the validity of Murdock's approach.

It pays to question why the *Ethnographic Atlas* is one of the very few sources of comparative anthropological data in the first place. One reason is that many anthropologists have serious doubts about the possibility of creating comparative measures

of culture at all. A second reason is that even if one does accept the possibility in theory, the practical challenges of deriving comparative data out of a highly diverse set of sources from different disciplines, traditions, methodologies, and points in time are daunting. Historians, anthropologists, and linguists have accused Murdock of seriously misrepresenting evidence, cherry-picking, using unreliable secondary sources, employing questionable categories, and committing factual errors (Davenport, 1988; Fage, 1961; Gleason Jr., 1960; Köbben, 1970; Weiner, 1982).

Beyond the source itself, its reception among economists is problematic. First, economists typically interpret *Ethnographic Atlas* data as a description of groups “prior to industrialization or European contact” (Giuliano & Matranga, 2021). This reading is incorrect and informed by essentialist notions of non-Western culture; as if, were it not for the impact of Western modernity, societies are completely static. In reality, identities are always contested; cultures never exist in isolation; boundaries shift; ethnic groups are not primordial units exhibiting specific cultural traits and staying constant over time (Barth, 1996). Ethnographic data provides snapshots at highly diverse moments in history, not descriptions of cultures fixed in time.

Second, the popularity of the *Ethnographic Atlas* among economists creates problems of its own. Over-use of any material causes any small errors within the data to be reproduced over and over again in different studies. As a result, the self-correcting nature of the scientific process goes lost. A wider use of sources is vital. Lowes (2021) provides a helpful overview of alternative and additional sources that may be considered.

The concerns discussed here are not confined to data borrowed from ethnography. Jerven (2011) makes similar points regarding the use of historic data by economics. A more general recommendation is therefore that economists using data collected in other disciplines add at least one member of this discipline to the research team, in order to provide the necessary context to the data.

41.4.4 Compression of History

None of the challenges we discussed so far are unique to studies of deep determinants of institutions, even if they may be particularly relevant there. A concern more specific about this literature is the way in which it infers insights about historical processes from data covering only a few points in time. In a typical set-up, current outcome variables are regressed on a (proxy for a) historical cause, driven by some historical instrument (Fig. 41.4). Findings of significant relations between these are considered to provide evidence of institutional persistence.

This approach has been criticized for “compressing history” (Austin, 2008). The argument is that showing the impact of historical determinants on current institutions requires more than just correlations between determinant and outcome. Without tracing how deep determinants created this impact over time, we cannot plausibly link determinants to outcomes (Abad & Maurer, 2021; Maseland, 2021).

This is both an empirical and theoretical issue. Empirically, decompressing history requires the analysis of periods in between historical causes and current

effects. Given the data constraints we discussed above, this is a major challenge. Yet, an increasing number of studies consider outcomes in intermediate periods, even if this is often limited to a few moments in time (Acemoglu et al., 2001; Nunn, 2008; Gorodnichenko & Roland, 2017). The results of Maseland (2021) show the relevance of this strategy. Replicating a large set of deep determinants studies while varying the time of measurement of outcome variables reveals that results are often not consistent over time. It matters when we measure.

On the theoretical side, authors of studies of determinants of institutions need to be clear about the mechanisms allegedly driving institutional persistence. Many authors do so. Olsson and Hibbs (2005), for instance, argue that an initial head start in economic development tends to be reproduced due to the endogeneity of economic growth, causing some societies to reach a phase of modern economic growth earlier than others, after which divergence in economic growth may be expected. Acemoglu et al. (2001, 2005b, 2021) explain persistence through interdependent economic and political institutions, giving rise to multiple equilibria. Unfortunately, there is also a sizable proportion of studies that are either silent or vague on what the theorized mechanism for persistence is. In their studies of the impacts of slavery and colonialism on institutional development, Nunn (2008) and Gennaioli and Rainer (2007) do not explain reasons for persistence at all. Spolaore and Wacziarg (2013) allude to various forms of intergenerational transmission, but do not address the reasons why institutions are transmitted intergenerationally.

The real challenge, however, lies in linking theory and empirics. Elaborate theoretical models that are not tested empirically are mere conjecture. From the empirical side, if assessing intermediate periods is only about repeating the set-up in which outcomes are regressed on historical causes for more moments in time, we do not learn much either. Successful assessment of mechanisms requires the translation of alternative mechanisms into alternative hypotheses that can be tested with intermediate data. Doing so is vital. As long as we do not address *how* historical determinants have come to affect current institutions, the deep determinants literature will not transcend the level of intriguing correlations.

41.4.5 Reading History Backwards

Research into institutional determinants seeks to explain contemporary outcomes by tracing these back to historical shocks. The risk of this approach is that history is read backward, with current outcomes being presented as the inevitable product of historical processes.

Consider Bisin et al.'s (2023) explanation of the long-run divergence between the Middle East and the West. Comparing the Middle East to the West, the authors employ a juxtaposition of entities that is the result of the uneven historical process they study, rather than an input. The idea of a uniform entity called the Middle East that develops in contrast to The West reproduces Us-Them divisions that are informed by the growing political economic power of European societies (Said, 1978). Resulting explanations in terms of secular versus religious equilibria in Bisin

et al. (2023) may be criticized for casting the account of divergence in terms that describe why the Middle East has not evolved like “Us” today, rather than describing its evolution in its own terms. Similar points may be raised about Gorodnichenko and Roland’s (2017, 2021) interpretation of individualist vs. collectivist societies, which echoes Orientalist notions of Eastern societies lacking dynamism and being inclined toward despotism (Said, 1978).

The problem with reading history backward is that we become oblivious to all the other scenarios that could alternatively have unfolded. Medieval city-state independence is associated with the development of stronger civic norms (Guiso et al., 2016). But could it also have led to more civil strife and a resulting erosion of social capital? Similarly, could resistance to more extractive colonial regimes alternatively have caused earlier de-colonization and eventually better protection of property rights? Without being open to such alternative scenarios, we risk seeing history as a teleological process inevitably leading to the existing outcomes.

This risk is not confined to persistence studies, but perhaps especially relevant here. The objective of studies into deep determinants of institutions is to establish the causal effects of historical events on contemporary outcomes. This reasoning contains two elements. First, an object (a tribe, country, region) is subject to a causal effect leading it to acquire a certain characteristic. Second, the object maintains this resulting characteristic throughout time. Persistent characteristics thus apply to persistent objects.

But which persistent objects? Authors have selected countries (Acemoglu et al., 2001), populations (Putterman & Weil, 2010), or ethnic communities (Alsan, 2015). But to what extent are these units the same as the past units that have received treatment? Researchers face the problem of what philosophers call *temporary intrinsics*, the incompatibility of the ideas that objects are defined by intrinsic characteristics and that characteristics of objects change over time (Miller, 2018). An object at time $t-1$ is only the same object as an object at time t if it has the same intrinsic properties. However, that means that the persistence of properties is implied by the persistence of the object; or, turning this around, the persistence of defining characteristics is what allows us to trace the object through time in the first place. Any claims of persistent characteristics are simply tautological if the characteristics that define the persistence of the unit are the same characteristics that the researcher claims to be persistent.

This is more than a philosophical issue. Consider Acemoglu et al. (2001), which traces former colonies’ institutions back to conditions during colonization. If we consider modern Nigeria (the unit for which institutional outcomes are measured) to be the heir of Lagos Colony (the unit receiving the historical treatment), we do so on basis of some institutional continuity between the two. However, if institutional continuity is what defines the connection between contemporary Nigeria and historical Lagos, the claim that colonies portray institutional persistence is assumed rather than observed. Similarly, if Putterman and Weil’s (2010) migratory index connects current populations to past ones through a process of intergenerational cultural transmission, the finding that past events matter for current outcomes because of cultural transmission is an identifying assumption rather than an analytical result.

How to deal with these issues? One way forward is to pay more attention to instances where persistence might be expected but is absent. Gingerich and Vogler (2023) offer such an analysis. Comparing the divergent effects of disruptions of local self-government in Poland and Brazil, they show that differences in duration and the scope of repression explain why similar historical shocks led to persistent effects in Poland but not in Brazil.

More fundamentally, we should address the process of creating history itself. History is written in the present by agents interpreting the past, creating a consistent narrative linking past events to present outcomes. Rather than seeking to avoid present bias, researchers may include the processes through which the past is brought into current ideas and behaviors into their research. A fascinating example of a study doing so is Ochsner and Roesel's (2024) analysis of the effects of sixteenth-century Ottoman rule on contemporary voting patterns in Austria. The study shows that voting for far-right parties is significantly more prevalent in areas previously under occupation by Ottomans. Yet, they also find that this pattern is a novel one, only emerging once the far-right FPÖ started to campaign against Turkish accession of the European Union, building on an existing collective commemoration of Ottoman occupation in the form of street names and monuments. Contemporary outcomes are thus directly linked to past events, but this connection only exists in and because of the active, collective recall of these events in the present, uniting current populations and historical victims of Ottoman oppression in a single "imagined community" (Anderson, 1983).

Contemporary outcomes are not inevitable products of a given chain of historical events. Persistence is present because history is actively invoked in current decision-making. An understanding of the process of (re-)production of effects of historical determinants in the present is at least as relevant as the identification of initial causes.

41.5 Concluding Thoughts

The combination of New Institutional Economic theory developed in the 1990s (Greif, 1993; North, 1990) with the econometric approaches associated with the "credibility revolution" in economics (Angrist & Pischke, 2010) has proven a fertile breeding ground for empirical studies explaining divergent institutional development. The resulting literature has produced a long list of historical, geographical, and cultural determinants of institutions. Two decades after the publication of the paper arguably kickstarting this literature (Acemoglu et al., 2001), it is time to take stock of what has been achieved. What have we learned so far?

The first and most important result of the literature on institutional determinants is the insight that institutions matter. Papers tracing institutions back to historical or geographical determinants initially did so to provide a source of exogenous variation in institutions, allowing credible inference about institutional effects on economic outcomes. As a result of these contributions, the insight that institutions matter has become widely accepted. Beyond this, the objectives of studies into institutional

determinants have gradually shifted from validating the result that institutions matter to shedding light on institutional development itself.

On this account, the literature's contributions are more mixed. What we have learned is that institutions have deep historical, cultural, and geographical roots. That has repercussions for policy. If institutions are embedded in their physical and cultural environment, swapping one institutional system for a better performing one overnight is not going to work (Boettke et al., 2008; Seidler, 2014). Policy-makers need to be aware of the social and geographical context and adapt recommendations accordingly. Yet, it is a fine line between stressing the contextual embeddedness of institutions and the suggestion that differences in institutional and economic performance are contextually pre-determined. A literature doggedly pursuing ever deeper determinants of development may easily give rise to fatalism and, worse, provide unintended legitimization of existing inequalities. Prominent authors are aware of this and argue against it, stressing that knowing more about the origins of institutions can only help to design effective policies targeting them (Nunn, 2020). Even so, this requires studying what accounts for the reproduction of institutional systems and under what conditions persistence does or does not materialize (Gingerich & Vogler, 2023; Maseland, 2021). These are the questions where the literature so far has been least prolific, however.

This chapter has shown that part of the reason for this is methodological. The credibility revolution in economics has emphasized causal identification, allowing the field of empirical institutional economics to bloom. However, the emphasis on identifying “clean” historical events to estimate causal parameters (Cantoni & Yuchtman, 2021) carries the risk of looking for what we can credibly estimate rather than what is important. The focus on causality in terms of initial roots rather than processes of reproduction has produced a literature that links historical variables to current institutional outcomes. Lack of reliable data has caused the part in between, describing the reproduction of institutional systems over time, to be inferred only. Direct analysis of the historical processes connecting past to present is rare. Hence, the charge of compression of history (Austin, 2008). For the empirical literature to go beyond producing lists of intriguing correlates of development, it needs to reconnect with the theoretical arguments about path-dependence, inertia, and reproduction of equilibria. Increasingly sophisticated arguments about institutional evolution in the literature (e.g., Acemoglu et al., 2021) deserve to be taken out of the realm of pure theory and be empirically evaluated. The revival of narrative approaches is to be welcomed in this regard (Weingast et al., 2020). Complementing econometrics with a narrative, case-study-based approach may be a good way to bring the historical process back into empirical research.

More in general, the literature would do well to reflect more thoroughly on the methodological assumptions and biases that go into its analyses. In this chapter, we have highlighted a few of such methodological concerns: the backward reading of history, the reduction of societies to a comparative existence, essentialism, the compression of history. These are all issues that have been thoroughly discussed in other disciplines, but are not prominent in the economic literature. Economics' increasing interest in data from other disciplines requires this to change. The idea

that one can bring in data from other fields without bringing in the methodological debates surrounding that data is a mirage. Studies of institutional determinants have led economists deep into the areas of other fields of science. It is time to take what these disciplines have to say about this data more seriously as well.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute.

References

- Abad, L. A., & Maurer, N. (2021). History never really says goodbye: A critical review of the persistence literature. *Journal of Historical Political Economy*, 1(1), 31–68.
- Acemoglu, D., & Robinson, J. A. (2019). *The narrow corridor: How nations struggle for liberty*. Penguin UK.
- Acemoglu, D., & Robinson, J. A. (2021). *Culture, institutions and social equilibria: A framework* (Working Paper No. W28832). National Bureau of Economic Research.
- Acemoglu, D., & Robinson, J. A. (2023). Weak, despotic, or inclusive? How state type emerges from state versus civil society competition. *American Political Science Review*, 117(2), 407–420.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. *American Economic Review*, 91(5), 1369–1401.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2005a). The rise of Europe: Atlantic trade, institutional change, and economic growth. *American Economic Review*, 95(3), 546–579.
- Acemoglu, D., Johnson, S., & Robinson, J. A. (2005b). Institutions as a fundamental cause of long-run growth. In *Handbook of economic growth 1* (pp. 385–472). Elsevier.
- Acemoglu, D., Egorov, G., & Sonin, K. (2021). Institutional change and institutional persistence. In *Handbook of historical economics* (pp. 365–389). Academic Press.
- Alatas, S. H. (1977). *The myth of the lazy native*. Frank Cass & Co.
- Alesina, A., Giuliano, P., & Nunn, N. (2013). On the origins of gender roles: Women and the plough. *Quarterly Journal of Economics*, 128(2), 469–530.
- Alsan, M. (2015). The effect of the Tsetse Fly on African development. *American Economic Review*, 105(1), 382–410.
- Anderson, B. (1983). *Imagined communities: Reflections on the origin and spread of nationalism*. Verso.
- Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2), 3–30.
- Ashraf, Q., & Galor, O. (2013). The “out of Africa” hypothesis, human genetic diversity, and comparative economic development. *American Economic Review*, 103(1), 1–46.
- Austin, G. (2008). The ‘reversal of fortune’ thesis and the compression of history: Perspectives from African and comparative economic history. *Journal of International Development*, 20(8), 996–1027.
- Barro, R. J. (1991). Economic growth in a cross section of countries. *Quarterly Journal of Economics*, 106(2), 407–443.
- Barth, F. (1996). *Ethnic groups and boundaries*. George Allen & Unwin.
- Becker, S., Boeckh, K., Hainz, C., & Woessmann, L. (2016). The empire is dead, long live the empire! Long-run persistence of trust and corruption in the bureaucracy. *Economic Journal*, 126(590), 40–74.
- Beugelsdijk, S., & Maseland, R. (2010). *Culture in economics*. Cambridge University Press.
- Bisin, A., & Moro, A. (2021). LATE for history. In *Handbook of historical economics* (pp. 269–296). Academic Press.

- Bisin, A., & Verdier, T. (2023). Advances in the economic theory of cultural transmission. *Annual Review of Economics*, 15, 63.
- Bisin, A., Rubin, J., Seror, A., & Verdier, T. (2023). Culture, institutions and the long divergence. *Journal of Economic Growth*, 29, 1–40.
- Boettke, P. J., Coyne, C. J., & Leeson, P. T. (2008). Institutional stickiness and the new development economics. *American Journal of Economics and Sociology*, 67(2), 331–358.
- Cantoni, D., & Yuchtman, N. (2021). Historical natural experiments: Bridging economics and economic history. In *Handbook of historical economics* (pp. 213–241). Academic Press.
- Chang, H.-J. (2011). Institutions and economic development: Theory, policy and history. *Journal of Institutional Economics*, 7(4), 473–498.
- Charron, N., Lapuente, V., Bauhr, M., & Annoni, P. (2022). Change and continuity in quality of government: Trends in subnational quality of government in EU member states. *Journal of Regional Research*, 53, 5–23.
- Chen, X., & Nordhaus, W. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21), 8589–8594.
- Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of The Royal Statistical Society Series B*, 82(1), 39–67.
- Cirone, A., & Pepinsky, T. B. (2022). Historical persistence. *Annual Review of Political Science*, 25, 241–259.
- Davenport, W. H. (1988). George Peter Murdock's classification of “consanguineal kin groups”. *Behavior Science Research*, 22(1–4), 10–22.
- Dell, M. (2010). The persistent effects of Peru's mining mita. *Econometrica*, 78(6), 1863–1903.
- Easterly, W., & Levine, R. (2003). Tropics, germs, and crops: How endowments influence economic development. *Journal of Monetary Economics*, 50, 3–39.
- Easterly, W., & Levine, R. (2016). The European origins of economic development. *Journal of Economic Growth*, 21, 225–257.
- Engerman, S. L., & Sokoloff, K. L. (1997). Factor endowments, institutions, and differential paths of growth among new world economies. In S. Haber (Ed.), *How Latin America fell behind* (pp. 260–304). Stanford University Press.
- Fage, J. D. (1961). Anthropology, botany, and the history of Africa. *Journal of African History*, 2(2), 299–309.
- Fernández, R., & Fogli, A. (2009). Culture: An empirical investigation of beliefs, work, and fertility. *American Economic Journal: Macroeconomics*, 1(1), 146–177.
- Feyrer, J., & Sacerdote, B. (2009). Colonialism and modern income: Islands as natural experiments. *Review of Economics and Statistics*, 91(2), 245–262.
- Fisman, R., & Miguel, E. (2007). Corruption, norms, and legal enforcement: Evidence from diplomatic parking tickets. *Journal of Political Economy*, 115(6), 1020–1048.
- Flyvbjerg, B. (2011). Case study. In *Sage handbook of qualitative research* (pp. 301–316). Sage.
- Gennaioli, N., & Rainer, I. (2007). The modern impact of precolonial centralization in Africa. *Journal of Economic Growth*, 12, 185–234.
- Gingerich, D., & Vogler, J. P. (2023, June 6). *When are junctures critical? The legacies and non-legacies of interruptions in local self-government*. <https://doi.org/10.2139/ssrn.4471223>
- Giuliano, P., & Matrangola, A. (2021). Historical data: Where to find them, how to use them. In *Handbook of historical economics* (pp. 95–123). Academic Press.
- Giuliano, P., & Nunn, N. (2021). Understanding cultural persistence and change. *Review of Economic Studies*, 88(4), 1541–1581.
- Gleason, H. A., Jr. (1960). Africa: Its peoples and their culture history. *Language*, 36(3), 470–475.
- Gledhill, J. (2018). Governance issues in development. In H. Callan (Ed.), *International encyclopedia of anthropology* (pp. 1–9). Wiley.
- Global Data Lab. (2023). *Subnational human development*. <http://www.globaldatalab.org>
- Gorodnichenko, Y., & Roland, G. (2017). Culture, institutions, and the wealth of nations. *Review of Economics and Statistics*, 99(3), 402–416.

- Gorodnichenko, Y., & Roland, G. (2021). Culture, institutions and democratization. *Public Choice*, 187, 165–195.
- Greif, A. (1993). Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *American Economic Review*, 83, 525–548.
- Guiso, L., Sapienza, P., & Zingales, L. (2016). Long-term persistence. *Journal of the European Economic Association*, 14(6), 1401–1436.
- Hall, P. & Soskice, D. (eds) (2001). *Varieties of capitalism: The Institutional Foundations of Comparative Advantage*. New York: Oxford University Press.
- Harootunian, H. (2005). Some thoughts on comparability and the space-time problem. *Boundary*, 32(2), 23–52.
- Harrison, L. (2000). *Underdevelopment is a state of mind: The Latin American case*. Madison Books.
- Henrich, J., et al. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(6), 795–815.
- Hibbs, D., Jr., & Olsson, O. (2004). Geography, biogeography, and why some countries are rich and others are poor. *Proceedings of the National Academy of Sciences*, 101(10), 3715–3720.
- Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage.
- Inglehart, R., et al. (2020). *World values survey: All rounds – Country-pooled datafile*. JD Systems Institute/WVSA Secretariat.
- Jerven, M. (2011). A clash of disciplines? Economists and historians approaching the African past. *Economic History of Developing Regions*, 26(2), 111–124.
- Kaufmann, D., & Kraay, A. (2008). Governance indicators: Where are we, where should we be going? *The World Bank Research Observer*, 23(1), 1–30.
- Kelly, M. (2019). *The standard errors of persistence* (CEPR Discussion Paper No. DP13783). CEPR.
- Köbben, A. (1970). Ethnographic atlas. *Bijdragen Tot De Taal-, Land- En Volkenkunde*, 126(1), 160–161.
- LaPorta, R., Lopez-De-Silanes, F., Shleifer, A., & Vishny, R. (1999). The quality of government. *Journal of Law, Economics & Organisation*, 15(1), 222–279.
- Levi, M., & Weingast, B. R. (2016, September 6). *Analytic narratives, case studies, and development*. <https://doi.org/10.2139/SSrn.2835704>
- Link, A. (2023, March 13). *Beasts of burden, trade, and hierarchy: The long shadow of domestication*. <https://ssrn.com/Abstract=4386758>
- Lowes, S. (2021). Ethnographic and field data in historical economics. In *The handbook of historical economics* (pp. 147–177). Academic.
- Lowes, S., Nunn, N., Robinson, J., & Weigel, J. (2017). The evolution of culture and institutions: Evidence from the Kuba kingdom. *Econometrica*, 85(4), 1065–1091.
- Maseland, R. (2013). Parasitical cultures? The cultural origins of institutions and development. *Journal of Economic Growth*, 18, 109–136.
- Maseland, R. (2018). Is colonialism history? *Journal of Institutional Economics*, 14(2), 259–287.
- Maseland, R. (2021). Contingent determinants. *Journal of Development Economics*, 151, 102654.
- Maseland, R., & Van Hoom, A. (2017). Culture at the country level. In H. van Herk & C. J. Torelli (Eds.), *Cross cultural issues in consumer science and consumer psychology*. Springer.
- Mayoral, L., & Olsson, O. (2022). *Floods, droughts, and environmental circumscription in early state development: The case of ancient Egypt*. Mimeo.
- Michalopoulos, S., & Papaioannou, E. (2014). National institutions and subnational development in Africa. *Quarterly Journal of Economics*, 129(1), 151–213.
- Michalopoulos, S., & Papaioannou, E. (2016). The long-run effects of the scramble for Africa. *American Economic Review*, 106(7), 1802–1848.
- Miller, K. (2018). Persistence. In *The Routledge encyclopedia of philosophy*. Taylor and Francis.
- Mitton, T. (2016). The wealth of subnations: Geography, institutions, and within-country development. *Journal of Development Economics*, 118, 88–111.

- Morgan, M. S., & Stapleford, T. A. (2023). Narrative in economics: A new turn on the past. *History of Political Economy*, 55, 395.
- Mokyr, J. (2014). Culture, institutions, and modern growth. In: *Institutions, property rights, and economic growth: The legacy of Douglass North* (pp. 151–191).
- Murdock, G. P. (1967). *Ethnographic atlas*. University of Pittsburgh Press.
- Norris, P., & Inglehart, R. (2002). Islamic culture and democracy: Testing the ‘clash of civilizations’ thesis. *Comparative Sociology*, 1(3), 235–263.
- North, D. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- Nunn, N. (2008). The long-term effects of Africa’s slave trades. *Quarterly Journal of Economics*, 123(1), 139–176.
- Nunn, N. (2020). The historical roots of economic development. *Science*, 367(6485), eaaz9986.
- Nunn, N., & Puga, D. (2012). Ruggedness: The blessing of bad geography in Africa. *Review of Economics and Statistics*, 94(1), 20–36.
- Nunn, N., & Wantchekon, L. (2011). The slave trade and the origins of mistrust in Africa. *American Economic Review*, 101(7), 3221–3252.
- Ochsner, C., & Roesel, F. (2024). Activated history: The case of the Turkish sieges of Vienna. *American Economic Journal: Applied Economics*, 16(3), 76–112.
- Olsson, O., & Hibbs, D. A., Jr. (2005). Biogeography and long-run economic development. *European Economic Review*, 49(4), 909–938.
- Olsson, O., & Paik, C. (2020). A western reversal since the neolithic? The long-run impact of early agriculture. *Journal of Economic History*, 80(1), 100–135.
- Oumer, A., Maseland, R., & Garretsen, H. (2020). Was De Montesquieu (only half) right? *Journal of Economic Behavior & Organization*, 173, 256–269.
- Page, S. E. (2006). Path dependence. *Quarterly Journal of Political Science*, 1(1), 87–115.
- Piekkari, R., & Welch, C. (2018). The case study in management research: Beyond the positivist legacy of Eisenhardt and Yin. In C. Cassel (Ed.), *The Sage handbook of qualitative business and management research methods* (pp. 345–358). Sage.
- Putterman, L., & Weil, D. N. (2010). Post-1500 population flows and the long-run determinants of economic growth and inequality. *Quarterly Journal of Economics*, 125(4), 1627–1682.
- Rathbun, B. C., Powers, K. E., & Anders, T. (2019). Moral Hazard: German public opinion on the Greek debt crisis. *Political Psychology*, 40(3), 523–541.
- Rodrik, D. (2003). Institutions, integration, and geography: In search of the deep determinants of economic growth. In D. Rodrik (Ed.), *Search of prosperity: Analytic country studies on growth*. Princeton University Press.
- Rodrik, D., Subramanian, A., & Trebbi, F. (2004). Institutions rule: The primacy of institutions over geography and integration in economic development. *Journal of Economic Growth*, 9, 131–165.
- Said, E. (1978). *Orientalism*. Pantheon Books.
- Searle, J. R. (2005). What is an institution? *Journal of Institutional Economics*, 1(1), 1–22.
- Seidler, V. (2014). When do institutional transfers work? The relation between institutions, culture and the transplant effect: The case of Borno in North-eastern Nigeria. *Journal of Institutional Economics*, 10(3), 371–397.
- Skarbek, D. (2020). Qualitative research methods for institutional analysis. *Journal of Institutional Economics*, 16(4), 409–422.
- Spolaore, E., & Wacziarg, R. (2013). How deep are the roots of economic development? *Journal of Economic Literature*, 51(2), 325–369.
- Summers, R., & Heston, A. (1991). The Penn world table (mark 5): An expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics*, 106(2), 327–368.
- Tabellini, G. (2010). Culture and institutions: Economic development in the regions of Europe. *Journal of the European Economic Association*, 8(4), 677–716.
- Thornhill, R., Fincher, C. L., & Aran, D. (2009). Parasites, democratization, and the liberalization of values across contemporary countries. *Biological Reviews*, 84, 113–131.

- Thornton, P., & Ocasio, W. (2008). Institutional logics. In R. Greenwood (Ed.), *The Sage handbook of organizational institutionalism* (pp. 99–128). Sage.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–240.
- Tsui, K. K. (2011). More oil, less democracy: Evidence from worldwide crude oil discoveries. *The Economic Journal*, 121(551), 89–115.
- Voth, H.-J. (2021). Persistence—myth and mystery. In *The handbook of historical economics* (pp. 243–267). Academic Press.
- Weiner, J. F. (1982). George Peter Murdock, atlas of world cultures (book review). *The Australian Journal of Anthropology*, 13(5), 461.
- Weingast, B. R., Rosenthal, J.-L., Levi, M., Greif, A., & Bates, R. H. (2020). *Analytic narratives*. Princeton University Press.
- Williamson, O. E. (2000). The new institutional economics: Taking stock, looking ahead. *Journal of Economic Literature*, 38(3), 595–613.
- Wolf, E. (2001). *Pathways of power: Building an anthropology of the modern world*. University of California Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Using Field Experiments to Understand the Impact of Institutions on Economic Growth

42

Omar Al-Ubaydli, Faith Fatchen, and John List

Contents

42.1	Introduction	1118
42.2	Prologue: Running Field Experiments in Institutional Economics	1120
42.2.1	Definitions	1120
42.2.2	Barriers to Running Field Experiments	1123
42.2.3	How Field Experiments Can Contribute to Institutional Economics	1125
42.3	A Survey of Selected Field Experiments in Institutions-Mediated Economic Growth	1126
42.3.1	Legal Systems	1126
42.3.2	Political Systems	1128
42.3.3	Informal Social Systems Including Religion	1131
42.4	Epilogue: Where Do Field Experiments Hold Most Promise for Institutional Economics and Economic Growth?	1134
42.4.1	Ideas Grounded in Laboratory Experiments	1134
42.4.2	Greenfield Field Experiments	1135
42.4.3	Mixed Methods	1137
42.5	Conclusion	1138
	References	1139

Abstract

Field experiments are a useful empirical tool that can be deployed in any sub-discipline—including institutional economics—to enhance the sub-discipline’s empirical insights. However, we here argue that there exist fundamental

O. Al-Ubaydli (✉)

Bahrain Center for Strategic, International and Energy Studies, Studies and Research Department, Riffa, Bahrain

George Mason University, Fairfax, VA, USA

e-mail: oalubaydli@derasat.org.bh; omar@omar.ec

F. Fatchen · J. List

University of Chicago, Chicago, IL, USA

e-mail: ffatchen@uchicago.edu; jlist@uchicago.edu

© The Author(s) 2025

C. Ménard, M. M. Shirley (eds.), *Handbook of New Institutional Economics*, https://doi.org/10.1007/978-3-031-50810-3_42

1117

barriers to the use of field experiments in understanding the impact of institutions on economic growth. Despite these obstacles, we present some significant scholarly contributions that merit exposition, while also proposing some future methods for using field experiments within institutional economics. While field experiments may be limited in answering questions in institutional economics with macroeconomic outcomes, there is great potential in employing field experiments to answer micro-founded questions.

Keywords

Institutional economics · Economic growth · Field experiments

42.1 Introduction

In terms of the questions it seeks to answer, economics as a discipline has developed a high degree of intellectual diversity, spanning important queries such as: what is the optimal exchange rate regime for a country to pursue (Mundell, 1961)? How much redistributive activity should governments undertake (Piketty, 2014)? And what is the nature and extent of discrimination in markets (List, 2004a)? However, an area of constant scrutiny and enlightenment since the late eighteenth century concerns the determinants of the wealth of nations: what factors allow an economy to realize a persistently high rate of growth (Barro, 2003)? Moreover, this inquiry is decidedly motivated by practical considerations that extend well beyond an intellectual desire to know how the world works. In other words, scholars seek knowledge of the underpinnings of economic growth because they wish to use that information to help economies grow and become more prosperous.

Institutional economics is a notable link in a long chain of methodological and topical evolutions within the broader economics discipline. While “institutions” is a somewhat amorphous concept in economics, new institutional economics (NIE) is structured around the notion that institutions, humanly devised rules, and norms—both formal and informal—are critical to understanding economies (Ménard & Shirley, 2022). NIE encompasses a diverse set of ideas, reflected by its most famous contributors including—among others—Douglass North, Oliver Williamson, and Elinor Ostrom.

Built upon the basis that institutions are fundamental, NIE approaches a diverse set of questions, diverging from old institutional economics in both the set of questions and methods leveraged by scholars. Unlike old institutional economics, NIE embraces some of the assumptions of neoclassical economics—subject to augmentation of behavioral assumptions and transaction costs. NIE additionally leverages the tools of econometrics to causally study questions of economic organization and development without ignoring institutional contexts (Ménard & Shirley, 2022). New institutionalists approach questions at many different levels—micro and macro—to understand the role of institutions. It is intellectually adjacent to—and intersects with—public choice economics, which analyzes many similar factors but

with a less ambitious selection of outcome variables. Thus, for example, where a public choice scholar may study the impact of voting rights on proximate tax policy, an NIE scholar wants to know what voting rights imply for long-run economic growth. In seeking to answer this question, they may wish to understand what are the formal or informal norms in place that lead to well-functioning voting systems.

To some degree, the genesis of institutional economics reflects the feeling among its practitioners that classical and neoclassical growth theory have failed to yield satisfactory answers to the core questions that Adam Smith and other founders of modern economics posed. In particular, it has brought to the attention of mainstream scholars the importance of a series of previously over-looked factors, such as legal systems, social norms, power structures, unique historical contexts, and so on.

Despite the rich and ultimately constructive exchanges between institutional economics and other sub-disciplines within economics, a separate literature has advanced contemporaneously and virtually independently: the use of experimental methods in economics. For those unfamiliar with experimental economics, the method can be succinctly summarized as the estimation of causal effects of groups of explanatory variables on dependent variables through the investigating scholar's direct manipulation of the explanatory variables, most frequently (though not definitively) via the use of randomized control. In short, experimental economists control the assignment mechanism, allowing them to answer heretofore quite complex questions using causal inference. This latter feature brings economics in line with the epistemological standards of the natural sciences by helping scholars to overcome the problems of endogeneity caused by relying on observational data (Al-Ubaydli & List, 2013).

Thus, for example, a non-experimental study of the causal effect of the interest rate on the demand for corporate credit would involve gathering data on realized interest rates and demand for corporate loans and inferring the statistical relationship between the two, aided by gathering data on control variables such as the business confidence and profit margins. In contrast, a study controlling the assignment mechanism (an experimental study) would witness the researcher directly and randomly manipulating the interest rate on corporate loans and observing the impact on demand, safe in the knowledge that any extraneous variation in factors that affect the demand for corporate loans is orthogonal to the interest rate variation and therefore does not bias the estimates of the causal effect. In the special case of natural field experiments, the environment in which the experiment takes place is the real, natural setting, and the participants are unaware of the fact that an experiment is under way (Harrison & List, 2004). This enhances the generalizability of the findings through a series of channels, most notably by eliminating the impact of Hawthorne effects, more formally referred to as experimenter demand effects (Levitt & List, 2007) as well as by controlling selection (Al-Ubaydli & List, 2013).

Having offered simple definitions of institutional economics and field experiments, it is clear that there are barriers to combining the two. The causal effects that institutional economics is interested in estimating involve explanatory variables that are generally very difficult for a researcher to manipulate, such as legal systems and social norms. Moreover, the outcome variables imply causal chains that may take

years or decades to materialize, such as the impact of property rights on economic growth. However, it is possible to use field experiments to understand elements of the large puzzle that links institutions to economic growth. Yet, in doing so we caution the reader and note that we must understand the complete set of mechanisms before generalizing the insights from a part of the puzzle to the whole. Indeed, without a keen sense of how the whole puzzle fits together, including mediation and moderation, one should take great care in generalizing and scaling ideas (List, 2024). This is especially important considering the many causes of variation in the outcomes that accompany many questions within the institutional economics corpus. The overarching message is that greater trust should be given to a policy that is well understood theoretically and one that has a simple mediation path. In addition, measuring key surrogates—proxies for long-run outcomes—and establishing relevant surrogacy conditions, can open an entirely new area of institutional economics research that leverages field experiments.

With those goals in mind, in this chapter, we survey a selection of studies, while noting that these ultimately represent watered-down versions of institutional economics. We focus on three main areas: legal systems, political systems, and social systems, where the latter can be thought of as informal social norms that also include religion (since there are only a handful of theocracies in the world at present). Considering the contextual sensitivity of the findings in these studies, and the breadth of inquiry, the literature does not yield a sufficient density of results relating to a certain question that allow for generalizable conclusions. Instead, our survey works more as an exposition of creative ways in which field experiments can be used under the umbrella of growth-centered institutional economics.

Given that our background is in field experiments, we conclude the chapter by presenting a series of ideas of future field experiments that could help scholars estimate the impact of institutions on economic growth. In some cases, our suggestions are based on adapting a previously-executed laboratory experiment to a field setting, while others involve totally novel designs. We hope that this class of open brainstorming can provide our fellow practitioners with inspiration for their own future research.

As a final remark, this chapter is written for people who have a limited background in institutional economics or field experiments, or possibly both. Given our general argument about the distinct nature of these two sub-disciplines, we believe that a large proportion of our colleagues satisfy this criterion.

42.2 Prologue: Running Field Experiments in Institutional Economics

42.2.1 Definitions

Normally, an economics paper would not require a lengthy section on semantics. In the case of this chapter, it is likely that readers lack a background in a least one of the two concepts of field experiments and institutional economics, and a full

understanding of the two is required for our main argument. Readers familiar with the definitions can skip to later sections.

42.2.1.1 Field Experiments

Field experiments are an epistemological method that spans the social sciences, while institutional economics is a combination of (1) a field of research that interests economists and (2) an associated set of epistemological techniques. *Prima facie*, field experiments can be used in institutional economics in the same manner that cross-country regressions can be used to study monetary policy, or surveys can be used to study how much people value protecting the environment. As we demonstrate below, upon careful consideration of the defining characteristics of institutional economics, it becomes evident that field experiments can rarely be applied to the study of institutions in the case of macro-outcomes such as economic growth, but are of use to new institutionalists seeking to understanding the micro-foundations of institution centered questions.

We begin by offering a clear definition of field experiments, based on Harrison and List (2004). In the social sciences, an empirical causal investigation is a process of gathering data to estimate the effect of varying a group of explanatory variables on the value of a dependent variable, where at least some of the explanatory and dependent variables are a form of human behavior. The effect being estimated is known as the casual effect of those explanatory variables on the dependent variable (Heckman, 2005).

In general, there are two types of data: observational and experimental (Al-Ubaydli & List, 2013). In the case of observational data, the data are generated organically and independently of the investigating researcher's efforts. In the case of experimental data, the variation in the explanatory variables is induced by the investigating researcher, who has control over the assignment mechanism; that is, who is placed in treatment and control. In this manner, an experiment usually involves the researcher inducing random variation in the explanatory variables (known as a randomized control trial), but this is not necessarily a defining characteristic; the key is the researcher's control of the explanatory variables, and their exercising that control as part of an effort to acquire knowledge. For example, in many of Vernon Smith's seminal laboratory experiments he controlled the environment but did not use randomization to identify a treatment effect. In this case, he used theory to generate the relevant counterfactual rather than data.

The simplest setting that social scientists can use for running experiments is a conventional laboratory, where the human behavior being studied is that of college students, the decision-making environment is abstract, and the rules governing their behavior are imposed. Conventional laboratory experiments transition into artefactual field experiments when the participating students are replaced by a non-standard subject pool, such as a representative sample of the general population. This then mutates into a framed field experiment when the decision-making is given real context. Critically, in all three of these classes, the participants—students or otherwise—consciously enroll in the study and are aware of the unnaturally high levels of scrutiny of their behavior by the overseeing research team.

The key distinction to be drawn is between these types of experiments and natural field experiments, commonly shortened to “field experiments.” In a natural field experiment, the setting is the natural environment where participants occupy natural roles within the setting, and—most importantly—the participants are unaware of their participation in an experiment and therefore there is no selection into or out of the experiment. While each class of experiment possesses a set of advantages and disadvantages relating to factors such as feasibility, ethicality, and epistemic value, a consensus is emerging that natural field experiments yield data that is more likely to provide generalizable insights. That means conclusions that apply to the real, high-stakes contexts that social scientists are most interested in studying, especially when their goal is to acquire knowledge that can be leveraged to manipulate the environment to the benefit of humans (Al-Ubaydli & List, 2013).

While a full exposition of the reasons for this is beyond the scope of this chapter, it can be concisely stated as follows: enrolling people to take decisions in abstract or unnatural settings coupled with unnaturally high levels of scrutiny can lead to distorted behavior that yields limited insights about behavior in natural settings over the entire population of interest with natural (and therefore low) levels of scrutiny (Levitt & List, 2007). A simple illustration is running an experiment to estimate the impact of an unexpected positive wealth shock on immediate charitable contributions. In a conventional laboratory experiment, people will donate an average of around 40% of their fortuitously acquired wealth, which is a large overstatement of donations in an isomorphic real setting (List, 2007).

The ability of experiments to overcome the problem of endogeneity has made them an increasingly popular tool in social sciences. In the case of using field experiments within institutional economics, we aim to delineate where precisely we see potential applications.

42.2.1.2 New Institutional Economics

The precursor to defining new institutional economics is defining institutions. The most famous definition is provided by Douglass North: “Institutions are the rules of the game in a society or, more formally, are the humanly devised constraints that shape human interaction” (North, 1990). Acemoglu (2008) expands by noting that this definition of institutions implies two things: they are humanly designed (as opposed to geographical phenomena, for example); and the constraints they impose on human behavior operate through incentives. Some of the core institutions that economists study include property rights, which fall under the broader umbrella of legal systems. Other formal institutions include markets, auction houses, online brokerages, and other systems for exchanging goods and services. One can add to these informal institutions, such as social norms.

This brings us to the definition of NIE. As alluded to in the introduction, institutional economists are united on the fundamental statement that institutions are critical to understand economies and societies (Ménard & Shirley, 2022). However, the approach and methods differ in important ways from what is referred to as old institutional economics. Both centered around the importance of institutions, but NIE leverages the

theoretical and analytical tools of neoclassical economics. With such tools at their disposal, NIE scholars position themselves for empirical testing.

To complete the definition of new institutional economics, it is important to expand upon the methods within the field. New institutionalists, unlike old, leverage formal modeling, rejecting some assumptions of neoclassical economics by introducing three tenets: transaction costs, bounded rationality, and opportunism (Ménard & Shirley, 2022). Notably, the field of institutional economics exhibits a commitment to narrative arguments like those seen in non-quantitative approaches to anthropology, history, political science, and sociology, featuring extensive expositions of instructive anecdotes, as well as formal case studies. One of the leading names associated with this approach is Elinor Ostrom, whose ethnography lies far away from the core methods found in Economics 101. However, as a whole, NIE embraces more of the core methods found in other circles such as formal modeling and modern econometric techniques.

42.2.2 Barriers to Running Field Experiments

The protracted nature of these definitions is necessary for one of the main arguments that this chapter makes, which is that field experiments are of limited use in understanding the macro-level outcomes often associated with institutional economics. The primary cause of this purported incompatibility with the “bigger” questions is that if one looks at an institutional economics paper dealing with such outcomes, and one identifies the key explanatory and dependent variables, a field experiment is usually infeasible, as controlling the assignment mechanism is difficult. In this regard, it is similar to many questions within mainstream macroeconomics.

Running field experiments in institutional economics is difficult for several reasons. First, the nature of the explanatory and dependent variables in institutional economics research typically makes a field experiment infeasible. In the case of the explanatory variables, many are too complex to be manipulated by a researcher, or require an unrealistic volume of financial resources. For example, scholars cannot—in a field setting—engineer wholesale changes in property rights (Davidson et al., 2018), judicial independence (Voigt et al., 2015), constitutional checks and balances (Fatás & Mihov, 2013), and so on. This is especially true because the sorts of institutions being studied are the ones that have an impact on some aggregate measure of societal welfare, and so represent pillars of society and the economy.

In the case of the dependent variables, the arguments inherent in an institutional economics paper usually imply outcomes that take years, decades, or even centuries to materialize, which would call upon the use of surrogates. As List (2025) discusses, the feasibility of a field experiment in such longitudinal settings revolves around generating treatments that affect surrogates and are the sole mediators of the treatment effect. In cases with high causal density, this is a difficult task because, as explained above, institutional economics is concerned with macro-societal outcomes such as the level of economic growth over a long period of time, the level of per capita income, the level of inequality, and so on. Thus, even if the requisite

control over the explanatory variables were to be secured, the practical constraints of generating useful surrogates are a difficult chore.

Furthermore, when the outcomes of interest are aggregate measures of societal welfare it creates problems of statistical power that are often insurmountable. Researchers cannot hope to have an institutional intervention where the unit of observation is a city or country, and to have control over the explanatory variables in enough cities or countries to generate satisfactory levels of statistical power. In fact, in many conventional institutional economics studies, the sample size is exactly one, as in the case study of Haiti in the period 1796–1820 (Ferrero, 2020). Moreover, when operating with units of observation as large as cities or countries, the problem of extraneous noise is amplified, making the need to secure a large sample more acute at a time when the feasibility is low. For example, culture and beliefs are critical mediating variables for the impact of institutions on economic growth, yet they can rarely be controlled for in a study because they are intangible and difficult to observe.

The challenge of sample size is made even more difficult by the importance of historical context to the arguments that some scholars make in institutional economics. A unique event in a certain location in history is often argued to be the driving force behind a pivotal institutional development, with the concomitant impact of economic growth, as in the case of England's Glorious Revolution of 1688 (North & Weingast, 1989). Reproducing these unique historical circumstances in a contemporary field experiment is impossible.

Finally, if one were to overcome the aforementioned feasibility constraints, there remain considerable ethical ones because when answering these macro questions the unit of analysis is generally an entire society. The process of randomization will mean that unlucky societies—be they cities or even entire countries—are potentially on the path of relatively low aggregate welfare outcomes for a significant period of time, which is morally unacceptable without research equipoise. For example, in the case of a study of judicial independence (Voigt et al., 2015), are we to condemn certain cities or countries to a system of kangaroo courts for the purposes of demonstrating how useful independent courts are? Informed consent is ruled out by natural field experiments, but even if informed consent were somehow compatible with a natural field experiment, it is unclear how one would secure it for an entire society.

Combining these factors highlights that the researcher interested in conducting field experiments in the area of institutional economics faces deep challenges, much as in many areas of macroeconomics (yet, for good examples of lab experiments in the area of macroeconomics please see the work of John Duffy and Charles Noussair, and coauthors). Phenomena such as inflation, unemployment, exchange rate regime, fiscal balances, and so on are ill-suited for field experiments for the same reasons as in institutional economics: they are not feasible, one cannot collect data, and it would be ethically unacceptable. Lest we be accused of hypocrisy, two of us wrote a review paper on field experiments in unemployment (Al-Ubaydli & List, 2019), yet if one examines the papers cited within the review, all of them are field experiments looking at a microeconomic aspect of unemployment. None have a dependent variable of aggregate unemployment, meaning that they are not macroeconomic papers. This foreshadows our next section, which argues that field experiments are of use in understanding such micro-foundations in institutional economics.

We close this section with a small anecdote that offers some validation for our claim. The experimental economics community (Economic Science Association) has an email listhost that members use to pose questions, advertise conferences, and so on. Members regularly send emails wherein they request references in a certain topic because they are authoring a survey paper and want to ensure maximum coverage. Fortunately, members are very cooperative, sharing the papers that they are familiar with. (The social norm is for the source of the original query to collate the suggested papers and then share a final list with the listhost.) In our case, we sent an email indicating our writing this chapter and asking for references. We received one reference only.

42.2.3 How Field Experiments Can Contribute to Institutional Economics

Despite the barriers that we argue that the scientist must overcome to conduct field experiments in institutional economics, we believe that there are several approaches in which field experiments can contribute to our understanding of institutional economics. We focus on two here.

The first is that—as intimated above—there exists a small but not insignificant minority of research questions in institutional economics that are potentially suitable for field experiments. In particular, the two scholars who shared a Nobel prize for institutional economics, Elinor Ostrom and Oliver Williamson, have conducted significant research where the dependent variable is a microeconomic phenomenon, and where the explanatory variables can be feasibly manipulated by an investigating researcher.

For example, Williamson (1967) analyzes optimal firm size as a function of various parameters within the organization, such as the firm's hierarchical structure and levels of compliance. Williamson (1988) examines how the characteristics of assets determine the optimality of debt versus equity finance for projects. Schlager and Ostrom (1992) study how the specific ownership structure of common-property resources affects usage incentives and outcome sustainability. Dietz et al. (2003) propose resolving non-local problems through a variety of strategies including dialogue between stakeholders and allowing for experimental solutions.

Naturally, one could make the argument that knowing how to structure a business well or how to manage a common-property resource effectively incrementally adds to our knowledge of how to make the economy grow. This is exactly what institutionalist studying micro-foundations do. This branch of NIE is critical to the fusion of NIE and field experiments.

This leads us to the second way in which field experiments can contribute to institutional economics which is by using them to understand microeconomic elements of the main economic growth-related institutions that scholars study in institutional economics, either in spirit or from a technical perspective. As mentioned above, this is the approach that we used in Al-Ubaydli and List (2019), and it will constitute the bulk of the survey that we will conduct in the next section.

At this point, it is worth clarifying what our approach has in common with the recently published (at the time of writing) paper by Callen et al. (2023), and how our

approach differs. Callen et al. (2023) is a survey of field experiments *on* institutions, which reveals an immediate difference: we are surveying field experiments in institutional economics. Their survey of the non-experimental literature leads them to the conclusion that certain institutional changes have been credibly demonstrated to have had persistent effects on the path of long-run development. Their focus then turns to using field experiments to study what determines institutional change, with a special emphasis on the value of measuring beliefs. We will cite many of the papers that Callen et al. (2023) cover in their highly valuable survey.

42.3 A Survey of Selected Field Experiments in Institutions-Mediated Economic Growth

At the outset of this survey, we drew readers' attention to the fact that we plan to cover selected papers based on our subjective view on what will be most beneficial to readers. This survey is not a systematic or comprehensive literature review.

42.3.1 Legal Systems

When analyzing the impact of legal institutions on economic growth, the most frequently cited element is property rights. As we argued above, randomizing property rights appears to be infeasible and morally questionable in certain cases. Nevertheless, Fabbri (2021) documents how this actually happened in rural Benin: the government formulated land tenure reforms that consisted of:

... formalizing customary tenure over land characterized by collective property and informal possession. This formalization was achieved by recording the set of rights each individual holds over land parcels, creating precise boundaries demarcation, and generating official land maps stored in a public repository. With the reform, the right-holders acquired formal, legally recognized use rights that can be traded, used as collateral, and defended in court against contenders. Thus, the reform produced a substantial shift toward a system of well-defined and individually assigned land rights.

Note that the unit of randomization was a rural village. As expected, almost a decade after the experiment took place, the study was unable to link the institutional reform to a central macroeconomic variable such as economic growth. However, Fabbri (2021) ran a series of artefactual and framed field experiments and found that people benefiting from the reforms exhibited more cooperative and trusting behavior within those experiments. These findings are consistent with the broad-based positive effect of property rights that scholars in institutional economics often argue.

Other studies are unable to randomize property rights, but they do randomize awareness and enforcement of property rights. Subramanian and Kumar (2024) randomly nudge Indian farmers into acquiring formal land titles, and they find that it increases agricultural investments, thereby raising various agricultural performance measures. The benefiting households also experience a net increase in per

capita consumption. Aberra and Chemin (2021) randomize access to the free services of a lawyer for 2 years in rural Kenya. The result was more secure property rights, thereby inducing greater investment and access to credit. Similar results were found in the context of a Western country (the United States) by Greiner and Pattanayak (2011), who randomized access to legal assistance for people facing eviction. Blair et al. (2019) is a field experiment in Liberia where teams of superior (better trained and equipped) policy officers had recurring patrols randomized by rural village, resulting in more secure property rights and less crime, though without any significant impact on trust in the police, courts, or government.

In the adjacent domain of common property resources, Cavalcanti et al. (2010) conduct a field experiment in six fishing communities in Brazil to investigate the impact of having a participatory management approach on people's willingness to individually contribute to solving the problems. One of the structural challenges they face is a small number of units of randomization: they randomly allocate three of the six participating villages into a participatory approach to managing the common resources, and three into the control. They find the participatory approach—via a variety of mechanisms including beliefs about the willingness of others to cooperate—results in higher levels of cooperation.

Exploring beyond property rights to broader studies of the legal system, field experiments have been deployed quite widely (Green & Thorley, 2014), though primarily as a means of assessing the impact of various factors on the performance of the legal system, without necessarily linking the performance of the legal system to aggregate societal welfare. For example, many field experimental studies demonstrate the important result that increasing perceived law enforcement for certain crimes (in this case via visually salient enforcement) leads to a decrease in the incidence of those crimes (Dur & Vollaard, 2019). Peyton et al. (2019) use a field experiment with a large urban police department and find that the alternative to hard enforcement—community-oriented policing—is effective in improving police legitimacy and residents' willingness to cooperate with police. Findley et al. (2015) use a field experiment to vary the properties of a request for anonymous incorporation, and they find that offers to “pay a premium” (illicit payments) reduce compliance, while alerting companies to international law has no effect. Kassin et al. (2019) find that recording crime suspects during interrogations in the United States had no effect on the important outcome variables such as the propensity to confess or to waive protections against self-incrimination (so-called Miranda rights), demonstrating the importance of features in which a formal system is operating in.

The courtroom also attracts the attention of researchers using field experiments. What must surely be one of the earliest legal field experiments was conducted by Heuer and Penrod (1989), which found that instructing jurors prior to the evidence part of the trial and providing them with a written copy of the judge's instructions has no discernible impact on a variety of outcome measures, such as ability to recall the judge's instructions, trial length, or juror satisfaction. The effect of juror note-taking was investigated in a field experiment by Heuer and Penrod (1994) and they also find no effect. Thompson et al. (2022) randomly vary the content of legal Wikipedia articles to show that judges apparently use Wikipedia to inform their judgments.

Ho and Ouellette (2020) show that the expansion of scientific peer review in patent determinations slightly improved search efforts and citations to the non-patent scientific literature, albeit at a considerable resource cost.

One of the oft-proposed methods for improving the functioning of the legal system—namely decreasing corruption—is greater transparency. Peisakhin and Pinto (2010) induce random variation in access to ration cards to demonstrate that India's freedom of information law is almost as effective as bribery in helping the poor secure access to basic public services. Worthy et al. (2017) compare freedom of information requests in English parish councils to simple asks and find that the former are significantly more effective. Karim (2020) induces transparency via household visits by police in Liberia, and she uses a field experiment to demonstrate a positive effect on trust in the legal system.

One huge sub-literature that intersects with field experiments in legal systems is field experiments on discrimination. Bertrand and Duflo (2017) provide a comprehensive survey, including several law-related findings such as: people illegally discriminate, sometimes consciously, other times not; the factors that contribute to higher levels of illegal discrimination; and so on. Another comparable large sub-literature examines tax compliance, but we defer an exposition to the section below that examines state capacity.

There are many other legal field experiments one could expand upon, but the pattern is clear. First, field experiments about legal systems are extremely diverse, and there is a high degree of context dependence. Accordingly, it is difficult to derive *general* conclusions beyond the ones that theory would point to *ex ante*, such as higher enforcement works, and transparency helps. Field experiments have helped rigorously demonstrate the sensitivity of the behavior of judicial actors (judges, jurors, lawyers, etc.) to seemingly innocuous factors, though running these kinds of experiments involves traversing an ethical minefield given the stakes associated with erroneous judicial decisions. In the context of property rights, the limited evidence does support the Northian argument that property rights are conducive of economic growth.

Though these papers provide little to no direct evidence of the effect of institutions on macroeconomic variables such as economic growth, they do provide micro-founded insights such as how specific institutions work. These examples illustrate where we see field experiments contributing to institutional economics.

42.3.2 Political Systems

The attributes of the political system have been a very popular area for field experimentation and the literature is several orders of magnitude larger than that in the legal arena, possibly because of the greater ability to experimentally manipulate key explanatory variables. However, just as in the case of legal systems, the studies—by our criteria—drift from the core of institutional economics because they are unable to link the interventions to aggregate societal welfare measures via surrogates of economic growth.

Following Fukuyama (2011) and Callen et al. (2023), when considering aggregate welfare, we divide the political system's characteristics into two primary

spheres: those that relate to the state's capacity, and those that relate to accountability within the state. Starting with the former, field experiments have studied two main determinants of state capacity.

The first is a functional one, which is the government's ability to collect tax revenues, which it can then allocate to the various responsibilities that the state assumes. The reviews by both Green and Thorley (2014) and Callen et al. (2023) provide excellent thorough surveys, which we selectively explore here. The simplest interventions include sending letters that increase the salience of audits (Kleven et al., 2011), or that remind people of their responsibility to pay taxes (Hallsworth et al., 2017). This latter example and others, such as Torgler (2013), build on the larger social sciences literature whereby moral suasion in the absence of material incentives is used to induce prosocial behavior (Ferraro & Price, 2013). In general, these informational interventions find statistically significant but economically modest positive impacts, though given that the cost of the intervention is usually trivial, it typically passes a cost-benefit test.

Notice that feasibility and potential ethical concerns limit scholars' ability to directly investigate the most obvious way of increasing compliance, i.e., increasing enforcement. Instead, as described above, they typically use workarounds, such as making existing enforcement levels more cognitively salient. A notable exception is De Andrade et al. (2013), which explores various ways of motivating informal businesses in Brazil to formally register (a stepping-stone toward paying taxes). They find that the informational treatments have no effect, while increasing inspections does substantially increase formal registration of businesses.

The second vein of state capacity-related field experimental research concerns increasing citizens' trust in government, as trust is widely regarded to be a great economic lubricant (Fukuyama, 2011). When people trust the government, a smaller volume of costly coercive effort is required to induce cooperation with the state, and to thereby effect the policies that the state prefers. Callen et al. (2023) report the findings of several field experiments where researchers worked with governments to increase citizens' trust in their governments. Haim et al. (2021) estimated the impact of an outreach program that helped Philippine citizens positively update their beliefs about the competence of their government, resulting in them providing the state with more timely information about the Covid-19 pandemic.

In another interesting piece of research in the second vein, Blattman et al. (2022) used a field experiment to demonstrate—in villages with certain attributes—improvements in government legitimacy and decreases in emergency calls in response to the state deploying full-time liaisons to assist with intra-community conflict resolution and to foster communication with the government. Likewise, Blair et al. (2022) study a separate intervention in Colombia designed to "... replace rebel governance by harnessing complementarities between state and communal authorities and by improving security and justice provision in areas once dominated by FARC, the country's largest rebel group." They also reported improved local conflict resolution and increased citizens' trust in some state institutions.

We next consider the other sphere of political systems, accountability, which has been a particularly vibrant area of research for scholars using field experimental

methods. The two relevant review papers we came across, Green and Thorley (2014) and Callen et al. (2023), together cite well over 20 studies conducted during the last 15 years examining a wide variety of aspects of accountability. Interestingly, the designs are fundamentally homogeneous: the scholars conjecture that a certain factor makes people more willing and able to hold their principals accountable, or makes the principals more willing and able to be held accountable by the people they serve. This factor can be individual level or group level, and it is randomly varied, with the outcome variables being either accountability activity (voting in elections, effort to communicate with senior officials, etc.) or some measure of satisfaction with the accountability process (trust in principals, quality of elected officials, etc.).

For example, Casey et al. (2021) intervene in elections in Sierra Leone by randomly changing the degree to which candidates are directly nominated by voters (at-large nominations) versus nominated by party officials. The latter, indirect route leads to candidates who are less likely to satisfy voters' preferences. These results echo the earlier findings of Beath et al. (2016), which randomly varied the electoral institutions of 250 Afghan villages to compare at-large elections with district-bound ones, finding that the at-large system produced superior outcomes in terms of politician quality (as measured by having completed high school). Olken (2010) drew comparable conclusions based on a field experiment in Indonesia.

Novel technological interventions are a popular choice for field experiments because when one starts with a baseline where the new technology is absent, and the technology is likely to be helpful, it is ethically palatable to randomly introduce that new technology due to research equipoise. Grossman et al. (2014) do this by making it cheaper for Ugandan citizens to send text messages to their representatives, and they find a high level of take-up that allowed marginalized populations to gain better access to politicians than would have been possible via traditional channels. Marx et al. (2021) also experiment with text messages, but as a voter mobilization tool in Kenya, and they find increased turnout but decreased trust. This latter finding is important in demonstrating the context-sensitivity of interventions in environments as complex and riddled with path-dependencies as national elections.

One recent area of research is the effect of partisan media on the quality of political institutions. Conroy-Krutz and Moehler (2015) compare home partisan to cross-cutting (opposing partisan) radio broadcasts in Ghana, finding that the latter were persuasive and worked to moderate voter views. Gerber et al. (2010) induce partisanship in the United States by reminding voters that they need to be registered with a party to vote in that party's primaries. The result is a strengthening of partisan identity and more partisan voting behavior. Porter and Rogowski (2018) demonstrate that partisan email writers during local elections were more likely to elicit a response from local election clerks. Both of these latter partisanship studies further show how complex political systems are in practice, as it is unclear whether the behavioral shift induced by partisanship is something that improves or diminishes the performance of political institutions.

There are also many papers that examine the effect of civic education on the political system, again due to the relative absence of feasibility and ethical barriers to

such experiments. Some occur in developing economies, such as Finkel and Lim (2021) and Finkel et al. (2023); while others occur in advanced economies, such as Claassen and Monson (2015) and Green et al. (2011). Brock-Petrossius (2023) is an interesting addition to this latter group that experimentally explores the impact of deep canvassing and finds that is effective in changing attitudes. In general, studies of the effect of civic education on the performance of the political system tend to find positive effects on self-reported measures of civic-mindedness, and on forms of political engagement.

The sub-topics we have covered here are a mere selection of what can be found in a diverse literature. We encourage those interested to explore the additional topics in Callen et al. (2023), such as political inclusiveness, political accountability interventions relating to politicians, the role of social media, participation on protests, and so on.

42.3.3 Informal Social Systems Including Religion

Today, in early 2024, theocracies play a very limited role in the global political system, yet religion remains important to daily life, thereby transforming it into a quasi-informal social system. Institutional economics pays great attention to informal social systems, with one of the most famous examples being the Weberian argument regarding the persistent effect of the Protestant work ethic on productivity in the twentieth century, well after the formal separation of church and state.

Before exploring the empirical evidence, it is worth briefly recapping the theory that guides how we should expect religion to affect aggregate welfare. In primitive societies, the belief in supernatural deities that reward righteousness and punish anti-social behavior can help overcome social dilemmas such as the free-rider problems, as adherents will fear divine retribution lest they behave in an anti-social manner (Fukuyama, 2011). Religion is also typically the first source of rule of law, where rule of law is defined as senior officials (including the head of state and head of government) being subject to the same restrictions as ordinary people; as these senior officials who have dominion over the masses must abide by religious restrictions. At various points in history, religious institutions have amassed political power independently of the state, helping them to act as a check on the state abusing its power. However, the key weakness that religion suffers from as an institutional source of superior performance is that it creates an in-group bias and exclusionary tendencies toward out-group members that may ultimately result in a net negative impact on aggregate societal welfare (Norenzayan & Shariff, 2008). This ambiguity—which puts religion in stark contrast to many other important institutions studied by institutional economists—enhances the value of empirical research in understanding its impact, and makes field experiments a potentially invaluable tool.

The main problem field experimental researchers face, however, is that adherence to a religion cannot be readily manipulated, nor can other important explanatory variables such as faith, edicts, acts of worship, and so on (Wright, 2018). Researchers are restricted to interventions such as priming religion in an individual and seeing how it affects their behavior, or manipulating the observability of an individual's religion

and seeing how it affects how others behave toward them. This latter approach has generated many studies of religious-based discrimination, such as Wright et al. (2013) and Valfort (2020); see Bertrand and Duflo (2017) for a more comprehensive list. These types of studies make salient the aforementioned in-group out-group problems, as religion ends up working as a vehicle that creates grounds for socially destructive animus in a wide variety of settings including labor markets, housing, socializing, charitable contributions, and so on. It is fair to say that religious discrimination studies account for a very large proportion of religious field experiments, primarily due to the ease of executing the basic audit design to which they mostly adhere.

The priming design is often used to demonstrate the potentially positive effect of religion on prosocial behavior. For example, Duhaime (2015) finds that the Muslim call for prayer induces a transient increase in prosocial behavior in Moroccan markets, while Al-Ubaydli et al. (2023) find that religious text messages induce higher levels of prosocial conservation in electricity consumption among residents of Qatar. Feldhaus et al. (2022) demonstrate that announcing the Catholic church's support for climate protection efforts induces higher charitable donations in the German Catholic Convention.

A unique study that likely has limited generalizability but that is highly instructive within the original context is the field experiment conducted by Clingingsmith et al. (2009), which involved randomizing participation in the greater pilgrimage in Islam, the *hajj*. The design leverages the fact that most Muslims only perform it once (if at all) during their lifetimes, and that it constitutes an extremely significant component of their cumulative religious activity. The authors find that it has a lasting, positive effect on tolerance, thereby inducing prosocial behavior which captures the importance of taking part in the institutional proceeding.

When it comes to informal social systems, institutional economics looks well beyond religion, too, especially as the world becomes more secular. There are far too many social norms to enumerate, though some of the more salient ones (Elster, 1989) include reciprocity (rewarding those who treat you well and punishing those who treat you badly), kin selection (treating your close kin better than non-kin members), in-group favoritism (behaving more favorably toward fellow members of a group you identify with, such as an ethnic or confessional group), and inequality aversion (being averse to an unequal distribution of resources within your group). As with the specific case of religious-based social norms, social norms in general are difficult for investigating researchers to manipulate because they are formed over decades (and sometimes even millennia), forcing researchers to use priming and salience. Yet, social norms acting as moderators of effects represent a key area of scholarly research (Levitt & List, 2007).

Additionally, it is worth affirming the potential value of experimental empirical work in studying social norms, as there exists theoretical uncertainty over the growth implications of several key social norms. Perhaps the most famous is the perpetual debate about whether collectivist or individualist cultures yield superior economic performance (Kyriacou, 2016). Some of the key social norms that have been studied using field experiments are as follows.

Bergquist et al. (2020) is one of many field experiments that leverages social norm-based moral suasion to induce prosocial behavior. It follows the typical design

of providing descriptive and injunctive social normative information, in this case relating to the importance of protecting the environment. The result was higher pro-environmental donations. Ferraro and Price (2013) provide information on people's water consumption compared to that of their neighbors and find that this creates an incentive for people to conserve water. Senderey et al. (2022) leverage social norms of responsibility toward others to induce higher vaccination rates in a field experiment. These examples are merely illustrative, as there exist numerous papers with analogous designs, many of which fall under the banner of "nudge" economics (Thaler & Sunstein, 2009).

Yoeli et al. (2013) take an indirect approach to studying the social norm of indirect reciprocity, i.e., the propensity to help people who help others, and punish those who mistreat others. A key moderating variable that determines the effectiveness of indirect reciprocity as a social norm is the observability of people's actions. The authors use a field experiment to manipulate observability in a utility company's blackout-reduction program, and they find increasing observability has a strong positive effect on people's prosocial behavior. Mujcic and Leibbrandt (2018) also use a field experiment to study indirect reciprocity, and show that it is effective in inducing prosocial behavior in the context of drivers in traffic permitting others to switch lanes.

Bandiera et al. (2020) use a field experiment to estimate the interaction between the degree of individualism-collectivism and the effectiveness of piece-rate incentives offered to workers. They find that piece rates are more effective for people who come from societies with an individualistic social norm, but it should be noted that individualism (or any prime thereof) is not randomized in this field experiment; instead, it is the strength of the incentives, and so this is technically not a randomized control field experiment in the context of estimating the impact of a social norm.

Scholars have also used informational interventions in field experiments to help correct misconceptions that might perpetuate out-of-date social norms. Bursztyn et al. (2020) do this in the case of the social norm of women staying out of the labor force in Saudi Arabia in accordance with their husbands' wishes. The authors show that when young married men learn that they have systematically underestimated acceptance of women working among their peer groups, they themselves become significantly more supportive of their own wives seeking employment. In the long-run, this can yield significant aggregate economic benefits as the labor market potential of women is unleashed.

The previous section on political systems also features several interventions designed to improve social norms related to conflict management in populations that may suffer from weak conflict management norms and/or weak state enforcement. For example, Blattman et al. (2014) showed how community education helped to establish better dispute resolution norms, though in such contexts, it is difficult to distinguish between a social norm (people have an improved mutual understanding about how to resolve conflicts) and human capital (people have acquired hard conflict resolution skills).

This selection of studies is—as in the case of political systems—far from exhaustive due to the wide range of social norms that have been scientifically investigated using field experiments. Though the range of social norms that could plausibly cause higher economic growth is considerably narrower, institutional economists have

allocated a smaller volume of their research efforts to studying such norms, as compared to formal legal and political institutions. This means that there are many opportunities to deepen our understanding of the effect of social norms, a point that we return to below in the epilogue. From the field experiments that we have surveyed, the breadth of inquiry and the importance of context together work to limit the ability to draw general conclusions, especially when combined with the structural barriers to running field experiments that we covered in previous sections. There is suggestive field experimental evidence that religion can be a force for the good, but the research designs preclude the general equilibrium analysis required to explore the major weakness that religions suffer from as a system for organizing behavior: the bias against out-group members. Continuing to explore these issues as the tools for running field experiments improve holds much promise.

42.4 Epilogue: Where Do Field Experiments Hold Most Promise for Institutional Economics and Economic Growth?

As field experimentalists, despite arguing that “true institutional economics” is difficult to address with field experiments, we remain hopeful that some of the barriers that we described can be overcome with a combination of technological advancement and lateral thinking. In this section, we first consider ideas grounded in institutional economics laboratory experiments, and then proceed to consider some more novel designs. We conclude the section by discussing the issue of mixed methods.

42.4.1 Ideas Grounded in Laboratory Experiments

Within institutional economics there is an evergreen debate concerning individualism versus collectivism, decentralization versus central control, markets versus hierarchies, socialism versus capitalism, and so on. Narrative, ethnographic, and empirical studies based on observational data have failed to resolve the debate, while field experiments have (to the best of our knowledge) failed to make a contribution. However, many relevant laboratory experiments have been run, including Al-Ubaydli et al. (2013) and Choi and Storr (2022), which supraliminally prime market experience among experimental participants and then observe the effect on trust. With a sufficient budget, these studies could be translated into field experiments. While the impact on truly aggregate welfare outcomes could not be observed and surrogates might be difficult to deliver, the design could be applied to small groups that have to live in comparative isolation for extended periods of time, such as military squads during training, generating informative data. More generally, as seen in religion-related field experiments, any laboratory experiment that relies on priming or salience can typically be translated into a field experiment in a straightforward manner.

A deep behavioral economics literature that has developed concerning decentralization versus central control revolves around whether preferences are defined over consumption levels or changes in consumption (Kahneman & Tversky, 1979). In a positive sense, the value disparity has considerable relevance. For example, such preferences call into question commonly held interpretations of indifference curves, make cost-benefit analysis illegitimate, change the procedure necessary to resolve damage disputes, and cause a reconsideration of several of our other deep intuitions, such as the invariance result of the so-called Coase theorem which relies on Hicksian equivalent surplus and Hicksian compensating surplus to be roughly equivalent (List, 2004b). While the theoretical paper of Kahneman and Tversky (1979) helped to set up the field of behavioral economics, Kahneman et al.'s (1990) laboratory experiment struck a deep empirical chord. The work has spawned several laboratory experiments that can be translated into a field experiment in a straightforward manner, and indeed that process has already begun (List, 2004b).

As an example, scholars took the baton from the important laboratory work in the 1980s and 1990s and used field experiments to show that the endowment effect is not a fundamental characteristic of preferences that is immutable by market experience (see, e.g., List, 2003). Lab and field studies replicated List's experience results (e.g., Engelmann & Hollard, 2010; Sutter, 2021) and new insightful theories were introduced (e.g., Köszegi & Rabin, 2006).

Another leading issue in institutional economics is the impact of property rights on long-run growth. Kimbrough et al. (2008) explore this issue in a laboratory experiment featuring a somewhat abstract trading game, where inducing a history of secure property rights encourages more impersonal trade in the long run and hence higher levels of aggregate societal welfare. Here, the nature of the intervention is a protracted historical experience, and so it is much more difficult than priming experiments to translate into a field experiment. One option is massively multiplayer online role playing games (MMORPGs), which can act as a bridge between the realism of the real world and the pliability of virtual worlds. Experiments relating to trade have been run in such environments (De Sousa & Munro, 2012), and while they are unquestionably less generalizable than field experiments in non-virtual settings, these are the sort of epistemic tradeoffs that inevitably arise when tackling difficult questions. More generally, virtual worlds are a likely underleveraged source of scientific knowledge in economics, and their potential value will probably increase as virtual reality technologies continue to improve.

42.4.2 Greenfield Field Experiments

The difficulties described above in conducting field experiments relating to informal social systems, especially religion, are not insurmountable, and Wright (2018) provides some sage strategies for overcoming the feasibility and ethical challenges relating to the manipulation of social norms. The most promising is randomizing access to oversubscribed programs relating to the social norm, as Clingingsmith et al. (2009) did in the case of the pilgrimage that Muslims perform typically once

during their lives. Ideally, as in the case of Islam's greater pilgrimage, the program would represent a highly significant component of the formation of that social norm. For many social norms, including religious ones, no intervention approaches the significance of the greater pilgrimage in Islam, but there are some unexplored (to the best of our knowledge) options.

One example is the programs designed to socialize immigrants so that they adopt the social norms of the country they are emigrating to. For example, Norway has a cultural orientation program that helps foster the integration of resettled refugees in local communities. This multi-month program is ripe for a field experiment, especially if one is to collect data on dependent variables of import. These would include individual-level outcomes for the immigrants, and society-level ones for the settlements that receive many immigrants. Norway likely does not have enough settlements for a high-powered study but were a similar program to be introduced to the United States, then over time, enough data at the aggregate level could potentially be collected that ultimately adds to our knowledge about the impact of informal institutions on economic growth.

If one affords us with a slightly expanded definition of institutional economics, then for our second greenfield example we would advocate going down the human capital route. This is because human capital is important for long-run growth, and schools are institutions that provide human capital. Field experiments have been used to explore the education production function and early skill formation, which are both linked to the long-run growth of an economy.

Consider Schweinhart (1993) analysis of the Perry Preschool Project, a preschool program launched in the 1960s. While the Perry Preschool Project began as early childhood framed field experiment to understand whether high-quality education benefits at-risk children, it soon became a longitudinal study documenting the lasting impact of the intervention on adult outcomes. Researchers, including James Heckman and colleagues, have reported that the project produced dramatic benefits for academic achievement, graduation rates, and earnings. It also lowered rates of crime, delinquency, teenage pregnancy, and welfare dependency when compared to the control group. We view future work using field experiments to provide such insights as key contributors to a more broadly defined institutional economics.

One feature that this example raises that has been mentioned in various places of this chapter is how to provide scientific insights into long-run problems. As aforementioned, economic experiments typically focus on estimating short-run substitution effects. Yet, in many cases understanding the long-term effects of an intervention is critical from a policy and scientific perspective. This is especially true within institutional economics, where key questions revolve around effects measured in the long run, such as economic growth. In such cases, it is fundamental to understand causal pathways.

To leverage causal pathways in the long-run, a first key consideration is an outcome's causal density. The causal density of an outcome variable refers to how well the determinants of the outcome are understood. An outcome with low causal density is one where the factors determining the outcome are well understood. For example, farmers have a near-perfect understanding of the determinants of crop yields, yielding

a low causal density. In contrast, the determinants of individual labor market outcomes or economic growth are much less well understood and thus have high causal density. For example, future employment and wage outcomes are a function of the numerous and interconnected inputs by multiple people and institutions.

When exploring outcome variables with low or high causal density, a key consideration is whether a feasibly observed proxy variable, called a statistical surrogate, can be measured in the short-run that estimates long-run outcomes. While the use of surrogates within economics is a developing area of research, we view its development as a key unlock that can usher in a new wave of field experiments in institutional economics. This is because development of attractive surrogates has the potential of moving the research agenda from one of the short-run measurements to a viable approach to examine macro outcomes of interest in institutional outcomes, such as long-term economic growth, or some alternative measure of aggregate welfare, such as per capita income, societal inequality, or political stability. As soon as relevant surrogates are developed, we envision field experiments as an attractive tool that will experience substantial growth amongst institutional economists.

42.4.3 Mixed Methods

While we are optimistic about the progress that can be made in using field experiments based on introspection about new designs and by leveraging new technologies, we remain realistic that the fundamental difficulties discussed throughout this chapter are unlikely to be overcome in the near future. Accordingly, we wish to extoll studies that use mixed methods to try to minimize the weaknesses associated with each type of data.

Orthodox economists have—in our opinion—been needlessly hostile to narrative and ethnographic arguments made within the profession and by other social scientists. While we are sympathetic to the view that a rigorous randomized control field experiment adds more to our knowledge of institutional economics than a purely narrative argument, we firmly subscribe to the view that a combination is superior. This is especially important because—unlike the physical phenomena that natural scientists study in the laboratory—the human phenomena at the heart of institutional economics exhibit very high degrees of context dependence. Accordingly, those looking to advance their knowledge would benefit considerably from understanding the historical context. Ethnographic work in tandem with field experiments, especially in-depth interviews, can be vital for generating hypotheses to test with field experiments, and for solving empirical puzzles.

In this regard, we are particularly supportive of taking the ethnographic and narrative work of scholars such as Elinor Ostrom and combining it with the sort of field experiments we surveyed in this chapter, i.e., the ones that do not present complete answers to institutional economics questions, but that contribute part of the answer based on a watering down of the original research question. Surveys are also growing in popularity in economics generally and can play an important role as a complement to imperfect field experiments in institutional economics. Blattman et al. (2021) use a variety of empirical methods—quantitative and qualitative; interviews, surveys,

observational data, and quasi-experiments—all conducted in the field to study institutions in Medellin, Colombia; and they combine this with theoretical modeling. We regard this approach to be a snapshot of the future of empirical institutional economics.

42.5 Conclusion

Institutional economics is a field of economics that gets to the heart of what the discipline is ultimately about: helping societies to improve their living standards. As with many foundational research questions in economics, there are no easy answers, and scholars are forced to combine a selection of imperfect epistemic tools to advance our understanding of the phenomena that underpin economic growth. The importance of historical context means that for the most part, the literature has relied on narrative and ethnographic methods, combined with the econometric analysis of observational data. While we collectively know a lot more today than a century ago about how institutions such as property rights, rule of law, and cooperative norms impact economic growth, there remain huge gaps in our understanding. The problems are made more challenging by the fact that the world is constantly evolving, and the determinants of economic prosperity are always mutating.

For the most part, as stand-alone contributions, field experiments face difficulties in institutional economics at the macro level: they are frequently infeasible, potentially unethical, and impossible to statistically power and find useful surrogates. Nevertheless, field experiments are a promising tool for understanding small pieces of a bigger puzzle such as economic growth with the caveat that the researcher must ensure an understanding of how the puzzle fits back together (List, 2024), and as already demonstrated by Williamson and Ostrom for understanding explanatory variables and outcomes that are feasibly manipulated.

The last decade in particular has seen an explosion of field experiments in the study of relevant factors such as property rights, state capacity, political accountability, and so on. Despite this progress, the literature remains nascent, and the studies remain context-dependent, limiting the reach where we can make general conclusions or offer a list of “regular” empirical findings. There also remains a significant gap to bridge regarding the ultimate impact of field experimental interventions on aggregate welfare metrics, most notably economic growth.

For this reason, we urge scholars to overcome any excessive affinity they might have with certain methodological approaches, and to embrace mixed methods that include field experiments and complementary approaches as the ideal path forward. For example, insisting on randomized field experiments to study the critical question of how property rights affect development yields limited work, and we discuss one study conducted in Benin. Rather than putting our proverbial intellectual eggs all in the experimental basket, when it comes to institutional economics, we argue that it makes sense to place weight on the conclusions drawn from narrative, ethnographic, and observational studies as well.

Acknowledgment of Funding Open access made possible through a generous donation in honor of the Ronald Coase Institute

References

- Abera, A., & Chemin, M. (2021). Does legal representation increase investment? Evidence from a field experiment in Kenya. *Journal of Development Economics*, 150, 102612.
- Acemoglu, D. (2008). Growth and institutions. In S. Durlauf & L. Blume (Eds.), *The new Palgrave dictionary of economics* (2nd ed.). Palgrave.
- Al-Ubaydli, O., & List, J. A. (2013). *On the generalizability of experimental results in economics: With a response to Camerer* (No. w19666). National Bureau of Economic Research.
- Al-Ubaydli, O., & List, J. A. (2019). How natural field experiments have enhanced our understanding of unemployment. *Nature Human Behaviour*, 3(1), 33–39.
- Al-Ubaydli, O., Houser, D., Nye, J., Paganelli, M. P., & Pan, X. S. (2013). The causal effect of market priming on trust: An experimental investigation using randomized control. *PLoS One*, 8(3), e55968.
- Al-Ubaydli, O., Cassidy, A. W., Chatterjee, A., Khalifa, A., & Price, M. K. (2023). *The power to conserve: A field experiment on electricity use in Qatar* (No. w31931). National Bureau of Economic Research.
- Bandiera, O., Dahlstrand-Rudin, A., & Fischer, G. (2020). “Incentives and culture: Evidence from a multi-country field experiment”. AEA RCT (2020): 4685–2.
- Barro, R. J. (2003). Determinants of economic growth in a panel of countries. *Annals of Economics and Finance*, 4, 231–274.
- Beath, A., Christia, F., Egorov, G., & Enikolopov, R. (2016). Electoral rules and political selection: Theory and evidence from a field experiment in Afghanistan. *The Review of Economic Studies*, 83(3), 932–968.
- Bergquist, M., Nyström, L., & Nilsson, A. (2020). Feeling or following? A field-experiment comparing social norms-based and emotions-based motives encouraging pro-environmental donations. *Journal of Consumer Behaviour*, 19(4), 351–358.
- Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. In *Handbook of economic field experiments* (Vol. 1, pp. 309–393).
- Blair, R. A., Karim, S. M., & Morse, B. S. (2019). Establishing the rule of law in weak and war-torn states: Evidence from a field experiment with the Liberian National Police. *American Political Science Review*, 113(3), 641–657.
- Blair, R. A., Moscoso-Rojas, M., Castillo, A. V., & Weintraub, M. (2022). Preventing rebel resurgence after civil war: A field experiment in security and justice provision in rural Colombia. *American Political Science Review*, 116(4), 1258–1277.
- Blattman, C., Hartman, A. C., & Blair, R. A. (2014). How to promote order and property rights under weak rule of law? An experiment in changing dispute resolution behavior through community education. *American Political Science Review*, 108(1), 100–120.
- Blattman, C., Duncan, G., Lessing, B., & Tobón, S. (2021). *Gang rule: Understanding and countering criminal governance* (No. w28458). National Bureau of Economic Research.
- Blattman, C., Duncan, G., Lessing, B., & Tobón, S., (2022). *Civilian alternatives to policing: Evidence from Medellín’s community problem-solving intervention Operación Convivencia*. No. w29692. National Bureau of Economic Research.
- Brock-Petrossius, K. (2023). *Changing dominant Carceral attitudes: A community organizing field experiment*. Doctoral dissertation, UCLA.
- Bursztyn, L., González, A. L., & Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in Saudi Arabia. *American Economic Review*, 110(10), 2997–3029.
- Callen, M., Weigel, J. L., & Yuchtman, N. (2023). *Experiments about institutions* (No. w31964). National Bureau of Economic Research.
- Casey, K., Kamara, A. B., & Meriggi, N. F. (2021). An experiment in candidate selection. *American Economic Review*, 111(5), 1575–1612.
- Cavalcanti, C., Schläpfer, F., & Schmid, B. (2010). Public participation and willingness to cooperate in common-pool resource management: A field experiment with fishing communities in Brazil. *Ecological Economics*, 69(3), 613–622.
- Choi, G. S., & Storr, V. H. (2022). The market as a process for the discovery of whom not to trust. *Journal of Institutional Economics*, 18(3), 467–482.

- Claassen, R. L., & Monson, J. Q. (2015). Does civic education matter?: The power of long-term observation and the experimental method. *Journal of Political Science Education*, 11(4), 404–421.
- Clingsmith, D., Khwaja, A. I., & Kremer, M. (2009). Estimating the impact of the hajj: Religion and tolerance in Islam's global gathering. *The Quarterly Journal of Economics*, 124(3), 1133–1170.
- Conroy-Krutz, J., & Moehler, D. C. (2015). Moderation from bias: A field experiment on partisan media in a new democracy. *The Journal of Politics*, 77(2), 575–587.
- Davidson, S., De Filippi, P., & Potts, J. (2018). Blockchains and the economic institutions of capitalism. *Journal of Institutional Economics*, 14(4), 639–658.
- De Andrade, G. H., Bruhn, M., & McKenzie, D. J. (2013, May 1). A helping hand or the long arm of the law? Experimental evidence on what governments can do to formalize firms. *Experimental evidence on what governments can do to formalize firms* (World Bank policy research working paper, (6435)).
- De Sousa, Y. F., & Munro, A. (2012). Truck, barter and exchange versus the endowment effect: Virtual field experiments in an online game environment. *Journal of Economic Psychology*, 33(3), 482–493.
- Dietz, T., Ostrom, E., & Stern, P. C. (2003). The struggle to govern the commons. *Science*, 302(5652), 1907–1912.
- Duhaime, E. P. (2015). Is the call to prayer a call to cooperate? A field experiment on the impact of religious salience on prosocial behavior. *Judgment and Decision making*, 10(6), 593–596.
- Dur, R., & Vollaard, B. (2019). Salience of law enforcement: A field experiment. *Journal of Environmental Economics and Management*, 93, 208–220.
- Elster, J. (1989). Social norms and economic theory. *Journal of Economic Perspectives*, 3(4), 99–117.
- Engelmann, D., & Hollard, G. (2010). Reconsidering the effect of market experience on the 'Endowment Effect'. *Econometrica*, 78(6), 2005–2019. <https://doi.org/10.3982/ECTA8424>
- Fabbri, M. (2021). Property rights and prosocial behavior: Evidence from a land tenure reform implemented as randomized control-trial. *Journal of Economic Behavior & Organization*, 188, 552–566.
- Fatás, A., & Mihov, I. (2013). Policy volatility, institutions, and economic growth. *Review of Economics and Statistics*, 95(2), 362–376.
- Feldhaus, C., Gleue, M., & Löschel, A. (2022). Can a Catholic institution promote sustainable behavior? Field experimental evidence on donations for climate protection. *Journal of Behavioral and Experimental Economics*, 98, 101855.
- Ferraro, P. J., & Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment. *Review of Economics and Statistics*, 95(1), 64–73.
- Ferrero, M. (2020). Accidental socialism: a natural experiment in Haiti 1796–1820. *Journal of Institutional Economics*, 17(3), 393–409.
- Findley, M. G., Nielson, D. L., & Sharman, J. C. (2015). Causes of noncompliance with international law: A field experiment on anonymous incorporation. *American Journal of Political Science*, 59(1), 146–161.
- Finkel, S. E., & Lim, J. (2021). The supply and demand model of civic education: Evidence from a field experiment in the Democratic Republic of Congo. *Democratization*, 28(5), 970–991.
- Finkel, S. E., Neundorff, A., & Rascon Ramirez, E. (2023). Can online civic education induce democratic citizenship? Experimental evidence from a new democracy. *American Journal of Political Science*, 68(2), 613–630.
- Fukuyama, F. (2011). *The origins of political order: From prehuman times to the French revolution*. Farrar.
- Gerber, A. S., Huber, G. A., & Washington, E. (2010). Party affiliation, partisanship, and political beliefs: A field experiment. *American Political Science Review*, 104(4), 720–744.
- Green, D. P., & Thorley, D. R. (2014). Field experimentation and the study of law and policy. *Annual Review of Law and Social Science*, 10, 53–72.

- Green, D. P., Aronow, P. M., Bergan, D. E., Greene, P., Paris, C., & Weinberger, B. I. (2011). Does knowledge of constitutional principles increase support for civil liberties? Results from a randomized field experiment. *The Journal of Politics*, 73(2), 463–476.
- Greiner, D. J., & Pattanayak, C. W. (2011). Randomized evaluation in legal assistance: What difference does representation (offer and actual use) make? *The Yale Law Journal*, 121, 2118.
- Grossman, G., Humphreys, M., & Sacramone-Lutz, G. (2014). “I wld like u WMP to extend electricity 2 our village”: On information technology and interest articulation. *American Political Science Review*, 108(3), 688–705.
- Haim, D., Ravanilla, N., & Sexton, R. (2021). Sustained government engagement improves subsequent pandemic risk reporting in conflict zones. *American Political Science Review*, 115(2), 717–724.
- Hallsworth, M., List, J. A., Metcalfe, R. D., & Vlaev, I. (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, 148, 14–31.
- Harrison, G. W., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4), 1009–1055.
- Heckman, J. J. (2005). The scientific model of causality. *Sociological Methodology*, 35(1), 1–97.
- Heuer, L., & Penrod, S. D. (1989). Instructing jurors: A field experiment with written and preliminary instructions. *Law and Human Behavior*, 13(4), 409–430.
- Heuer, L., & Penrod, S. (1994). Juror notetaking and question asking during trials: A national field experiment. *Law and Human Behavior*, 18(2), 121–150.
- Ho, D. E., & Ouellette, L. L. (2020). Improving scientific judgments in law and government: A field experiment of patent peer review. *Journal of Empirical Legal Studies*, 17(2), 190–223.
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>. 1990. *Prospect theory: An analysis of decision under risk. Rationality in action: Contemporary approaches*. New York: Cambridge University Press.
- Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, 98(6), 1325–1348.
- Karim, S. (2020). Relational state building in areas of limited statehood: Experimental evidence on the attitudes of the police. *American Political Science Review*, 114(2), 536–551.
- Kassin, S. M., Russano, M. B., Amrom, A. D., Hellgren, J., Kukucka, J., & Lawson, V. Z. (2019). Does video recording inhibit crime suspects? Evidence from a fully randomized field experiment. *Law and Human Behavior*, 43(1), 45.
- Kimbrough, E. O., Smith, V. L., & Wilson, B. J. (2008). Historical property rights, sociality, and the emergence of impersonal exchange in long-distance trade. *American Economic Review*, 98(3), 1009–1039.
- Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S., & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica*, 79(3), 651–692.
- Kőszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences*. *The Quarterly Journal of Economics*, 121(4), 1133–1165. <https://doi.org/10.1093/qje/121.4.1133>
- Kyriacou, A. P. (2016). Individualism–collectivism, governance and economic development. *European Journal of Political Economy*, 42, 91–104.
- Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2), 153–174.
- List, J. A. (2003). Does market experience eliminate market anomalies? *The Quarterly Journal of Economics*, 118(1), 41–71. <https://doi.org/10.1162/00335530360535144>
- List, J. A. (2004a). The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics*, 119(1), 49–89.
- List, J. A. (2004b). Testing neoclassical competitive theory in multilateral decentralized markets. *Journal of Political Economy*, 112(5), 1131–1156. <https://doi.org/10.1086/422564>
- List, J. A. (2007). On the interpretation of giving in dictator games. *Journal of Political Economy*, 115(3), 482–493.

- List, J. A. (2024). Optimally generate policy-based evidence before scaling. *Nature*, 626(7999), 491–499.
- List, J. A. (2025). *Experimental economics: Theory and practice*. University of Chicago Press, forthcoming.
- Marx, B., Pons, V., & Suri, T. (2021). Voter mobilisation and trust in electoral institutions: Evidence from Kenya. *The Economic Journal*, 131(638), 2585–2612.
- Ménard, C., & Shirley, M. M. (2022). *Advanced introduction to new institutional economics*. Edward Elgar Publishing.
- Mujcic, R., & Leibbrandt, A. (2018). Indirect reciprocity and prosocial behaviour: Evidence from a natural field experiment. *The Economic Journal*, 128(611), 1683–1699.
- Mundell, R. A. (1961). A theory of optimum currency areas. *The American Economic Review*, 51(4), 657–665.
- Norenzayan, A., & Shariff, A. F. (2008). The origin and evolution of religious prosociality. *Science*, 322(5898), 58–62.
- North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.
- North, D. C., & Weingast, B. R. (1989). Constitutions and commitment: The evolution of institutions governing public choice in seventeenth-century England. *The Journal of Economic History*, 49(4), 803–832.
- Olken, B. A. (2010). Direct democracy and local public goods: Evidence from a field experiment in Indonesia. *American Political Science Review*, 104(2), 243–267.
- Peisakhin, L., & Pinto, P. (2010). Is transparency an effective anti-corruption strategy? Evidence from a field experiment in India. *Regulation & Governance*, 4(3), 261–280.
- Peyton, K., Sierra-Arévalo, M., & Rand, D. G. (2019). A field experiment on community policing and police legitimacy. *Proceedings of the National Academy of Sciences*, 116(40), 19894–19898.
- Piketty, T. (2014). *Capital in the twenty-first century*. Harvard University Press.
- Porter, E., & Rogowski, J. C. (2018). Partisanship, bureaucratic responsiveness, and election administration: Evidence from a field experiment. *Journal of Public Administration Research and Theory*, 28(4), 602–617.
- Schlager, E., & Ostrom, E. (1992). Property-rights regimes and natural resources: A conceptual analysis. *Land Economics*, 68, 249–262.
- Schweinhart, L. J. (1993). *What the high/scope perry preschool study reveals about developmental transitions and contextual challenges of ethnic males*. <https://eric.ed.gov/?id=ED371257>
- Senderey, A. B., Ohana, R., Perchik, S., Erev, I., & Balicer, R. D. (2022). Encouraging COVID-19 vaccination through behaviorally informed reminders: Results from a national randomized field experiment in Israel. *Behavioral Science & Policy*, 8(1), 25–32.
- Subramanian, A., & Kumar, P. (2024). Property rights, factor allocation and household welfare: Experimental evidence from a land titling program in India. *Journal of Development Economics*, 167, 103238.
- Sutter, M. (2021). Does market experience eliminate market anomalies? (by John A. List). In *The art of experimental economics* (pp. 176–181). Routledge.
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- Thompson, N., Flanagan, B., Richardson, E., McKenzie, B., & Luo, X. (2022). *Trial by internet: A randomized field experiment on Wikipedia's influence on judges' legal reasoning*. Available at SSRN 4327890
- Torgler, B. (2013). A field experiment in moral suasion and tax compliance focusing on underdeclaration and overdeduction. *FinanzArchiv/Public Finance Analysis*, 69, 393–411.
- Valfort, M. A. (2020). Anti-Muslim discrimination in France: Evidence from a field experiment. *World Development*, 135, 105022.
- Voigt, S., Gutmann, J., & Feld, L. P. (2015). Economic growth and judicial independence, a dozen years on: Cross-country evidence using an updated set of indicators. *European Journal of Political Economy*, 38, 197–211.

- Williamson, O. E. (1967). Hierarchical control and optimum firm size. *Journal of Political Economy*, 75(2), 123–138.
- Williamson, O. E. (1988). Corporate finance and corporate governance. *The Journal of Finance*, 43(3), 567–591.
- Worthy, B., John, P., & Vannoni, M. (2017). Transparency at the parish pump: A field experiment to measure the effectiveness of freedom of information requests in England. *Journal of Public Administration Research and Theory*, 27(3), 485–500.
- Wright, B. R. (2018). Field experiments in religion: A dream whose time has come. *Journal for the Scientific Study of Religion*, 57(2), 193–205.
- Wright, B. R., Wallace, M., Bailey, J., & Hyde, A. (2013). Religious affiliation and hiring discrimination in New England: A field experiment. *Research in Social Stratification and Mobility*, 34, 111–126.
- Yoeli, E., Hoffman, M., Rand, D. G., & Nowak, M. A. (2013). Powering up with indirect reciprocity in a large-scale field experiment. *Proceedings of the National Academy of Sciences*, 110-(supplement_2), 10424–10429.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Index

A

- Academic literature, 91, 137, 511
- Accademia dei Segreti*, 918
- Access costs, 908, 909
- Accountability, 20, 186, 192, 203, 205, 211, 212, 277, 653, 664, 666, 719, 730, 731, 749, 750, 929, 933, 1011, 1092, 1095, 1129–1131, 1138
- Acemoglu, D., 212–216, 225, 228–230, 241, 242, 244, 245, 248, 249, 257, 269, 321, 331, 333, 335, 372, 381, 454, 457, 530, 710, 711, 717, 721, 731, 732, 735, 736, 738–741, 744–748, 751, 763, 768, 773, 779, 787, 789, 790, 795, 798, 800–802, 804, 813, 842, 858, 876, 898, 901, 928, 929, 942, 947, 980, 1012, 1013, 1015, 1090, 1092, 1096, 1097, 1102–1104, 1107–1108, 1110, 1122
- Action arena, 80, 82, 84, 86, 88, 95–98
- Action-outcome linkages, 82, 89, 91
- Action situation, 82–95
- Actor, 5, 8, 14, 16, 26, 29, 53, 54, 66, 67, 82–84, 89, 91, 96, 107, 112, 119, 134, 135, 166, 169–171, 180, 211, 228, 237–239, 245, 248, 249, 252, 253, 257, 258, 271, 278, 285, 321, 324, 328, 329, 336, 392, 410, 531, 536–538, 554, 559, 664, 707–710, 716, 720, 722, 732, 747, 762, 804, 814, 841–843, 886, 889, 929, 932, 933, 947, 987, 994, 1014, 1062, 1070, 1081, 1128
- “Actor-Centered Institutionalism” framework, 81
- Adaptability, 288, 360–362, 366, 368–371, 374, 380–382, 481, 489, 661, 943
- Adaptation, 51, 52, 54, 55, 57, 58, 60, 67, 68, 241, 254, 382, 393, 397, 398, 406, 407, 410, 420, 421, 424, 425, 427–432, 435, 436, 451, 480, 482–485, 487, 488, 490, 491, 514, 534, 556, 560, 602, 604, 606, 630–634, 642, 644, 657, 659, 660, 707, 722, 727, 745, 750, 751, 765, 766, 780, 781, 884
- Administrative regimes, 12, 490, 492–493, 495, 497
- Agency costs, 280, 508–511, 513, 515, 525, 654
- Agenda powers, 144–146
- Aggregation rules, 88–90
- Agriculture, 7, 14, 64, 187–189, 204, 241, 343, 408, 464, 674, 675, 677, 678, 680, 681, 683–684, 690, 691, 693, 694, 704, 730, 732, 737, 740–742, 744, 796, 797, 888, 900, 933, 957, 966, 1081, 1091
- Aid and institutions, 749
- Airlines, 409, 452, 598, 606, 623, 638, 639, 644
- All-Civic equilibrium, 883
- Alliances, 110, 189, 451, 458, 484, 486, 489, 492, 523, 525, 547, 632, 946
- Allocation rules, 93
- American blockholders, 518
- American legitimacy light, 524
- Analytical narratives, 751, 1101, 1102
- Anarchy, 186, 238, 242, 243, 796, 928
- Anglo-American legal practice judges, 276
- Anhui, 940
- Anticorruption campaign, 937
- Anticorruption policy, 1002, 1016, 1019
- Anti-expropriation institutions, 928, 929, 946, 947
- Anti-takeover laws, 511, 524
- Antitrust, 59, 64–65, 179, 585
- Archaic proto-state, 873–875, 883, 885, 886, 889–891
- Aristotle, 238, 239, 248
- Arrondissements, 108, 109, 117

- Asia, 132, 135, 145, 147, 149, 511, 610, 711, 739, 741, 742, 746, 747, 758, 759, 856, 899, 956, 958, 967
- Asset holders, 166, 322, 481, 831–833, 837, 839–841, 847, 848, 852
- Asset specificity, 54, 56, 58–60, 62, 63, 420, 422–427, 434, 436, 439–442, 448–454, 456, 458–465, 481, 487, 496, 537, 539, 546, 547, 1064
- Association for the Study of Religion, Economics, and Culture (ASREC), 76, 954
- Attributes, 53–56, 58, 78, 83, 84, 88, 91, 92, 94, 95, 133, 241, 320, 367, 392–394, 396–397, 419–421, 423–425, 427, 431, 433, 436–441, 443, 450, 454, 455, 461, 479, 480, 482, 496, 537, 551, 674, 676, 683, 686, 687, 692, 815, 819, 937, 1128, 1129
- Audiencia*, 197, 198
- Aufklärung*, 902
- Authoritarian, 8, 133, 134, 136, 186, 187, 209–230, 239, 240, 242, 245, 247, 249–251, 254, 258, 331, 744, 747, 780, 930, 1068
- bargain, 227
- elections, 219, 220
- regimes, 186, 210, 211, 214, 219, 220, 222
- Authoritarianism, 8, 134, 208–211, 220, 222, 250, 251
- Authority rules, 88, 89, 91
- Autocracy, 8, 210, 215, 240, 242, 244, 245, 254, 255, 257, 720, 722, 723, 959, 1092
- Autocratic regimes, 8, 219, 225
- Autogolpe, 151, 152
- Automobile industry, 456, 547
- Autonomous adaptations, 54, 58
- Aversion to the custom, 716
- B**
- Backsliding of reforms, 239, 254–256, 258, 720
- Baconian program, 903
- Baconian theology, 1040, 1041
- Balancing power, 8, 228–229
- Bandwagon effect, 715
- Bangladesh, 609, 641, 711, 991
- Barter system, 42
- Barzel, Yoram, 213, 392, 396, 481, 683, 686, 813
- Bayesian formalization, 1084
- Bayesian persuasion, 220, 221
- Bayesian process tracing, 19, 1060, 1074–1079, 1084, 1085
- Bayes' theorem, 1077
- Behavioral economics, 247, 1135
- Beijing-style market communism, 253
- Beliefs, 3, 5, 26, 29–32, 116, 220, 222, 226, 251, 252, 271, 272, 328, 330, 332, 336, 601, 604, 612, 703, 708, 713, 737, 738, 742–747, 749–752, 761–763, 766, 769, 772, 787–791, 819, 825, 832, 836–838, 841, 842, 845, 899, 901, 907, 912, 919, 930, 931, 939, 979, 993, 994, 1003, 1004, 1033, 1034, 1093, 1094, 1124, 1126, 1127, 1129
- Belief systems, 29–34, 750, 752, 791
- Bentham, Jeremy, 294, 311, 371, 1051
- Better Business Bureau, 824, 828, 829, 987
- Big idea, 5, 49
- Big push, 703, 723
- Bill of Rights, 177, 858
- Biological production, 675
- Biological transmission, 1095
- BlackRock, 527
- Blackstone, William, 295, 296, 305, 899, 900
- Blockchain, 271, 287
- Blockchain-based transfer, 287
- Blockholders, 510, 511, 518
- Board of directors, 62, 63, 85, 506, 508, 510, 516, 524
- Bonanza farms, 685
- Bonus, 109, 623, 626, 628–635, 642, 722, 830
- Boundary rules, 88–90
- Bounded rationality, 12, 53, 83, 248, 421, 437, 476, 477, 480, 538, 541–545, 550, 559, 703, 1123
- Bourgeois convention, 888, 889
- Bourse of Amsterdam, 827
- Brazilian Amazon, 323, 324, 328
- Bribery, 277, 405, 511, 792, 1007–1009, 1018, 1128
- British Enlightenment, 905–907, 910
- British legal framework, 762
- British slave colonies, 774
- Buddhism, 930, 931
- Bundling, 13, 287, 574, 578, 582–583, 586
- Bureaucratic costs, 57, 58, 60, 420, 451
- Bureaucrats, 210, 220, 223, 227, 662, 665, 667, 836, 930–932, 943, 1013
- Business environment, 933, 936, 943–947
- Byzantine Empire, 837
- C**
- Capitalism, 268, 537, 761, 769, 861, 930, 940, 1064, 1096, 1134
- Capitalist systems, 536
- Capital markets, 355, 397, 513, 515, 517, 735
- Cascade-based models, 715
- Cascading effects, 715–717
- Case law, 363, 364, 368, 381, 740

- Case study, 10, 16, 19, 134, 321, 335, 375, 380, 407, 437, 438, 453, 454, 464, 483, 496, 497, 635, 684, 751, 1011, 1059–1085, 1101–1102, 1104, 1110, 1123, 1124
 advantage of, 1061, 1063
 individual theories, 1067
 process tracing, 1074
 Queralt's theory, 1070, 1071
 typological theory, 1066
- Cash rent contracts, 14, 677, 679, 686, 689–691, 693
- Casting, 111, 257, 779, 1040, 1052, 1108
- Catholic church, 373, 717, 736, 801, 954–956, 967, 1132
- Catholic schools, 967
- Causal identification, 19, 1060, 1061, 1064, 1065, 1110
- Causality, 455, 461, 712, 721, 723, 749, 751, 752, 1047, 1090, 1102, 1110
- Causal relationships, 323, 691, 714, 741, 1069
- Censorship, 211, 220–225, 918, 967
- Central Europe, 136, 717, 731, 805
- Centralization, 118, 202, 204, 528, 633, 713, 875, 929, 930, 947
- Changing climate, 674
- China, 4, 17, 33, 170, 172, 210, 218, 219, 221, 222, 224, 225, 227, 228, 230, 253, 256, 366, 377, 381, 448, 513, 531, 719, 730, 731, 734, 744, 746, 747, 750, 752, 812, 814, 831, 836, 839, 848–852, 860, 906, 916, 927–948, 962, 968, 969, 1008, 1034, 1072
 future, 945–948
 government and business in history, 929–933
 private sector (*see* Private sector development in China)
 state and private sectors (*see* State and private sectors in post Mao China)
- China's National Market Supervision Bureau, 936
- Chinese Communist Party (CCP), 212, 216, 225, 744, 934, 937, 943, 945, 947
- Christianity, 852, 917, 954, 969
- Christians, 119, 373, 742, 956, 963, 966
- Church, 18, 373, 395, 407, 409, 417, 731, 735, 736, 801, 804, 852, 854, 858, 900, 954–956, 958, 959, 961–964, 967, 1079, 1131, 1132
- Civic-Bourgeois-Sharer Game, 886–889
- Civic-Grabber-Sharer Game, 880–884
- Civil conflicts, 711, 1052
- Civil law, 274, 276, 299, 301, 364–372, 374–378, 380, 381, 641, 738–740, 766
- Civil War, 173, 175, 177–179, 250, 255, 800, 801, 805, 831, 834, 836, 858, 993, 1052, 1053, 1081, 1083
- Clarity, 14, 43, 87, 285, 615, 624, 625, 634–637, 643–645
- Classical game theory, 707, 708, 715, 721
- Clientelism, 717, 984
- Climate change, 14, 480, 506, 512, 527
- Closed autocracy, 240, 245, 254
- Coalescing shareholders, 513, 517–519
- Coalitional presidents, 139
- Coalitions, 8, 109, 122, 137–141, 153, 154, 203, 210, 211, 215, 225, 226, 228, 229, 240, 269, 270, 347, 408, 745, 746, 801–803, 819–821, 931, 940, 957
- Coase, Ronald, 5, 6, 8, 10, 17, 37–45, 49, 50, 52, 53, 228, 324, 368, 392, 407, 410, 422, 432, 448, 449, 451, 454, 457–459, 466, 476, 478, 479, 481, 487, 528, 682, 683, 685, 710, 732, 733, 791, 792, 795, 933, 976, 1061, 1135
- Coasean approach, 228, 476
- Coase theorem, 43, 685, 710, 795, 1135
- Coastal adaptation, 884
- Codes of behavior, 713
- Codification, 363–365, 381, 908
- Coercion-constraining institutions (CCIs), 16, 812–814, 831–844
 balancing coercive powers, 833–835
 balancing coercive power with economic power, 837, 838, 840
 origin and dynamics of, 840–844
 ruler with an independent coercive power, 835, 836
 ruler with conditional coercive power, 836
- Coercive state, 872
- Co-evolutionary model, 15, 702, 707, 718, 722, 880
- Coevolution of institutions and culture, 721
- Co-investments, 487
- Collective action, 8, 10, 18, 84, 92, 212, 220, 225–229, 297, 326, 329, 330, 333, 344, 355, 356, 393, 410, 707, 744, 746, 797, 804, 838, 844, 851, 872, 876, 898, 932, 976, 979, 980, 982–984, 986, 990, 991, 993, 994, 1013, 1096
- Collective-choice rules, 96
- Collective conservatism, 716
- Collusion, 13, 600, 610–613, 617, 661, 945
- Colonial heritage, 732, 737–742
- Colonial powers, 711, 738–741, 1092
- Colonization, 10, 244, 332, 360, 366, 371, 372, 740, 775, 781, 836

- Combine harvester, 680
- Commitment, 13, 14, 51, 56, 58, 60, 62, 82, 116, 148, 195, 213–218, 221, 228, 267–289, 321, 331, 333, 405, 441, 442, 453, 488, 491, 537, 539, 545, 547, 549, 555–558, 560, 653, 658, 660, 677, 710, 731, 765, 781, 793, 795, 811–861, 976–979, 987, 1003, 1070, 1101, 1123
- Common law, 65, 283, 296–299, 301, 307, 308, 365–372, 374, 375, 377, 378, 381, 641, 685, 693, 736, 738–740, 857, 1038, 1042, 1047
- Common pool, 7, 9, 77, 81, 90, 92, 94, 95, 161, 164, 165, 169, 171, 174–177, 179, 215, 295, 297, 298, 308, 309, 312, 339–356, 987, 1061, 1065
- Common property, 9, 76, 88, 94, 191, 295, 341–351, 353, 355, 356, 1125, 1127
- Commons arrangements, 320, 324–326
- Communism, 32, 33, 253, 791, 889
- Community responsibility systems (CRS), 734, 827, 857, 861
- Comparative existence, 1103, 1104, 1110
- Competition, 3, 7, 14, 27, 28, 30, 40, 41, 44, 50, 54, 64, 96, 105–123, 154, 161–163, 166–168, 177, 179, 180, 202, 222, 246, 269, 274, 281, 285–287, 325, 326, 334, 341–343, 353, 356, 371, 394, 397, 398, 401, 407, 422, 435, 452, 455, 456, 485, 495, 514, 521, 527, 539, 541, 556, 568, 569, 571–573, 575, 578–581, 583–586, 588, 604, 606–610, 656, 657, 660, 663, 665, 666, 668, 683, 704–706, 734–736, 744, 745, 748, 750, 766, 790, 792, 822, 827, 830, 836, 837, 851, 852, 857, 880, 881, 883, 884, 934, 936, 941, 945, 946, 955, 956, 968, 985, 1070
- authority, 585
- policy, 495, 568
- Competitive exploitation, 340
- Complementarity, 14, 522, 631, 637, 645, 668, 705, 906
- Complements, 48, 521–522, 531, 548, 568, 632, 786, 831, 989, 1031
- Complexity, 9, 19, 74, 245, 269, 279, 281, 299, 300, 305–308, 408, 420, 424, 439, 452–454, 456, 487, 495, 497, 541, 556, 636, 653, 685, 693, 820, 831, 1102, 1103
- Compression of history, 1106–1107, 1110
- Conditional modernization theory, 250, 723
- Conflict, 14, 16, 50, 61, 93, 130, 134, 137, 139, 140, 144, 145, 153, 154, 173, 175, 191, 214, 274, 285, 288, 297, 310, 312, 313, 321, 327, 334, 335, 342, 347, 350, 353, 356, 361, 363–365, 367, 369, 382, 400, 404, 420, 427, 432, 434, 435, 437, 448, 482, 485, 520, 543, 544, 634, 711, 718, 735, 745, 746, 768, 773, 790, 793–796, 798, 799, 801–803, 805, 833, 834, 836, 837, 842, 847, 851, 854–857, 881–884, 886, 887, 890, 956, 995, 1011, 1052, 1083, 1092, 1095, 1103, 1129, 1133
- Conflict-prone culture, 718
- Confucian, 930–932, 934, 939, 941, 962, 969
- Confucianism, 744, 930–932, 955, 962, 969
- Confucian moral codes, 931, 932, 934
- Conservation, 14, 304, 311, 312, 343, 346, 351, 352, 355, 356, 676, 691–693, 1132
- Constitutional-choice rules, 96
- Constitutional democracy, 199–203, 205
- Constitutional stability, 171, 172
- Constitutions, 2, 16, 31, 98, 134, 142, 144, 151, 171, 172, 193, 329, 332, 731, 748, 766, 775, 777, 1034
- Consumer experience, 576
- Contract-enforcement institutions (CEIs), 16, 358, 716, 729, 812, 814–831, 843–852, 854, 857, 858
- CCIs (*see* Coercion-constraining institutions (CCIs))
- demand for public-order and designed CEIs, 845
- designed CEIs, 824–831
- historical development of markets, 849–859
- market expansion, 849
- merits, 816
- organic CEIs, 817–819, 821
- political development, 849
- political institutions, 849–859
- private-order CEIs, 817, 824–831
- public-order CEIs, 821, 823, 824
- supply of public-order and designed, private-order CEIs, 846–848
- Contracting institutions, 735, 928, 929, 946
- Contract law, 9, 55, 56, 269, 270, 272–278, 280–286, 288, 434, 480, 484, 507, 522–523
- Contracts, 4–6, 11–17, 28, 41, 42, 45, 53–55, 59, 63, 176, 190, 242, 268, 269, 271, 272, 274–276, 280, 283–285, 299, 349, 362, 364, 371, 375, 378, 392, 418–425, 427–436, 438–441, 448–452, 457–460, 462, 464, 476, 478–488, 490–497, 519, 523, 537, 538, 547, 560, 567–579, 597–617, 622–645, 652–665, 667, 668,

- 673–694, 704, 706, 710, 731–736, 739, 742, 747, 767, 792, 795, 822, 824, 827, 850, 873, 898, 928, 929, 934, 938–941, 946, 962, 979, 980, 982, 996, 1013, 1064, 1066
- formal, 604, 606
- forward, 603
- spot, 603
- Contractual commitment, 267–289, 441, 987
- Contractual design, 51, 449, 457–459, 490, 640, 645
- Contractual incentives, 627, 631, 640
- Contractual relationship, 190, 283, 284, 394, 401, 419, 420, 424, 443, 463, 626, 637, 638, 653, 656, 668, 823, 829, 850
- Contractual rigidity, 655–657, 659, 660
- Contractual technology, 632, 640–642
- Convention, 31, 149, 295, 708, 709, 876, 877, 886–889, 1016, 1132
- Cooperation, 7, 16, 53, 117, 138, 139, 176, 188, 189, 192, 321, 356, 403, 409, 479, 485, 488–491, 520, 539, 547–549, 599–601, 603, 610, 617, 622–624, 631, 634, 636, 642–644, 713, 734, 737, 743, 750, 791, 794–803, 805, 820, 833, 834, 842, 872, 880, 883, 939, 940, 976, 977, 979, 982, 984, 986, 990, 993, 1015, 1127, 1129
- Cooperative adaptations, 54, 57, 58, 60, 660
- Co-option of elites, 711
- Coordination, 7, 9, 12, 40, 41, 52, 96, 105–123, 172–174, 188, 189, 217, 222, 225, 252, 253, 274, 275, 280, 286, 328, 437, 448, 450, 452, 458, 459, 462, 465, 477, 479, 484–494, 541, 548, 569, 570, 633, 667, 708, 721, 737, 793, 804, 820–822, 824, 838, 873, 875, 979, 1005, 1006, 1017, 1053, 1054
- of votes, 107
- Coordination of elites within districts
 - optimal nomination, 116, 117
 - strategic contributions, 115, 116
- Corporate finance, 360–362, 369, 377–381
- Corporate governance, 2, 12, 369, 374, 377, 378, 451, 505–531, 675, 823, 830
- Corporate laws, 12, 510, 520–521
- Corporate lawsuits, 510, 520–521
- Corporate social responsibility, 455, 513, 527, 531
- Corruption, 7, 10, 14, 18, 151, 165, 166, 169, 180, 197, 223, 275, 277, 278, 329, 331, 381, 394, 403, 405, 615, 652–654, 657, 660, 661, 664–668, 747, 749, 823, 851, 857, 861, 936–938, 942–944, 947, 985, 993, 1001–1019, 1064, 1093, 1100, 1128
- norms, 1007, 1008, 1014
- Corrupt political leaders, 719
- Cost-effective finance, 62
- Court, 5, 8, 14, 16, 31, 55, 56, 60, 75, 89, 98, 152, 173, 176, 192–197, 199, 269, 272, 274–283, 285–288, 295, 297, 302, 307, 309, 328, 347, 348, 350, 362–368, 370, 371, 377, 380–382, 409, 430, 441, 460, 480, 482, 490, 491, 493, 581, 600, 602, 613, 622–627, 629, 631–634, 636–642, 645, 736, 739, 740, 742, 747, 751, 794, 813, 816, 821–824, 828, 831, 832, 839, 846, 857, 900, 910, 911, 914, 917, 940–942, 944, 946, 954, 961, 963, 976, 987, 988, 1003, 1011, 1012, 1015, 1035–1038, 1042, 1046, 1048, 1049, 1051, 1124, 1126, 1127
- procedures, 275
- Covenantal theory, 81
- Credible/credibility, 60, 120, 192–194, 221, 283, 323, 333, 402, 465, 624, 630, 633, 643, 817, 982, 984, 1109, 1110
- commitment, 213, 288, 331
- Critical juncture, 19, 244, 745, 746, 786, 789, 802, 804, 1052–1053, 1061–1063, 1084
- Crop share, 14, 674, 677–679, 685–687, 689, 691–694
- Cross-district coordination, 117, 118
- Cross-group network effects, 567, 569–571, 574
- Crowding out, 640–642, 1014
- Crude evolutionary theory, 720
- Cultural and institutional changes, 902, 904, 905, 908, 909
- Cultural arguments, 1094
- Cultural dynamics, 718
- Cultural traditions, 930
- Cultural traits, 704, 712, 788, 964, 969, 1106
- Cultural transmission, 719, 969, 1095, 1108
- Culture, 1, 4, 13, 16–20, 29, 95, 188, 332, 364, 367, 373, 494, 544, 560, 615, 635, 703, 709, 713, 718, 733, 742–744, 762, 897–921, 928, 932, 934, 939, 945, 954, 957, 984, 986, 992, 1031–1037, 1040, 1042–1046, 1048, 1052–1055, 1092–1094, 1099, 1100, 1102, 1104–1106, 1124
- Curiosity cabinets, 917
- Custom combiners, 681
- Custom contracting, 680, 681, 688
- Custom harvesting, 688

D

Darwinian natural selection, 703
 Darwinian process, 703
 Debt, 62, 63, 174, 193, 283, 362, 375, 378, 379, 382, 515, 766, 815, 825, 840, 843, 854, 962, 979, 987, 1070–1073, 1079, 1081, 1094, 1125
 Decentralization vs. centralization, 528
 Decision rights, 55, 451, 459, 482, 483, 486, 487, 491, 631, 633
 Decree authority, 130, 142–155
 Decrees, 7, 135, 136, 139, 142, 143, 189, 1042
 Dedicated assets, 426
 Deep determinants, 1089–1111
De facto property rights protection, 934
 Deferral culture, 744
De jure property rights, 335
 Delegated powers, 193, 194, 197
 Delegation of authority, 13, 93, 144, 433, 623, 633
 Democracy, 6, 8, 106, 112, 132, 133, 135, 136, 138, 147, 150, 154, 199–203, 237–258, 614, 722, 723, 731, 749, 751, 761, 795, 852, 958, 959, 980, 984, 1064, 1068, 1092
 Democratic accountability, 205, 719, 1092
 Democratization, 147, 210, 214, 215, 230, 239, 244, 246, 248, 250, 252, 254, 256, 257, 723, 795, 1092, 1095
 Demsetz, Harold, 38, 63, 296, 321–323, 335, 346, 347, 478, 479, 490, 493, 556, 791, 792
 Deng Xiaoping, 933, 934
 Descriptive norms, 1003–1005, 1009, 1010, 1017
 Designed CEIs, 821, 824, 826–831, 843–848, 850–852, 854, 855, 857
 DeSoto, Hernando, 302, 303
 Despotism, 242, 243, 1108
 Determinants of institutions, 744, 745, 1089–1111
 compression of history, 1106, 1107
 countries to grids, 1099, 1100
 cross-sections to case-studies, 1001, 1102
 cultural determinants, 1093–1095
 data quality, 1105, 1106
 historical determinants, 1092
 independence of observations, 1104, 1105
 natural environments, 1091, 1092
 operationalizing institutional diversity, 1103, 1104
 persistent, 1095–1097
 Developing countries, 90, 180, 403, 610, 653, 655, 658, 732, 745–748, 750, 752, 861

Development, 8, 13, 31, 33, 62, 75, 81–84, 180, 216, 230, 238, 250, 253, 276, 304, 322, 332, 355, 359–382, 541, 546–549, 643, 651–669, 703–706, 713, 721, 722, 729–752, 762, 769, 780, 977, 986, 993, 995
 Dictatorship, 8, 210, 211, 214, 218–222, 224, 226–228, 244, 248–250, 252
 Diffusion theory, 551–553, 555
 Digital ecosystems, 13, 567, 588
 Digital Markets Act (DMA), 583, 589
 Digital platforms, 459, 566, 568, 570, 580, 586–588
 Direct/within-group network effects, 566
 Discretion
 benefits, 662–664
 detrimental effects of, 664–666
 Dishonesty, 1008, 1010, 1018
 Dispute resolution, 5, 55, 60, 191, 269, 272, 282, 285–288, 298, 613, 941, 944, 1133
 Dissemination, 17, 28, 272, 614, 904, 905, 987
 “Divide-and-rule” framework, 228
 Document-topic matrix (DTM), 1039, 1045, 1047
 Dominant stockholders, 507, 508, 511, 513, 516, 530
 Doorstep conditions, 242, 244, 334, 335, 789
 Double commitment problem, 710
 Double complement rule, 139
 Downsian equilibrium, 249
 Dynamic efficiency, 720
 Dynamic political game, 719
 Dynamic trajectories, 720
 Dysfunctional institutions, 133, 719

E

Early English Books Online-Text Creation Partnership (EEBO-TCP), 1035–1037, 1051
 Early marriage, 716
 Econometrics, 19, 222, 333, 437, 454, 481, 712, 1029–1055, 1101, 1105, 1109, 1110
 Economic change, 3, 4, 26, 29–34, 761, 762, 768
 Economic classics of federalism, 161–164
 Economic development, 9, 15, 17, 19, 64, 134, 176, 238, 239, 244, 249, 250, 253, 255, 257, 258, 333, 361, 372, 375, 400, 704, 713, 733, 761, 796, 900, 912, 919, 927–948, 1003, 1032, 1049, 1051, 1090, 1095, 1099, 1107

- Economic growth, 17, 28, 160, 177, 200, 201, 230, 253, 322, 332, 360, 704, 711, 742, 760–764, 766, 792, 861, 898, 899, 955, 956, 960, 961, 1107, 1117–1138
- Economic markets, 346, 976, 978–981, 986
- Economic organization, 27, 28, 48, 52–55, 67, 68, 268, 401, 450, 484, 850, 851, 1118
- Economic policy, 38, 65, 142, 179, 761, 764, 985
- Economic progress, 8, 334, 890, 909, 920, 976
- Economic system, 29, 38–45, 55, 409, 479, 531, 771, 943
- Economic theory, 3, 38, 39, 42–45, 81, 90, 160, 162–166, 218, 268, 422, 581, 614, 708, 720, 722, 762, 786, 872, 1002, 1109
 economic classics of federalism, 162, 163
 market-preserving federalism, 165, 166
 modern economics of federalism, 164, 165
- Effective magnitude, 109–112, 115
- Effective number of votes per voter, 111, 122
- Efficiency, 39, 55, 60, 61, 90, 94, 96, 190, 223, 297, 353, 355, 356, 361, 362, 369, 379–381, 402, 405, 454–456, 601, 606, 652, 654, 668, 705, 711, 720, 745, 793–795, 805, 813, 815, 820, 821, 824, 828, 829, 858, 859, 945, 946, 990, 1015, 1090
- Elections, 6, 98, 110, 111, 113, 114, 117, 138, 140, 141, 152, 173, 201, 210, 219, 227, 239, 240, 244, 245, 249, 252, 258, 376, 656, 659, 717, 736, 773, 775, 777, 780, 929, 985, 992, 994, 1130
- Electoral autocracy, 240, 244, 245
- Electoral coordination
 coordination of votes within districts, 112
 “M + 1 rule”, 112–115
- Electoral cycle, 135, 140–141
- Electoral district, 108, 112, 115, 984
- Electoral fraud, 219, 220, 252
- Electoral institutions, 6, 105–123, 137, 148, 173, 210, 252, 1130
- Electoral persuasion, 107, 119–121
 office-seeking competitors with credible promises, 119
 office-seeking competitors without credible promises, 120
 rent-seeking competitors without credible promises, 120
- Electoral rules
 number of seats per district, tier and segment, 109, 110
 number of votes per voter, 110, 111
 translating votes into seats, 108, 109
- Electoral segment, 108
- Elite dominance, 737–739
- Ellickson's analysis, 305
- Embeddedness, 12, 406, 466, 476, 1067, 1091, 1093, 1096, 1097, 1110
- Empires, 185–205, 230, 852, 853, 965
- Empirical expectations, 1004, 1005, 1009, 1017
- Empirical research, 11, 58, 62, 86, 90, 222, 227, 419, 437, 438, 452–456, 461, 463
- Empirical testing, 58, 67, 768, 1123
- Employment protection, 526
- Encyclopedia, 909, 917
- Endogenous federalism, 170–175
- Endogenous institutions, 715, 718, 765, 788, 792
- Endogenous matching, 676, 690, 691
- Endowment, 321, 332, 335, 342, 372, 373, 376, 377, 710, 732, 737, 738, 740–742, 745, 768, 770, 771, 774, 841, 1070, 1092, 1135
- Enforcement, 6, 8, 9, 16, 87, 88, 189, 190, 193, 203, 240, 268–275, 278, 280, 282–288, 320, 322, 329, 330, 335, 336, 341, 346, 347, 353, 360–362, 372, 375, 378, 379, 381, 393, 399–402, 404–406, 409, 430–432, 454, 479, 482, 490, 491, 520, 531, 600–604, 607, 610–614, 617, 622, 623, 625, 627, 629–632, 636–639, 641–645, 684, 694, 704, 731, 733–, 735, 738, 740, 742, 748, 762, 763, 800, 812–816, 818, 820–831, 838, 850, 859–861, 880, 898, 900, 928, 938, 939, 942, 946, 983, 987–989, 996, 1013–1016, 1018, 1064, 1126–1129, 1133
- Engels, Friedrich, 874, 889
- Engineers, 76, 280, 426, 456, 905, 911, 912, 916
- England, 18, 19, 189, 199–201, 243, 257, 279, 304, 305, 313, 351, 363, 365, 371, 376, 377, 678, 686, 717, 735, 736, 739, 742, 745, 769, 770, 772, 789, 796, 800–802, 804, 849, 851–860, 900, 904–907, 909, 910, 912, 916, 918, 930, 957, 1030, 1033, 1038, 1041, 1042, 1044, 1047, 1049, 1050, 1052, 1053, 1124
- English caselaw, 19, 1036, 1044, 1048, 1049, 1052
- English Reports (ER), 1035–1036
- Enlightenment, 17, 743, 747, 748, 899, 902–910, 913–915, 917–920, 931, 1034, 1118
- Environmental management, 676
- Environmental trades, 312

- Equilibria, 28, 90, 98, 188, 189, 192, 213, 216–217, 242, 251–252, 258, 569, 707, 708, 721, 788, 789, 792–794, 833, 834, 842, 876, 882, 887, 1005, 1062, 1095, 1107, 1110
- Equilibrium dependence, 1095, 1096
- Equilibrium-of-the-game approach, 708–709, 712, 721
- Equipment contracting, 687–689
- Equity, 62, 63, 343, 345, 362, 367, 374–376, 378, 380, 486, 547, 559, 740, 747, 939, 942, 944, 945, 947, 1042, 1071–1073, 1125
- Essentialism, 1110
- Essentialist, 1106
- Ethnicity, 218, 713, 770, 992, 993
- European CEIs, 850
- European enlightenment, 910
- European feudal system, 853
- Evolutionary economics, 537, 552, 554, 722
- Evolutionary equilibria, 707
- Evolutionary game theory, 815, 877
- Evolutionary models, 702, 707, 718, 722, 880
- Evolutionary theory, 476, 720, 721, 791, 794
- Ex ante investment, 419–421, 424, 425, 427–431, 436, 514, 515
- Exchanges, 2, 3, 9, 10, 16, 34, 39, 41–44, 48–50, 52, 54, 55, 63–65, 84, 89, 226, 270, 271, 280, 283, 288, 301, 343, 346, 354, 355, 379, 392, 393, 399, 400, 405, 420, 432, 449, 461, 476, 486, 523, 539, 548, 558, 626, 639, 677, 682, 732–734, 736, 811–861, 881, 898, 961, 963, 976–980, 987, 988, 991, 995, 1002, 1011, 1015, 1035, 1070, 1103, 1118, 1119, 1124
- Excludability, 92–93, 914
- Executive compensation, 519
- Exogenous shocks, 382, 709, 712, 714, 717, 721–722, 788, 790, 803, 887, 920, 1050, 1065
- Expectations, 32, 74, 79, 96, 123, 151, 178, 198, 272, 335, 494, 512, 634, 660, 708, 709, 713, 716, 720, 721, 737, 743, 815, 819–820, 836, 845, 874, 992, 1004, 1005, 1009, 1010, 1017, 1061–1063, 1066, 1080
- Experiment, 18, 20, 81, 84, 220, 222, 455, 544, 610, 613, 712, 745, 771, 781, 794, 842, 882, 978, 1007, 1014, 1018, 1040, 1119–1124, 1126–1137
- Experimental economics, 1119, 1125
- Ex post* adaptation, 55, 425, 427–430, 436, 480, 490, 560
- Ex post* contractual adaptation, 430–431
- Ex post* performance, 419–421, 424, 425, 427, 441
- Expressive power of the law, 1012
- Expropriation, 15, 17, 191, 200, 204, 215, 228, 246, 251, 375, 381, 427, 436, 461, 530, 655, 731, 734, 735, 737, 744, 747, 763, 831, 832, 841, 928, 929, 938, 942–948, 959, 1079
- Extension of franchise, 717
- Externalities, 27, 42, 89, 164, 203, 304, 321, 329, 341, 356, 392, 396, 407, 410, 422, 461, 488, 506, 509, 512, 527, 552, 557, 567, 575, 588, 656, 709, 910, 982, 1062, 1096
- Extractive institutions, 213, 216, 217, 739, 741, 744, 745
- Extreme poverty, 730
- Extrinsic incentives, 18, 988, 994
- F**
- Factor endowments, 332, 335, 732, 737, 740–742, 745, 763, 770, 771, 773, 775, 781, 1070, 1092
- Fallback option, 630, 632, 640, 641
- Farming, 14, 188, 189, 332, 674, 675, 678, 682, 684–686, 688, 691, 694, 797, 840, 873, 878, 880, 883–890, 900, 930, 933
- communities, 188
- economies, 889
- Farmland contracts, 677, 678, 684, 685
- Federalism, 6, 7, 118, 121, 135, 159–180, 203, 736, 929
- Federalism in the United States, 162, 173, 175–179
- Articles of Confederation 1781–1788, 175–176
- democracy in the civil war, 178
- failure, 178, 179
- under the new constitution, 176–178
- Federalist* 46, 172, 178
- Fellow-townsmen, 939, 940
- Female genital mutilation, 716
- Feudalism, 16, 748, 769, 802, 842, 859
- Fictional player, 720
- Field experiments, 20, 220, 222, 613, 645, 962, 1017, 1117–1138
- Finance, 2, 10, 18, 19, 50, 60–65, 92, 115, 119, 122, 174, 175, 331, 360–362, 368–382, 407, 511, 521, 642, 657, 734, 773, 856, 933, 942, 962, 985, 1050, 1051, 1060, 1066, 1070, 1080–1083, 1125

- Financial development, 10, 360–362, 367–382
- Financial institutions, 360, 373, 374, 376, 518, 524, 531, 773, 780, 959–963, 969
- Firm boundaries, 11, 448, 449, 453, 454, 464, 465, 622
- Firms, 4, 5, 10, 12–14, 27, 39, 40, 45, 50, 58, 60, 62, 74, 75, 163, 164, 166, 223, 227, 282, 342, 348, 350, 351, 353, 375–381, 393, 397–403, 407, 410, 418, 419, 422–424, 426, 433–435, 437, 441–443, 448–460, 463–466, 476, 478, 480, 485, 487, 490, 493–495, 497, 511–518, 521, 535–560, 583, 584, 587, 588, 598, 600–604, 612, 616, 622, 623, 633, 636, 639, 642, 658, 663, 665, 666, 674, 675, 681, 690, 718, 719, 744, 824, 826, 830, 854, 928, 933–947, 976, 978–983, 985–987, 1007
- development routines, 543
- market experience, 545
- sales force, 557
- size, 527
- specific skills, 62
- First mover, 270, 272, 555, 556, 911
- First Opium War, 377
- Fiscal decentralization, 170, 172
- Fiscal federalism, 168, 203
- Fiscal interest model, 167–170
- Fisheries, 90, 342–344, 346, 351–355, 884
- Focal level of analysis, 83
- Focal-point effect, 189
- Foot binding, 716
- Forage, 873, 877–880, 885–887, 889, 890
- Forbearance, 296, 311, 480, 482, 483, 486, 487, 490, 494
- Foreign aid, 15, 204, 749
- Foreign direct investment (FDI), 935
- Foreign intervention, 247, 248, 258
- Formal Bayesian analysis, 1079–1084
- Formal contracts, 11, 14, 269–273, 288, 431, 434, 459, 460, 492, 604, 606, 623, 624, 636, 638, 644, 742, 830, 941, 942, 946
- Formal institutions, 2, 4, 16–18, 130, 242, 254, 288, 406, 703–705, 713, 718, 734, 738, 743, 750–752, 789, 790, 816, 927–948, 980, 985, 987–989, 994–996, 1003, 1014, 1016, 1092, 1093, 1095, 1099, 1100, 1122, 1136
- Formal legal system, 302, 402, 614, 1013
- Fortune reversals, 723
- Forward integration into distribution, 59, 61, 62
- Forward-lookingness of agents, 720
- Foster markets, 162, 170, 831
- Franchise, 52, 61, 63–65, 205, 210, 214, 215, 246, 283, 440, 640, 717, 742, 746, 767, 775–780, 1066
- Franchising, 12, 484, 489, 495, 539, 548, 631, 632, 640, 943
- Free-rider problem, 92–93, 510, 840, 1131
- French Civil Code, 363, 366
- French Revolution, 201, 202, 363, 365, 367, 746, 748, 789, 804–806, 844, 1034
- Frequency-dependent behavior, 100, 1010
- Functional institutional equilibrium, 719
- Functional institutions, 718
- ## G
- Game-theoretic model, 121, 211, 227, 716, 1101
- Game theory, 13, 81, 85, 86, 238, 247, 296, 497, 599, 611, 707, 708, 715, 721, 805, 816, 817, 877, 880
- Gatekeeping, 513, 517, 519
- Gender-biased norms, 716
- General Data Protection Regulation (GDPR), 587
- Generalized trust, 976, 984, 988, 990, 1003, 1011
- Gentry, 7, 187, 194–201
- Geography, 372, 713, 732, 735, 796, 841, 899, 917, 993, 1092, 1097, 1098, 1100, 1104
- German code, 364, 366
- German-style codetermination, 512
- Gini coefficient, 883, 890
- Global modernization theory, 252–253
- Glorious Revolution, 213, 246, 736, 789, 800, 801, 852, 858, 900, 930, 1046, 1052, 1124
- Governance, 10–12, 48, 50–56, 58, 60, 62–66, 68, 394, 395, 400, 401, 403, 406, 418–426, 431–443, 452, 454, 456, 458–462, 466, 477, 479, 480, 482, 484, 485, 487–494, 497, 505–531, 558, 559, 565–589, 598–600, 603, 613, 622, 623, 636, 644, 645, 660, 675, 707, 800, 812, 816, 823, 830, 840, 843, 847–849, 851, 853–855, 859–861, 872, 875, 891, 930, 959, 1002, 1003, 1066, 1067, 1070, 1090, 1093, 1100, 1129
- forms, 3
- structure, 432, 450, 455, 456, 461, 462
- Government, 4, 6, 42, 43, 45, 86, 106, 120, 121, 129–154, 160–180, 197, 200, 202, 210, 220, 221, 241, 244, 252, 274, 278, 302, 308, 314, 330, 334, 335, 341, 342, 348, 352, 355, 368, 395, 405, 435, 654,

660–667, 710, 717, 739, 744, 747–750, 763, 766, 769, 780, 800, 860, 900, 901, 904, 920, 930–948, 976–979, 982–986, 988, 991, 993, 996

Governmental opportunism, 655, 661

Government crises, 147–153

Government-employed lawyers, 281

Government expropriation, 932, 942

Gradual approach, 714

Gradual change, 15, 702, 714, 717, 721, 722

Grande Encyclopédie, 908

Great Enrichment, 897–920

Great Famine of 1958–1960, 940

Great Plains, 680, 681, 685

Growth, 15, 17, 20, 28, 31, 33, 76, 160, 170, 177, 179, 190, 200–202, 213, 214, 216, 217, 224, 230, 243, 253, 322, 327, 329, 331, 332, 360, 379, 397, 459, 528, 536, 614, 675, 691, 704–706, 711, 712, 730–733, 735, 737–745, 747, 750, 751, 758, 760–769, 771–773, 780, 781, 792, 801, 802, 813, 829, 841, 853, 855, 857, 861, 898–901, 903, 904, 907, 910, 912, 915, 920, 928–929, 934–938, 940, 942, 945, 947, 948, 954–956, 959–962, 964, 970, 978, 980, 981, 991, 1090, 1097, 1098, 1101, 1107, 1117–1138

Guanxi, 939–942, 946

Gulf shrimpers, 352

H

Hajj, 1132

Harvests and stocks falling, 352

Haute couture, 552, 553

Hawk-Dove-Bourgeois Game, 886

Hawk-Dove equilibrium, 881

Hawk-Dove Game, 881

Hedge funds, 517–519

Heuristic model, 11, 477–481

Hierarchies, 5, 6, 11, 55, 56, 58, 73–99, 241, 312, 432, 436, 450, 451, 455, 477–479, 483, 484, 490, 492, 495, 547, 548, 598, 633, 960, 988, 1134

High-powered incentives, 55, 628, 910–919

High quality institutions, 719

History, 20, 30–32, 34, 134, 173, 175, 179, 187–189, 191–194, 198, 199, 210, 216, 229, 242, 246, 254, 255, 257, 302, 321, 329, 333, 334, 367, 380, 405–410, 524, 537, 548, 635, 674, 711, 721, 723, 735, 740, 744, 746, 781, 786, 791, 814, 855, 856, 861, 917, 920, 929–933, 957, 964,

966, 1015, 1031, 1036, 1040–1042, 1047, 1050, 1053–1055, 1062, 1098, 1101, 1102, 1106–1110, 1131, 1135

Hobbes, Thomas, 17, 795, 831, 872–874, 883, 887, 888, 891, 917, 983

Hofplatz model, 886

Hold-up, 421, 427, 429, 431, 435, 436, 478, 479, 488

Honesty, 279, 718, 719, 820, 1008, 1010, 1014, 1017, 1018

Horizontal cleavages, 711

Horizontal corporate governance, 511–513

Horizontal externality, 423, 424

Horizontal governance, 508, 511, 516, 518, 531

Horizontal mergers, 584, 585

Household responsibility system (HRS), 33, 933

Human actors, 29, 53, 66, 67

Human asset specificity, 62, 426, 453, 462

Human behavior, 5, 49, 74–76, 78, 762, 912, 989, 1003, 1121, 1122

Human capital, 18, 29, 230, 272, 275, 276, 279–281, 284, 326, 373, 426, 427, 433, 450, 453, 456, 459, 461, 464, 530, 634, 665, 736, 742, 744, 752, 763, 770, 775, 779, 793, 832, 838, 842, 912, 916, 963–969, 1090, 1136

Hunter-gatherer bands, 188

Hybrids, 6, 11, 12, 56, 58, 60, 132, 136, 145, 450, 459, 466, 475–498, 632, 854

constitutional formats, 130–133, 135–138, 145, 148, 149, 153, 154

resilience, 489

Hypothetico-deductive method, 1045, 1048

I

Ideological prejudices, 710

Illegal, 86, 277, 304, 395, 397, 402, 404, 584, 587, 613–617, 664, 827, 933, 937, 1011–1013, 1128

Illegal drugs, 402

Illegal organizations, 404

Incentives, 3, 5–8, 10, 14, 18, 28, 44, 50, 55, 57, 77, 79, 80, 84, 87, 89, 92–94, 114, 118–120, 140, 146, 154, 160–164, 166–174, 178, 214, 217, 223, 224, 246, 268, 270–272, 276–278, 280, 286, 320, 322, 327, 334, 342, 346, 348, 355, 356, 365, 397, 399, 421, 423–425, 427, 430–437, 449, 451, 460, 479, 482, 484, 489, 510, 518, 543, 544, 547, 560, 573, 575–578, 587, 604, 622, 625, 627, 631, 632, 635, 663, 684, 685, 694, 719, 736,

- 743, 750, 787, 792, 815, 821, 823, 826, 830, 833, 843, 898, 901, 910–916, 919, 932, 934, 941, 942, 945–948, 958, 983
- Incentive structure, 27–29, 33, 272, 443, 912
- Inclusive institutions, 212, 213, 244, 269, 332, 736, 737, 745, 747, 800
- India, 114, 166, 170, 190, 210, 256, 323, 324, 367, 381, 382, 613, 709, 711, 730, 734, 740, 774, 932, 956, 965, 968, 991
- Indirect network effects, 567
- Individualism, 294, 374, 705, 788, 846, 847, 851, 852, 859
- Individualism-collectivism, 1133
- Individual Transferable Quotas (ITQs), 354
- Induced institutional innovation, 15, 702, 706–707, 721
- Industrial clusters, 747, 939, 940
- Industrial enlightenment, 17, 902–909, 913–920
- Industrial organization, 11, 38, 39, 45, 58, 64, 66, 421, 422, 437, 442, 460, 610
- Industrial revolution, 17, 19, 200, 214, 238, 241, 735, 736, 758, 859, 899–904, 906, 910–913, 919, 920, 967, 1048, 1049
- Industry effects, 554–555
- Inequality, 17, 191, 241, 249, 251, 255, 256, 258, 710, 738, 740, 767, 771, 773, 775, 777, 779, 874, 880, 883, 889–891, 899, 992, 993, 1094, 1096, 1103, 1123, 1132, 1137
in material wealth, 890
- Informal institutions, 2, 4, 16–18, 288, 406, 703–705, 713, 738, 751, 789, 927–948, 995, 1016, 1099, 1100, 1122, 1136
- Informal norms, 9, 18, 27, 30, 32, 33, 250, 281, 295, 305, 406, 409, 678, 746, 790, 928, 1003, 1007–1009, 1011, 1013–1016, 1019
- Informal social systems, 1131–1135
- Information, 2, 4, 8, 12, 28, 52, 75, 90, 91, 145, 162, 210, 211, 219–228, 240, 247, 270, 271, 284, 287, 330, 341, 344, 348, 353, 374, 399, 401, 430, 434, 441, 449, 450, 461, 463, 481, 486, 496, 497, 510, 513, 516, 520, 522, 528, 537, 541–545, 548–550, 552, 555, 558, 559, 576, 588, 611, 612, 623, 644, 656, 704, 709, 715, 733, 734, 751, 762, 775, 777, 813, 815–818, 823–825, 830, 839, 840, 845, 846, 875, 906, 915, 922, 946, 978, 984, 986, 987, 994, 1005, 1009, 1012, 1017, 1035, 1039, 1050, 1055, 1075, 1076, 1080, 1083, 1084, 1128, 1129, 1133
- Informational control, 219, 220
- Informational dilemma, 223
- Information control, 8, 220, 223–225
- Information distribution, 513, 516, 517
- Information manipulation, 211, 220–222, 225, 258
- Information rules, 88, 89, 91
- Injunctive norms, 1003–1010, 1012
- Innovation, 3, 15, 37, 230, 279, 287, 326, 332, 409, 410, 451, 456, 488, 521, 536, 538–542, 545, 548–560, 898–901, 903, 910–912, 914, 919, 920
- Input sharing, 687
- Insecure property rights, 327, 328, 333
- Institution, 9, 15, 17, 78, 79, 199, 212, 275, 280, 282, 344, 510, 513, 516, 519, 521, 523, 536, 548, 550, 558, 667, 708, 709, 714, 715, 743, 760, 766, 770, 788, 791, 795, 797, 803, 819–821, 826, 827, 838, 841, 843, 844, 847, 853, 860, 874, 877, 891, 900, 938–940, 996, 1065, 1121
- Institutional analysis, 3, 5, 60, 73–99, 273, 466, 514, 751, 787, 793
- Institutional analysis and development (IAD) framework, 5, 76, 82–84
- Institutional change, 5, 8, 15–16, 26–29, 31, 33, 212, 244, 322, 325, 332, 347, 406, 701–723, 750, 755, 760, 761, 765, 769, 773, 779, 785–806, 842, 876, 901–910, 920, 945, 969, 1031, 1051, 1126
as an incremental process, 717–719
as gradual process, 714
stepping stone approach, 716
- Institutional constraints, 6, 96, 406, 520
- Institutional divergence, 712, 1103–1104
- Institutional dualism, 705
- Institutional dynamics, 715, 718, 847, 848, 886
- Institutional economics, 1–20, 213, 268, 392–394, 407, 410, 682–684, 721, 730, 953–970, 975–996, 1003, 1059–1085, 1101, 1110, 1118–1126, 1128, 1131, 1132, 1134–1138
barriers to running field experiments, 1122–1125
definitions, 1120
field experiments, 1121
greenfield field experiments, 1135–1137
informal social systems, 1131–1134
laboratory experiments, 1134, 1135
legal systems, 116–1128
mixed methods, 1137
NIE, 1122
observational and experimental, 1121

- Institutional economics (*cont.*)
 optimal firm size, 1122
 political systems, 1128–1130
- Institutional framework, 5, 27, 29, 78, 80–82, 653, 655, 666, 731, 732, 745, 748–750, 752, 763, 767
- Institutional functionalism, 786, 790–794
- Institutional investors, 509, 510, 512, 517–519, 527
- Institutionalized ruling party, 212, 216, 227–230
- Institutional legitimacy, 523–527
- Institutional persistence, 718, 743, 876, 884–889, 1093, 1095, 1097, 1104, 1106–1108
- Institutional quality, 658, 662, 666, 667, 741, 745, 977, 1014, 1090, 1092, 1097, 1099, 1103, 1105
- Institutional reinforcement, 715
- Institutional stagnation, 721, 722
- Institutional structure of production, 5, 37–45
- Institutions, 9, 15, 26, 27, 29, 30, 34, 39, 77–79, 197, 199, 248, 275, 280, 282, 344, 450, 510, 513, 514, 516, 519, 521–523, 548, 550, 558, 667, 709, 714–716, 718, 760, 770, 786, 787, 874, 898–902, 910, 912, 914, 919, 920
 conflict theory of, 795
 of control over, 225
 definition, 79, 212
 formal and informal, 789
 under parliamentarism, 137
- Instrumental variable (IV), 238, 333, 455, 608, 641, 749, 1063, 1097, 1098, 1101, 1102
- Insurance, 123, 283, 309, 310, 399, 400, 440, 459, 518, 558, 574, 691, 829, 854, 873, 878, 900, 962, 1076
- Intangible assets, 379, 427, 433, 453
- Intellectual property, 9, 241, 286, 300, 311, 312, 547, 583, 773, 909, 1051
- Interactions between peers, 566, 713
- Interest groups, 14, 166, 169, 369, 393, 395, 404–409, 653, 656, 722, 736, 792, 945
- Inter-generational transmission of norms, 988
- Inter-group competition, 883, 884
- Interjurisdictional competition, 161–163, 166, 168, 179, 180
- Internal organization, 12, 52, 56, 60, 419, 420, 422, 424, 431–443, 449, 453, 454, 478, 830, 856
- Internal team processes, 543–544
- International collaboration, 948
- International diffusion, 553–554
- Intrinsic motivation, 18, 815, 912, 917, 918, 990, 1004, 1005, 1014
- Intrinsic norms, 977, 987–989, 992–996
- Inventions, 300, 899, 909–914
- Investor protection, 10, 360–362, 370, 372, 374, 377–380
- Irreversibility, 488, 489, 720
- Islam, 954, 955, 957, 965, 969, 970, 1132, 1136
- Islamic banking, 962
- Islamic banks, 962
- Islamic inheritance law, 788, 960
- Islamic law, 788, 959–961, 963
- ## J
- Jewish, 373, 742, 819, 966
- Judaism, 733, 761, 954, 966, 969
- Judges, 197, 274–278, 280–282, 297, 363–369, 464, 520, 538, 677, 739, 816, 823, 835, 854, 911, 942, 961, 1010, 1013, 1127, 1128
- Judicial human capital, 276
- Judicial review, 364, 368
- Judiciary, 31, 268, 269, 274, 328, 348, 363–365, 368, 370, 377, 381, 382, 406, 736, 858, 929, 941, 995
- ## K
- Keju system, 931, 932
- Kinship groups, 717, 733, 946, 977
- Kinship ties, 17, 939
- Knowledge diffusion, 899, 904, 909, 913, 914
- Knowledge economy, 241, 243, 254
- ## L
- Labor, 29, 42, 61, 62, 64, 65, 76, 92, 115, 168, 177, 191, 230, 282, 284, 294, 301, 322, 327, 332, 343, 345, 350, 477, 479, 506, 512, 515, 524, 610, 626, 665, 681, 682, 686–688, 692, 694, 704, 705, 743, 745, 750, 761, 767, 769–771, 773–775, 778–790, 797–800, 822, 874–876, 899, 900, 914, 933, 988, 1101, 1103, 1132, 1133, 1137
- Laboratory experiments, 645, 1121, 1134–1135
- Labor contracts, 681–682, 694
- Land attributes, 674, 676, 683, 692
- Landless Peasants' Movement (MST), 330
- Land titling programs, 302
- Late Paleolithic or Neolithic population, 880, 890

- Laws, 2, 3, 7, 9, 12, 16, 44, 85–87, 89, 175, 189, 201, 244, 275, 278, 282, 283, 308, 332, 342, 348, 360, 361, 367, 368, 374, 377–380, 511, 513, 520, 524, 526, 566, 633, 667, 731, 740, 742, 766, 767, 777, 801, 816, 819, 821, 823, 854, 885, 894, 902, 928, 935, 938, 958, 961, 1003, 1009, 1011–1016, 1019, 1035, 1036, 1038, 1039, 1041, 1042, 1044–1051, 1053–1055
- Laws and private growth, 935–936
- Lawsuits, 510, 513, 520–521, 602
- Lawyers, 43, 75, 77, 274, 278–282, 299, 517, 623, 636, 677, 1040, 1128
- Leadership, 7, 8, 141, 186–190, 192–194, 198, 199, 201, 202, 204, 205, 224, 227, 249, 250, 668, 707, 805, 836, 875, 899, 911, 929, 995
- Left-right political orientation, 524
- Legal
- activities, 402
 - adaptability, 382
 - education, 279
 - institutions, 269, 274, 276–279, 284, 285, 288, 360–363, 372, 374–379, 382, 959–963
 - origin, 360–362, 367–370, 374–377, 380, 381
 - owner, 399
 - profession, 269, 280, 281
 - reasoning, 281
 - and regulatory systems, 821
 - rights, 409
 - rules, 283
 - services, 278, 281
 - status, 409
 - system, 4, 9, 402, 1120, 1122
 - traditions, 10
 - wage, 398
- Legislative rules, 140–141
- Legislative shield, 140
- Legislators, 18, 91, 134, 139, 141, 146, 148, 169, 278, 368, 583, 584, 616, 644, 739, 792, 823, 1012, 1081
- Legitimacy, 14, 18, 93, 133, 134, 330, 507, 509, 511–514, 518, 523–527, 531, 711, 713, 736, 763, 801, 803, 804, 836, 841, 843, 860, 900, 931, 945, 953, 956–958, 965, 1013, 1127, 1129
- Legitimizing, 801, 956–959
- Legitimizing agents, 801, 956
- Legitimation, 801, 955–959, 1110
- Levirate, 717
- Lewisian theory, 705
- Libecap, Gary, 9, 10, 94, 297, 301, 310, 312, 322–324, 326–328, 339–356, 396, 402, 461, 481
- Liberal Democratic dominance, 116
- Linz, Juan, 133–135, 137, 139, 148, 151, 153, 240, 257
- Litigation, 276, 277, 280, 281, 283, 285–288, 309, 348, 371, 434, 437, 636, 645, 655, 734, 911, 940, 1052
- Lobbying, 115, 326, 340, 341, 348, 350, 352, 355, 356, 656, 929, 943, 991
- Local average treatment effects (LATE), 1098
- Local government, 4, 7, 17, 161, 163, 164, 187, 197, 200, 202, 203, 256, 309, 405, 717, 747, 933, 935, 936, 938–947
- Local leadership, 186, 189, 198, 199, 204, 205, 929
- Local order, 187–192, 195, 197, 199
- Local politics, 7, 118, 185–205, 330, 350, 665, 790
- London Exchange, 825
- Long-term contracts, 59, 392, 419, 422, 431, 435, 436, 450, 451, 458, 460, 462, 464, 606, 684
- Long-term trajectories, 720
- Lordship, 196
- Low-cost litigation, 280
- Lower bound approach, 638, 639
- Low-level equilibrium traps, 703
- Low-powered incentives, 55, 61, 434, 910–919
- M**
- M + 1 rule, 113–117, 122
- Machine learning (ML), 19
- effects of caselaw, 1049
 - hypothetic-deductive methodology, 1045
 - hypothetico-deductivists, 1046
 - leveraging text-as-data, 1035
 - measuring novelty using words, 1042
 - topic modeling, 1038
 - VARs, 1050
 - word embeddings, 1043
- Mafia, 395, 404, 613–616
- Maghribi traders, 16, 733, 742, 819, 994, 1101
- Magna Carta, 16, 786, 802–804, 806, 853, 856, 857
- Majority run-off systems, 139
- Make-or-buy decisions, 11, 49, 52, 59, 447–466
- “Make-or-buy” tradeoff, 10, 11, 457, 478, 481
- Malthusian context, 1051

- Management, 4, 40, 63, 88, 90, 130, 188, 220, 224, 277, 281, 282, 284, 286, 287, 297, 298, 308, 342–345, 349, 353, 356, 403, 433, 451, 466, 492, 497, 509, 517, 521, 522, 528, 534, 550, 553, 557, 566–568, 601, 634, 635, 654, 667, 676, 682, 692, 693, 934, 979, 1061, 1127, 1133
- Managerial labor markets, 513, 515, 529
- Managerial practices, 14, 493–494, 497
- Managers, 12, 40, 87, 163, 280, 362, 401, 433, 434, 436, 451, 453, 482, 492, 493, 506–522, 524, 525, 528, 530, 539, 542, 559, 599, 623, 635, 636, 642, 656, 830, 934, 938, 976, 978, 979, 981, 988
- Mandarins*, 7, 187, 194, 195, 197
- Manipulation, 110, 166, 211, 219, 220, 222, 225, 258, 626–632, 662, 664, 691, 693, 820, 1119, 1135
- Manufacturer products, 566
- Market creation, 12, 536
- Market economy, 2, 8–10, 34, 39, 277, 278, 280, 476, 477, 480, 494, 497, 598, 732, 745, 812, 814, 822, 827–829, 831, 839, 845, 851, 852, 934, 1103
- Market exchange, 48, 50, 54, 55, 61, 64, 65, 394, 399, 437, 461, 977, 1103
- Market expansion, 575, 812, 814, 845, 849, 854, 861
- Marketing, 12, 177, 345, 353, 458, 460, 536, 538, 539, 541, 543, 549, 551–556, 559, 560, 674–676
- Market-preserving federalism, 165–166
- Markets, 4–6, 10–12, 16, 41, 48, 51–53, 55, 60, 64, 73–99, 162, 170, 286, 312, 322, 330, 340, 345, 360, 368, 370, 372, 373, 380, 382, 393, 399, 401, 419, 422, 448, 450, 465, 477, 478, 482, 483, 486, 487, 497, 507, 511, 513, 514, 535–560, 575, 588, 589, 598, 603, 610, 613, 614, 688, 709, 730, 733, 734, 737, 738, 760, 765, 766, 770, 795, 800, 801, 811–861, 874, 898, 920, 946, 976, 978–981, 983, 986, 988, 989, 1041, 1071, 1103, 1118, 1122, 1132, 1134
- Market-supporting institutions, 16, 812, 844–849
- Marshall Plan, 723
- Measurement, 11, 151, 353, 393, 398–402, 440, 442, 443, 449, 460, 461, 463, 496, 601, 643, 662, 692, 879, 888, 919, 1062, 1067, 1068, 1094, 1107, 1137
- Media, 118, 221–223, 226, 253, 256, 392, 511, 552, 553, 555, 583, 586, 587, 656, 751, 929, 943, 994, 1130
- Mediterranean cultures, 1094
- Merchants, 365, 572, 733, 734, 742, 745, 748, 801, 819, 820, 827, 828, 832, 838, 843, 851, 857, 930, 932, 957, 966
- Merger control, 583–585
- Meta-punishment, 709
- Methodology, 3, 4, 432, 440, 496–497, 611, 638, 655, 657, 659, 712, 751, 1031, 1045, 1051, 1090
- Metropolitan institutions, 769, 772
- Microeconometrics, 20, 38, 39, 42, 360, 407, 418, 422, 432, 465, 632, 969, 985, 1002, 1003, 1005, 1049, 1050, 1055, 1124, 1125
- Microeconomics, 38, 39, 42, 418, 422, 432, 465, 1002
- Micro-institutions, 476, 477, 495, 497, 498, 707
- Military Revolution, 842, 844, 854, 858
- Mineral rights, 342, 347, 348, 353, 356
- Missionaries, 958, 967, 968
- Mistakes, 247–248, 250, 257, 258, 276, 456, 520, 709, 850
- Mistrust, 18, 737, 744, 976–986, 988, 989, 993, 995, 996
 economic costs of, 979–981
 political costs of, 982–986
- Mobilization, 105–123, 150, 222, 249, 341, 1015, 1130
- Model parliament, 857
- Modern democracies, 8, 201, 204, 239
- Modern economics of federalism, 164–165
- Modern economic theory, 39
- Modernization, 238, 241–244, 249, 250, 252–256, 258, 657, 723, 1071
- Modern natural language programs, 1037
- Modes of organization, 2, 4, 5, 48, 55
- Monopoly pricing, 568, 582
- Monotonous transformation, 722, 723
- Moral hazard, 187, 188, 192–195, 197, 310, 538, 574, 663, 683, 685–689, 692, 694, 749, 818, 900
 costs, 688
 model, 685
- Moral norms, 1003, 1007, 1012
- Movies, 606, 640
- Muerte Cruzada*, 135, 150, 153
- Multihomers, 572, 579
- Multinational diffusion, 554
- Multi-plant economies, 422

- Multiple equilibria, 28, 189, 213, 216–218, 251, 716, 787, 793, 794, 1005, 1095, 1107
- Multiple levels of analysis, 76, 78–80
- Muslim, 819, 833, 839, 843, 848, 852, 860, 948, 956, 958, 960, 961, 965, 966, 1132, 1135
- Mutual causation, 1011–1012
- Myanmar, 711
- N**
- Napoleon, 363, 364, 366, 748, 804, 916
- Napoleonic Code, 364, 366
- Narratives, 20, 615, 713, 714, 722, 723, 751, 1032, 1094, 1101, 1102
- Nash equilibrium, 570, 600, 612, 708, 720
- National Bureau of Economic Research, 44
- National Bureau of Statistics (NBS), 936–937
- Nationally-oriented president, 139
- National politics, 7, 8, 173, 186, 187, 194, 202, 204, 205
- National state, 186, 188, 189, 203, 204, 704
- Nation-building, 711
- Natural experiments, 238, 323, 405, 455, 645, 712, 764, 771, 781, 1049, 1050, 1055, 1101, 1102, 1104
- Natural field experiment, 1119, 1122, 1124
- Natural monopoly, 61, 63–65, 826, 830
- Natural resource endowments, 376
- Natural States, 334, 789
- Neoclassical economics, 3, 4, 18, 48, 51, 90, 392, 406, 421, 463, 685, 792, 1118, 1123
- Netherlands, 200, 603, 735, 736, 739, 742, 759, 911
- Networks, 13, 17, 18, 211, 253, 393, 401, 405, 451, 458, 484, 486, 572, 614, 632, 733, 747, 803, 819, 845, 850, 905–908, 939–941, 946, 977, 987, 989–996, 1034, 1035, 1044
- effects, 13, 566, 567, 569–574, 580, 582, 584, 588
- industries, 64
- New constitution, 176–178
- New economic history, 712, 721
- New institutional economics (NIE), 2–20, 76, 161, 212, 213, 392, 393, 410, 418, 476, 537, 559, 706, 730, 732, 750, 790, 953–970, 975–996, 1003, 1060, 1101, 1109, 1118, 1122, 1123
- New markets
- acquiring experience, 545
- bounded rationality, 542
- commitment, 556
- diffusion theory, 552
- disparity of goals, 542
- firms develop routines, 543
- internal team processes, 543
- international diffusion, 553
- New product development (NPD), 355, 537–544, 546, 547, 549, 558
- New world economies, 332, 771
- No-confidence votes, 131
- Non-contractibilities, 480, 483, 488, 489
- Non-democracies, 8
- Nonvested benefits, 62
- Normative expectations, 1004, 1009, 1010, 1017
- Norms, 2, 3, 5, 9, 16, 18, 27–33, 78, 79, 82, 84, 85, 87, 95, 188, 203, 242, 250, 270, 281, 313, 403, 406, 460, 520, 599, 614, 708, 716, 718, 731, 733–745, 819, 845, 878, 880, 928, 966, 977, 984, 987, 988, 990, 994, 995, 1001–1019, 1064, 1092–1095, 1099, 1100, 1108, 1118–1120, 1122, 1132–1136, 1138
- nudges, 1017, 1018
- of trust, 987–989
- North, Douglass, 2, 3, 5, 6, 13–16, 25–34, 49, 79, 132, 177, 178, 189, 201, 213, 214, 216, 241–244, 246, 322, 332, 334, 351, 372, 400, 477, 478, 489, 530, 675, 678, 685, 691, 713, 731–739, 741–743, 745, 747, 763, 767, 769–772, 775, 778, 781, 787, 789, 790, 792, 796, 798, 814, 816, 825, 831, 858, 874, 879, 898, 900, 901, 910, 928, 929, 938, 946, 947, 976, 979, 994, 1003, 1014, 1015, 1061, 1090, 1092, 1103, 1109, 1118, 1122, 1124
- No-till agriculture, 693
- O**
- Office-seeking competitors with credible promises, 119–120
- Office-seeking competitors without credible promises, 120
- Oil and gas, 297, 342, 348, 350, 352, 355, 356
- “Old Poor Law”, 900
- Oligarchies, 191
- Online dispute resolution (ODR), 285, 287, 288
- Open access, 9, 320, 339–342, 356
- Open-access fisheries, 343
- Open access order, 242, 334, 335, 739, 789
- Operational rules, 96, 97

- Opportunism, 12, 14, 53, 66, 421, 431, 434, 435, 451, 463, 464, 488, 522, 537, 538, 546–550, 556, 558, 560, 571, 654–656, 658, 660, 661, 976, 988, 994, 1002, 1015, 1123
 Opportunistic acts, 719
 Opportunistic behavior, 12, 14, 18, 286, 420, 428, 431, 432, 434, 442, 476, 488, 604, 655, 658–660, 663, 718, 898, 976–983, 987–989, 994–996
 Optimal nomination, 115–117
 Optimal standardization, 298–300
 Organic CEs, 817, 819, 821–823, 828, 831, 845–847, 849, 850, 852
 Organizational mechanisms, 271
 Organizations, 2, 9–13, 26–28, 30, 32, 33, 39, 41, 74, 78, 96, 117, 173, 242, 274, 275, 279, 280, 404, 420, 421, 436, 476–483, 528, 544, 553, 613, 615–617, 633, 656, 682, 694, 749, 821, 822, 829, 838, 841–844, 851, 858, 904–906, 908, 914, 929, 931, 947, 976, 978, 979, 986, 991, 994, 1051, 1064, 1067
 Organized crime, 600, 613–615, 830
 Ostrom, Elinor, 3, 5, 9, 73–99, 295, 308, 325, 341, 344, 356, 749, 751, 872, 914, 976, 987, 1061, 1063, 1065, 1066, 1118, 1123, 1125, 1137, 1138
 Outsourcing, 448, 455, 456, 462, 463, 466, 546, 547, 549, 550, 606, 631, 638, 639, 652
 Override, 136, 145–147, 273
 Ownership, 11, 12, 52, 60, 187, 189, 190, 198, 204, 294, 308, 332, 341, 345, 348, 355, 361, 371, 378, 380, 401, 404, 433, 450, 451, 455, 458, 459, 464, 484, 489, 494, 497, 507–510, 513, 517, 518, 525, 526, 530, 682, 683, 687, 692, 693, 695, 830, 877, 885, 886, 933, 934, 936–938, 944–947, 1125
 dispersion, 525
 separation, 530
- P**
 Package veto, 146
 Pakistan, 613, 711
 Parliament, 18, 131, 134, 137, 142, 143, 154, 199–201, 214, 228, 246, 365, 371, 380, 736, 800, 801, 803, 836, 843, 851, 856, 858, 900, 906, 911, 912, 984
 Parliamentarism, 130, 131, 134, 137, 138, 147
 Parliamentarization of presidential systems, 150–153
 Partial veto, 146
 Particularism, 1008, 1009, 1015
 Partisan compatibility of presidents and assemblies, 138–139
 Party system, 114, 118, 135, 137, 138, 140, 141, 173
 Past interactions, 637–639, 644
 Patents, 17, 854, 906, 909–912, 914, 938, 944, 948
 Path dependence, 12, 30–33, 216–218, 241, 243, 252, 405–407, 410, 541, 546, 548, 550, 554, 557, 560, 1048, 1063, 1065, 1096, 1110
 Patronage, 17, 250, 717, 793, 915, 1073
 Payoff rules, 88, 89, 91
 Peace of Constance, 854
Peninsulares, 197
People, 5, 8, 9, 40, 53, 92, 99, 163, 173, 186–192, 194, 195, 197–203, 205, 296, 300, 313, 328, 329, 405, 507, 515, 516, 537, 542, 549, 552, 558, 610, 703, 708, 716, 718, 719, 733, 734, 742, 743, 768, 769, 817, 824, 846, 874, 875, 890, 899, 903, 907, 911, 916, 920, 931, 933, 934, 940, 941, 943, 944, 960, 977, 980, 987, 993, 1003, 1004, 1006–1008, 1011, 1012, 1014, 1015, 1017, 1034, 1074, 1094, 1105, 1120–1122, 1126–1131, 1133, 1137
People's Daily, 933
 People's Republic of China, 940
 Performance, 2, 4, 6–8, 25–34, 45, 54, 87, 90, 130, 133, 154, 159–180, 195, 205, 253, 270, 284, 321, 329, 365, 404, 406, 420, 421, 427, 430, 433, 435, 436, 456, 463, 476, 493, 494, 497, 529, 537, 542, 558, 616, 623, 626–629, 633, 636, 637, 662–664, 713, 719, 742, 760–762, 764, 765, 780, 781, 818, 822, 827, 898, 910, 934, 945, 956, 980, 984, 991, 996, 1003, 1030, 1049–1051, 1096–1100, 1103, 1110, 1126, 1127, 1130, 1131
 Persistence, 332, 372, 707, 712, 718, 737, 743, 788, 876, 883, 884, 891, 966, 1093, 1095–1100, 1104, 1106–1110
 Personal norms, 1004, 1007, 1008, 1014, 1016, 1019
 Persuasion, 105–123, 219–221, 1015
 Physical and material conditions, 79, 91–92, 94
 Physical asset specificity, 426, 453, 462
 Piece rate, 627–632, 642, 676, 682, 1133
 Pivotal agent, 716
 Platform exclusivity, 578

- Platform governance, 588
Platform liability, 587
Platforms, 1, 4, 11–13, 271, 272, 287, 288, 410, 449, 451, 456, 458–460, 565–589, 884
Plural forms, 494–495
Plural governance, 548
Plurality rule, 121, 139
Podestaria system, 835
Policymaking, 81, 98, 138, 139, 141–142, 145, 153, 161, 163, 167, 168, 179, 382, 758, 1002, 1016, 1110
Policymaking powers of presidents, 141–142
Political accountability, 20, 186, 719, 1131, 1138
Political agency models, 719
Political and adaptability mechanisms, 361, 362, 368–369, 374, 380–382
Political centralization, 713, 875
Political coalition-formation, 226
Political Coase Theorem, 710
Political decentralization, 186, 202, 256
Political development, 201, 223, 814, 831, 832, 843, 849, 851, 860, 957, 969
Political economics, 9, 15, 18, 26, 34, 59, 204, 212, 220, 341, 664, 708, 710, 711, 713, 714, 721–723, 768, 786, 792, 802, 955–959, 969, 995, 1016, 1060, 1075, 1094, 1107
Political incentives, 166–170
 endogenous federalism, 170, 171, 173
 fiscal interest model, 167, 168, 170
Political institutions, 3, 4, 6–8, 12, 16, 166, 167, 175, 200–202, 210, 212, 213, 216, 243, 257, 321, 328, 333, 372, 376, 523–525, 531, 713, 744, 747, 790, 801, 802, 812–814, 820, 831, 840–845, 849–861, 900, 919, 960, 963, 969, 1096, 1102, 1107, 1130, 1134
Political legitimacy, 954
Politically connected firms, 227, 658, 665, 938, 941, 942, 945
Political markets, 6, 334, 978, 979, 983, 986
Political opponents, 14, 652–656, 668
Political parties, 27, 96, 215, 228, 405, 407, 736, 801, 929, 947, 959, 976, 978, 979, 986
Political polarization, 255, 934
Political power, 192, 203, 213, 216, 321, 330, 333, 372, 525, 710, 742, 767, 768, 773, 775, 873, 874, 891, 898, 957–959, 965, 1096, 1131
Political regimes, 8, 230, 237–258, 709
Political status quo, 216, 711
Political systems, 34, 132, 151, 186, 194, 200, 227, 376, 382, 709, 717, 733, 780, 835, 843, 911, 946, 947, 984, 1120, 1128–1131, 1133
Political transaction costs, 333, 342, 356
Populism, 518, 524, 992
Pork-barrel politics, 717
Position rules, 88, 89
Poverty, 15, 76, 730, 738, 982, 1094
Power, 7–9, 16, 18, 19, 61, 63, 87, 93, 118, 133–136, 139, 141, 142, 145, 146, 164, 165, 171, 173, 175–180, 188–200, 202–205, 210–230, 239, 240, 243, 245–247, 250, 275, 283, 321, 330, 332–335, 341, 365, 370, 372, 381, 418, 419, 431–433, 435, 436, 439, 448, 449, 510, 525, 570, 576, 585, 602, 604, 661, 685, 709–712, 730, 736, 737, 745, 746, 768, 773, 796, 800, 802, 803, 813, 814, 823, 830–861, 873–875, 879, 891, 900–902, 906, 917, 919, 920, 930, 932, 934, 943, 944, 947, 956–959, 965, 984, 1012–1014, 1032, 1054, 1065, 1067, 1070, 1083, 1094, 1096, 1107, 1119, 1124, 1131, 1138
Pre-announcements, 556
Predictions, 58–59, 82, 90, 222, 246, 258, 354, 369, 370, 438, 463, 464, 546, 557, 571, 584, 601, 616, 625, 631, 637, 645, 794, 798
Preference falsification, 716
Present discounted value (PDV), 193, 637
Presidentialism, 118, 130, 131, 133–135, 138–141, 147, 150
Presidential or parliamentary government, 6
Presidential reelection, 148
Presidential vetoes, 144, 145, 147
Pre-state communities, 873
Price competition, 573, 663
Price controls, 331, 392, 396, 398, 399, 401
Price discrimination, 61, 392, 422, 423, 573, 574, 582, 604, 612
Price parity clauses, 568, 575, 578, 580–581, 586
Price theory, 38, 39, 50, 422
Principal-agent framework, 18, 685, 687
Principal-agent model, 220, 663, 689, 706, 721
Principal-agent problem, 485
Principal and agent, 14, 268
Private farming, 933
Private firms, 86, 622, 666, 744, 747, 877, 928, 933–947
Private international law (PIL), 269, 285

- Private investment, 203, 743, 744, 928, 933, 937, 938, 947
- Private land use controls, 306, 307
- Private law, 44, 269, 281, 285, 288, 364, 520
- Private-order institutions, 5, 50, 51, 817, 824–831, 843, 859, 860, 919
- Private property, 8–10, 17, 94, 176, 196, 205, 320, 325, 330, 341, 342, 344, 346, 356, 360, 362, 365, 368, 374, 376, 380, 381, 732, 735, 741, 787, 791, 871–891, 935, 938, 942, 944
- Private protection of their property rights, 216
- Private sector development in china
contracts enforcement, 939–942
protection against expropriation, 942–944
- Private sectors, 17, 405, 598, 654, 655, 659, 928, 929, 932, 933, 935, 937–946, 948, 995, 1008
- Process tracing, 19, 751, 752, 1060, 1061, 1065, 1066, 1068, 1069, 1073–1079, 1083–1085
- Procurement, 14, 52, 57, 449, 456, 458, 493, 608, 611, 637–639, 654, 659, 662–668, 1003
- Production stages, 451, 674, 880
- Product markets, 51, 52, 513–515, 521, 522, 524, 525, 529, 583
- Professional control, 279
- Professionalism and norms, 520
- Professional organization, 275, 279, 280
- Prohibition, 65, 148, 280, 365, 395, 404, 575, 584, 586, 717, 769, 788, 791, 853
- Propaganda, 211, 212, 219–223, 920, 964
- Property
constraints on use, 305–308
informal property, 295, 296, 311–313
initial acquisition of, 298
and “personhood”, 303, 304
transfer of property, 301–303
- Property law, 9, 17, 293–313, 377, 583, 935, 936, 938, 941
categorization in, 299, 301
civil law and common law systems of, 299
as formal system, 295, 296, 305
land use controls, public and private, 306, 307
nuisance law, 305, 308
numerus clausus, 299–301
and personhood, 304
politics of change, 297
recording and registration systems, 301
regulation of property, 310
law scholars, 306
takings claims, 309
- Property rights, 6, 9–12, 16, 17, 20, 180, 189, 190, 192, 195–198, 200, 204, 320–326, 328, 330, 331, 341–347, 351–356, 476, 478, 481–488, 493, 497, 530, 538, 682–683, 718, 735, 873, 928, 930, 932, 933, 935, 937–940, 942–944, 1122
- “Property rights” (PR) approach, 418
- Proportionality, 6, 107–109, 115, 121–123, 439, 1010
- Prorating, 349–351
- Protection against expropriation, 928, 942–945
- Protection of property rights, 216
- Protestant, 371, 961, 962
- Protestant Ethic* hypothesis, 954
- Protestantism, 18, 761, 804, 955, 964, 965, 969
- Protestant Reformation, 735, 736, 742, 963
- Proto-states, 873, 875, 883, 885, 886, 889, 891
- Proximate determinants, 333, 1090
- Proxy fights, 517–519, 527
- Public contracts, 14, 287, 652–661, 664, 667–669
renegotiations, 657–661
specificities, 653–657
- Public firms, 508, 509
- Public goods, 7, 81, 92, 96, 107, 119–121, 123, 160–164, 166, 169, 170, 174, 176, 177, 188, 202, 203, 246, 250, 278, 279, 281, 285, 287, 613, 707, 717, 795, 797, 836, 837, 839, 844, 850–852, 872, 881, 891, 898, 942, 967, 982, 984–986, 989, 994, 1005, 1080, 1081, 1083, 1096
- Public governance, 586–588
- Public-order CEIs, 50, 821–824, 826, 827, 846–852, 857, 858
- Public policy, 48, 50, 53, 59, 62, 64–66, 68, 166, 437, 511, 985, 986, 1017
- Public procurement, 14, 615, 653, 654, 657, 659, 662, 664–669
- Public scrutiny, 652, 653, 655
- Public service, 93, 163, 188, 190, 197, 202, 203, 311–312, 403, 654, 659, 668, 780, 943, 1016, 1128
- Puritanism, 1039, 1052
- Q**
- Qin Dynasty, 930, 931, 947
- Quasi-central banks, 765
- Quasi-parameters, 714–715
- Queralt’s theory, 1070, 1071, 1073, 1074, 1079, 1080, 1082–1084

R

- Radical change, 241, 709, 710, 715, 717, 749, 1053
- Radin, Margaret, 294, 303, 304
- Radius of trust, 977, 979–985, 987, 990
- Raised farming platforms, 884
- Randomization, 323, 1121, 1124, 1126, 1127
- Reciprocity, 294, 401, 486, 994, 1011, 1132, 1133
- Reelection, 27, 135, 148–150, 154
- Reformation, 294, 373, 735, 736, 742, 790, 801, 804, 806, 918, 955, 957, 958, 963–965, 967, 968
- Regression discontinuity designs, 455, 662, 712, 1101
- Regulation, 2, 7, 10, 12, 13, 38, 40, 43, 44, 63–65, 85, 86, 94, 102, 164, 177, 179, 263, 271, 308, 309, 330, 339–358, 375, 378, 391–410, 436, 442, 455, 460, 464, 489, 493, 503, 509, 516, 565–589, 661, 662, 665, 668, 748, 800, 823, 835, 868, 895, 925, 939, 974, 983, 985, 1003, 1092, 1099, 1115
- Relational contracts, 4, 13, 14, 272, 273, 288, 431, 460, 480, 483, 485, 487, 490, 491, 496, 497, 597–604, 606, 608, 610, 613, 615, 617, 622, 624, 625, 627, 629, 630, 632–643, 645, 653, 816, 980
- Relational enforcement, 625, 632, 641
- Relational incentives, 625, 627–631, 632, 635, 640, 641, 645
- Relationships
 long-term, 598, 604, 606–608, 614
 value of, 603–605
- Relationship-specific investments, 11, 418, 421, 424, 425, 427, 428, 431, 435, 450, 453, 454, 460–462, 465, 934
- Religion, 4, 17, 18, 194, 241, 244, 373, 376, 406, 709, 733, 761, 762, 953–970, 993, 994, 1040, 1052, 1120, 1131, 1132, 1134, 1135, 1140
- Religious authorities, 954, 956–959, 965, 966
- Religious institutions, 17, 955–970, 987, 994, 1131
- Religious legitimacy, 956–958, 961
- Religious legitimization, 957, 958
- Remediableness, 66
- Reneging temptation, 638, 639
- Renegotiations, 14, 654, 657–661, 663, 667–669
- Rents, 7, 9, 89, 95, 119–121, 165–170, 180, 187, 188, 192–194, 213, 215, 216, 223, 226, 227, 230, 270, 327, 334, 343, 345, 347, 349, 397, 408, 420, 426–428, 430, 448, 450, 461, 462, 464, 515, 608, 611, 640, 655, 676, 710, 772, 920, 943, 984, 988, 1014, 1018, 1080
- Rent-seeking competitors with credible promises, 120–121
- Representative assemblies, 187, 199–203
- Reputation, 17–19, 54, 95, 120, 150, 174, 192, 215, 217, 270, 273, 286, 425, 431, 488, 492, 554, 557, 558, 603, 613, 614, 663, 682, 684, 685, 742, 816, 818–820, 822, 825, 827–830, 838, 858, 910–912, 915, 928, 939, 940, 943, 946, 987, 989, 1005, 1018
- Resource-based view (RBV), 11, 451, 463, 479
- Resource use, 309, 311, 312, 320, 321, 328, 345–347
- Return on assets (ROA), 936
- Revealed preferences approaches, 1094
- Revenue sharing system, 7, 169, 170
- Reversals and uncertainty, 936–938
- Risk aversion, 51, 52, 423, 676, 685, 689–693
- Risk neutral contracting parties, 685
- Risk-taking, 374, 559, 590, 945
- Robinson, J.A., 212, 214, 225, 230, 241, 242, 245, 248, 249, 269, 331, 333, 335, 372, 486, 711, 717, 721, 735, 736, 738, 739, 741, 744–746, 763, 779, 781, 787, 789, 790, 795, 800–802, 813, 842, 876, 904, 929, 930, 947, 1015, 1096, 1097, 1102, 1104
- Role of government, 55, 347, 368, 928
- Role of individuals, 16, 721, 804–805
- Roman law, 190, 362, 363
- Rudiments, 53–59
- Rule configurations, 88–91, 99
- Rule of law, 8, 17, 87, 240, 242, 286, 328, 334, 336, 719, 730, 731, 736, 747, 813, 838, 898–900, 928, 936, 941, 946–948, 980, 985, 988, 1013, 1014, 1016, 1095, 1131, 1138
- Rule-ordered relationships, 87
- Rules, 2, 3, 5–7, 9, 10, 12, 14, 16, 27, 29–33, 44, 50, 63, 74, 77–80, 82, 84–99, 106–108, 111, 119, 121, 122, 130, 135, 137–141, 148, 150, 151, 154, 166, 171, 172, 175, 180, 190, 191, 211–213, 217, 226, 239, 251, 255, 258, 273–279, 282–287, 293, 297, 299, 300, 302, 307, 336, 347, 349, 351, 353, 364, 365, 374, 375, 404, 406, 408, 481, 488, 491, 494, 520, 522, 523, 527, 584, 587, 588, 613, 654, 655, 664, 666, 667, 685, 693,

- 703–705, 707–709, 713, 731, 734,
737–739, 742, 743, 746, 749, 750, 762,
766, 787, 794, 795, 820, 823, 824, 831,
835–837, 844, 854, 858, 873, 874, 877,
908, 910, 930, 935, 938, 943, 946, 984,
1064, 1095, 1096, 1121, 1122
- Rules-in-form, 79, 87
- Rules-in-use, 79, 86, 87, 89, 90, 94
- S**
- Samaritan's dilemma, 748, 750
- Scaffolds, 32, 33
- Scale, 78, 163–165, 217, 224, 243, 308, 313,
323, 334, 343–345, 353, 398, 400, 424,
449, 479, 528, 556, 567, 570, 574, 588,
682, 732, 761, 769, 770, 791, 812, 831,
834, 843, 845, 851, 855, 860, 873, 885,
891, 901, 946, 980, 1077, 1078,
1084, 1090
- Scandinavian civil law, 365, 374
- Science of choice, 49–51, 61, 62
- Science of contract, 49–51
- Scientific revolution, 899, 918, 1041
- Scientific societies, 905, 907, 908
- Scope rules, 88, 89
- Seasonality, 683, 684
- Sedentism, 885
- Selectorate theory, 226
- Self-enforcement, 270, 273, 288, 627, 630–632,
638–639, 643
- Self-enforcing, 7, 160, 162, 171, 175, 177–180,
273, 287, 356, 421, 425, 432, 460, 490,
622, 623, 629, 630, 632, 634, 635,
639–640, 642, 708, 709, 713, 715, 721,
813, 816, 820, 822, 825, 940
- approach, 709
- equilibria, 721
- federalism, 177
- institution, 709
- property of institutions, 708
- Self-governance, 816, 840, 843, 847, 848, 851,
853–855, 859–861, 1093
- Self-identification, 711
- Self-preferencing, 576–578, 586
- Self-reinforcement, 1062, 1096, 1097
- Self-reinforcing sanctions, 709
- Self-undermining process, 715
- Semi-presidential, 131, 145
- Senior managers, 506–508, 515, 516, 519
- Sequential development, 543
- Serfdom, 16, 707, 745, 786, 789, 790, 795,
797–798, 800, 804, 806, 842
- Severance pay, 62
- Shadow banking system, 940
- Shadow of the future, 623, 631, 637–640, 642
- Shared understanding, 84, 86, 625, 635, 636,
643–645, 845
- Shareholders, 12, 63, 362, 369, 371, 375, 378,
380, 401, 507, 509–513, 516–525, 528,
530, 531, 665, 740, 765
- Shareholder voice, 517–519
- Shipley, William, 914
- Shirking, 271, 275, 330, 435, 509–511, 514,
520, 521, 548, 987
- Simple autocratic empire, 196
- Simple contracts, 284, 675, 677, 684–685
- Simple contractual schema, 59–61
- Simple market exchange, 48, 50, 54, 61, 64, 65
- Simulation models, 721
- Singlehomers, 572
- Site specificity, 426, 441, 453, 462, 463
- Slavery, 173, 175, 178, 213, 332, 742, 744, 769,
770, 772, 773, 780, 790, 791, 805, 874,
890, 1107
- Small-scale fishers, 353
- Smith, Henry, 294, 298
- Social capital, 374, 713, 761, 899, 906, 907,
977, 986–994, 1099, 1108
- Social choice theory, 81, 137, 138
- Social customs, 705
- Social distance, 18, 977, 987, 990, 992–994, 996
- Social heterogeneity, 857, 993, 994
- Social institution, 9, 295, 297, 748, 791, 873
- Social interactions, 16, 17, 374, 743, 873, 874,
886, 990
- Socialist market economy, 934
- Socialization, 713, 982, 990, 1095
- Socially efficient institution, 714
- Social networks, 18, 571, 572, 614, 734, 803,
819, 977, 987, 989–992, 994, 1018
- Social norms, 18, 599, 708, 722, 816, 876, 879,
884, 908, 995, 996, 1003–1005, 1007,
1010–1013, 1017–1019, 1120, 1132
- Social relationships, 294, 820, 850, 989, 995
- Social system, 548, 552, 555, 733, 1120, 1131,
1132, 1135
- Society capacity, 943
- Soil conservation, 14, 476, 692, 693
- Sourcing systems, 603, 608, 610
- Spatial regression discontinuity designs, 712
- Specialization gains, 688
- Specificity of assets, 479, 483, 486
- Spin dictators, 222, 244

- Spot markets, 55, 56, 58, 61, 419, 422–424, 432, 440, 442, 448, 450, 490, 598, 601, 603, 606, 608, 631, 632, 638
- Spulber's conception, 539
- Standard economic theory, 42–44, 50
- State and private sectors in post Mao China
early reforms (1978–1992), 933–934
laws and private growth (2007–2012), 935, 936
reversals and uncertainty (2012–present), 936–938
support to private activity (1992–2007), 934
- State-building, 205, 959, 1064, 1080, 1081, 1083
- State capacity, 20, 258, 986, 1070, 1128, 1129, 1138
- State dominance, 929
- Stated preferences approaches, 1094
- State enforcement, 320, 322, 1133
- State-owned enterprises (SOEs), 282, 642, 744, 932, 934–940, 944, 945, 947, 948
- Static game, 720
- Statistical studies, 1062, 1063
- Stepping stone approach, 716
- Stigler's theory, 424
- Stockholders, 506–511, 513, 516–518, 520, 525, 530
- Storage system, 886
- Strategic centers, 486, 492
- Strategic default, 603
- Strategic mobilization, 107
- Strategic voting, 112–115, 140
- Subnational governments, 160, 162–166, 169–174, 177, 179, 180, 186, 202, 943
- Sub-Saharan Africa, 150, 228, 229, 323, 366, 713, 730, 967, 968, 1099, 1100
- Suffrage institutions, 775, 779, 780
- Supplier opportunism, 539, 558
- Supply/supplier, 52, 53, 59, 60, 92, 271, 321, 342, 424, 426, 428–430, 436, 448, 461, 464, 485, 492–494, 530, 539, 544, 546–548, 558, 560, 604, 608, 624–626, 631–633, 636, 639, 645, 653, 661–665, 818, 830, 994
- Supply chains, 455, 457, 489, 492, 599, 602, 604, 609, 610, 617, 632
- Supply reliability, 423
- Surrogates, 1120, 1123, 1124, 1128, 1134, 1137, 1138
- Survey, 13, 16–18, 20, 32, 58, 113, 161, 162, 174, 179, 212, 221, 222, 273, 301, 438, 441, 453, 496, 503, 546, 568, 601, 608, 636, 641, 642, 666, 667, 691, 693, 712, 721, 722, 737, 745, 749, 786, 796, 805, 816, 823, 934, 940, 943, 945, 954, 955, 959, 969, 977–979, 985, 986, 988–990, 1003, 1007, 1008, 1033, 1094, 1100, 1120, 1121, 1125, 1126, 1128, 1129, 1134, 1137
- Swiss Confederation, 854
- Systematic seasonality, 683
- ## T
- Takeovers, 284, 401, 510, 516–519, 524
- Taoism, 930, 931
- Taxes, 42, 89, 92, 121, 163, 169, 170, 175, 180, 191, 196, 197, 199, 245, 310, 398, 407, 573, 765, 800, 803, 834, 836, 838, 840, 844, 854, 856, 857, 942, 943, 979, 982, 985, 995, 1071, 1073, 1080, 1081, 1092, 1129
- Technological change, 326, 435, 453, 459, 515, 541, 546, 547, 731, 760, 769, 911
- Technological mechanisms, 271, 287
- Technological progress, 17, 744, 899, 900, 903, 906, 909–911, 919, 920
- Temptations to deviate, 601–603, 606, 608
- Texts, 363, 364, 958, 967, 1031–1038, 1045, 1047, 1054, 1055, 1095
- The Enlightenment, 17, 743, 747, 748, 899, 902–904, 906, 908, 918
- The Great Instauration*, 917
- Theory of contracts, 13, 706
- Theory of the firm, 5, 39, 58, 59, 213, 228, 448, 460, 465, 529
- The Wealth of Nations*, 38, 49, 403, 1118
- Third parties, 14, 16, 118, 268, 270–272, 274, 282, 292, 301, 304, 305, 328, 358, 371, 434, 440, 441, 446, 487, 489, 529, 533, 546, 576–578, 622, 626, 641, 654–656, 682, 727, 731, 734, 796, 816, 820, 822, 847, 938, 939, 974, 976, 988, 989, 1018, 1058
- Third-party enforcement, 268, 482, 622, 641, 642, 734, 939, 989
- Tillage, 691–693
- Timeliness, 626, 680, 683, 684, 687, 688
- Tipping point, 705, 715, 718
- Topic modeling, 1034, 1035, 1038–1042, 1047, 1048, 1052
- Tort law, 282, 305, 363, 583, 586–588
- Total factor productivity (TFP), 937, 947
- Township and village enterprises (TVEs), 33, 750, 933
- Trade opportunities, 719
- Traditions in contact design, 689

- Transactional attributes, 420, 424, 439, 440, 443, 455, 461
- Transaction cost economics (TCE), 5, 47–68, 75, 401, 418, 449–451, 476, 622, 652, 721
- Transaction costs, 2, 5, 6, 9–12, 16, 19, 41–44, 47–68, 81, 228, 268, 271, 272, 280, 282, 285, 287, 288, 392, 393, 396–402, 409, 410, 419, 437, 449, 452, 454, 459, 476, 479, 539, 542, 706, 73, 734, 736, 743, 792, 793, 979, 1015, 1051, 1118, 1123
- and agriculture, 683–684
- and property rights, 682–683
- and risk aversion, 689–691
- simple contracts, 684–685
- Transaction-specific investment, 546
- Transitional economies, 90
- Transnational enforcement, 287–288
- Trial-and-error behaviors, 703
- Trust, 7, 13, 14, 18, 187, 192, 193, 198, 199, 270, 273, 374, 409, 455, 460, 530, 544, 566, 574, 599, 600, 603, 613, 614, 622–624, 626, 630, 633, 635, 636, 642, 713, 719, 732, 734, 738, 743, 761, 763, 816, 932, 934, 940, 941, 975–, 996, 1011, 1016, 1095, 1127–1130, 1134
- institutions and norms of, 987, 989
- intrinsic norms of, 992–994
- social networks and, 990, 991
- Trustworthiness, 270, 272, 734, 976–978, 986, 988, 989, 992, 993, 995, 996
- Turning point, 139, 175, 478, 705, 706, 957
- Two-sided platform, 568, 570, 572, 579
- Tying, 13, 431, 574, 578, 582, 585, 947
- Typological theory, 19, 1060, 1064–1074, 1084
- U**
- Uncertainty, 2, 12, 26, 29, 54, 60, 64, 118, 226, 248, 268, 302, 310, 356, 404, 420, 424, 430, 431, 450–453, 456, 460, 461, 464, 465, 482, 483, 487, 516, 537, 543, 545, 549, 550, 552, 557, 559, 560, 683, 710, 733, 936–938, 945, 948, 1132
- Undernomination, 116
- Unitization, 322, 340, 349, 351, 461
- Unit of analysis, 50, 53, 54, 1050, 1100, 1104, 1124
- Unit of observation, 1048, 1124
- Universalism, 1008, 1009, 1015
- Unlimited supply of labor, 704
- Upper tiers, 108, 117
- Urgency powers, 144
- Useful knowledge, 17, 898–891, 903–905, 908–910, 912, 914, 916, 918
- Uzbekistan, 216, 711
- V**
- Varieties of Democracy database (V-DEM), 132, 133, 136, 145, 146, 148, 240, 242–245, 254–257
- Vector autoregression (VAR), 1049
- Vertical corporate governance, 509–511, 513, 516, 517, 519, 520, 528, 529, 531
- Vertical governance, 419, 442, 508, 519, 531
- Vertical integration, 11, 13, 48, 52–60, 64, 67, 271, 392, 394, 399–400, 417–443, 450, 452–456, 458, 460–466, 478, 481, 529, 537, 539, 540, 546, 549, 560, 598, 606, 684, 823
- Vertical market restrictions, 59, 61, 62, 64, 65
- Veto, 7, 133, 136, 141, 144–147, 178, 321, 329, 330, 376
- Village communities, 187–191, 196
- Vineyard contracts, 686
- Violence, 167, 172, 218, 219, 222, 240, 320, 321, 327, 330, 334–336, 399, 402, 404, 603, 614, 718, 733, 735, 795, 802, 831, 835, 855
- Virtuosi, 917, 918
- W**
- Wallis, J.J., 167, 168, 170, 177, 180, 213, 331, 334, 402, 736, 739, 789, 900
- Waqf*, 959, 960
- War of the Roses, 858
- Weak society, 242, 929, 932, 947
- Wealth inequality, 241, 874, 880, 889–891
- Weber, Max, 113, 373, 440, 453, 456, 760, 761, 764, 874, 875, 930, 954, 955, 963, 1004, 1007, 1008, 1016, 1022, 1094, 1131
- Weingast, B.R., 6, 7, 159–180, 213, 214, 251, 269, 270, 288, 331, 334, 533, 736, 739, 789, 793, 800, 813, 814, 820, 823, 831, 858, 900, 928, 929, 1101, 1102, 1110, 1124
- Welfare system, 80
- Western Europe, 132, 145, 147, 149, 707, 734, 741, 742, 745, 747, 751, 759, 789, 797–800, 899, 919, 957, 959, 964, 965, 1047
- Why Nations Fail*, 212–213, 269
- Williamson, Oliver E., 2, 3, 5, 10, 11, 13, 38, 40, 41, 44, 47–68, 75, 102, 213,

-
- 268–271, 284, 373, 388, 401, 405, 406,
419–421, 423, 426, 427, 430, 431,
433–436, 449, 450, 453, 456, 458,
462–466, 476–482, 484, 485, 490–492,
497, 498, 500, 503, 512, 515, 528, 530,
536, 538, 539, 544, 549, 559, 560, 599,
622, 623, 652, 657, 659, 661, 668, 684,
734, 814–816, 822–824, 928, 940, 946,
976, 1066, 1091, 1093, 1118, 1125,
1138
Williamsonian branch, 2
Wissenschaft, 904
Word embeddings, 1043, 1044, 1054