The Curious Culture of Economic Theory

Ran Spiegler

The MIT Press
Cambridge, Massachusetts
London, England

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Palatino by Westchester Publishing Services.

The author's preference is to use binary pronouns in this work. The MIT Press recognizes all gender identities and recommends the use of gender-neutral pronouns.

Library of Congress Cataloging-in-Publication Data

Names: Spiegler, Ran, author.
Title: The curious culture of economic theory / Ran Spiegler.
Description: Cambridge, Massachusetts : The MIT Press, [2024] | Includes bibliographical references and index.
Identifiers: LCCN 2023028463 (print) | LCCN 2023028464 (ebook) | ISBN 9780262548229 (paperback) | ISBN 9780262379021 (epub) | ISBN 9780262379038 (pdf)
Subjects: LCSH: Economics. | Economics—Sociological aspects.
Classification: LCC HB171 .S7155 2024 (print) | LCC HB171 (ebook) | DDC 330.01—dc23/eng/20230713
LC record available at https://lccn.loc.gov/2023028463
LC ebook record available at https://lccn.loc.gov/2023028464

Contents

Preface

This book is a collection of essays about the professional culture of economic theory. When is a theoretical result "taken seriously" for economic applications, and how do theorists try to influence this judgment? What determines whether a new theoretical subfield adopts a "foundational" or an "applied" style? Why have theory papers become so long, and how do journals and readers handle this trend? How do theorists respond to economists' taste for "rational" explanations of human behavior? Each question addresses the norms that economic theorists apply as they produce, evaluate, and disseminate research. The essays in this book explore these questions and others. Through them, I hope to illuminate our culture—at least as I have experienced it since the turn of the century.

In a strange way, the book is a product of the COVID-19 pandemic. Lockdowns, school closures, and travel restrictions disrupted my cherished work habits as an economic theorist (no more sketching models in cafes or proving theorems in airport lounges), and suddenly gave a comparative advantage to a different kind of project that is not *in* economic theory but *about* it: a project that would allow me to mull over an idea for as long as I wanted and implement it in brief, unpredictable spurts of activity.

At the same time, the pandemic intensified the kind of introspection that writing about one's own culture demands. When the crisis went global in March 2020, several members of my international research community decided they were not going to sit this one out. Theorists who hadn't shown a strong bent for policy-oriented research suddenly began composing pieces about how to do viral tests more efficiently, or how to make epidemiological models better at accounting for behavioral responses to mitigation policies. Some of these pieces were garden-variety applied theory inspired by the situation, but others had a direct

policy pitch. Some of us claimed to have temporarily abandoned economic theory altogether, realizing there were more important and urgent things.

This reaction was short-lived. But from my subjective perspective, it seemed to reflect a deep-seated anxiety about the role of theorists within the economics profession and in society at large. Theorists regularly live with this anxiety: witness our constant attempts to write papers that would appeal to the "general reader" (translation: labor economists; they are not "general," and they have better things to do than read our papers). The COVID-19 crisis brought this anxiety to the surface.

This combination of factors impelled me to try something I had wanted to do for a long time: write about economic theory in a style that I thought I had seen in other disciplines but not in my own. It would involve a bit of intellectual history, but it wouldn't be a "proper" history-of-economic-thought treatise. It would have its share of polemic, but it wouldn't campaign for any particular position. Its selection of topics and commentaries would be subjective, but the discussions would be grounded in objective, pedagogically oriented exposition of concrete pieces of economic theory. It would occasionally get technical, but it wouldn't be written exclusively for connoisseurs. Conversely, while it would present concrete examples of economic theory in a deliberately accessible manner, it wouldn't be a popular-science book. And it would make some use of my own work experience, but it certainly wouldn't be a "scientific autobiography."

Instead, it would be a collection of "cultural criticisms" by a working theorist: not a philosopher or historian who perceives this culture from afar; nor an aristocrat of the profession who has lost touch with the everyday business of economic theory. Too often, our community leaves the task of "talking about the profession" to its mightiest big shots. But isn't it more interesting to hear the perspective of active theorists outside the profession's house of lords? True, in recent years, social media is filled with academic commentary by a growing and diverse crowd of economists. Yet, there is still a big difference between the brief, jumpy Twitter thread, however sharp and articulate, and the measured, longer-breathed, and carefully organized essay form—the genre to which the chapters in this book belong.

There is an additional factor behind this book. In the years 2015–2021, I served as a coeditor and then chief editor of the journal *Theoretical Economics*. This experience has given me several opportunities to

muse over our professional culture and occasionally try to nudge it ever so slightly.

Who is the intended audience of this unusual "cultural criticism of economic theory"? Obviously, I will be happy if members of my research community of economic theorists read it and find it thought-provoking. Hopefully, they'll be intrigued by the "cultural criticism" spin on classics from the last quarter-century and find it worthwhile to assign as a complementary reading in (core or advanced) graduate-level economic-theory courses. However, I am also targeting economists from other subfields, who often look at theorists with varying mixtures of bemusement, puzzlement, and disapproval. I know that I would be very curious to read an introspective analysis of the professional culture of, say, applied microeconomics. By the same token, I hope that academic economists of various stripes will take an interest in the present text. Philosophers and historians of science may use the book's content as valuable raw material for their more professional and systematic discourse on the methodology and sociology of contemporary economics. Finally, I have tried to pitch the occasional technical discussions at a level that readers with minimal graduate-level exposure to economic theory will be able to grasp. Those readers, who frequently encounter rants about economic theory in popular and social media, might be curious to learn a bit about what this curious culture looks like from the inside.

I am grateful to Yair Antler, Oren Danieli, Kfir Eliaz, Nathan Hancart, Elhanan Helpman, Michele Piccione, Ariel Rubinstein, Heidi Thysen, and Dan Zeltzer for their comments on an earlier draft of the book, and for their general support for this project. I also benefited from comments on specific chapters by Duarte Gonçalves, Stephen Morris, and Philipp Strack. Tuval Danenberg helped preparing the index and bibliography and offered excellent additional comments on the substance. Finally, I wish to thank Emily Taber, the MIT Press editor, for valuable exchanges that helped me improve the book. All remaining lame self-referential jokes are mine.

Ran Spiegler
Tel Aviv, September 2022

1 Apps and Stories (an Introduction)

The Oppressors Have Become the Oppressed

In the epilogue of their blockbuster book *Mostly Harmless Econometrics* (2009), Josh Angrist and Steve Pischke write, "If applied econometrics were easy, theorists would do it."[1] As academic jokes go, this one is reasonably funny. But coming at the end of a book that didn't display the slightest interest in economic theory (and why would it?), the joke feels gratuitous. It prompts the reader to look for some hidden resentment behind the joke.

Such resentment against economic theory and economic theorists is something the authors could have picked up during their formative years as graduate students. The late 1980s were peak years in terms of the status of economic theory within the broader economics profession. The field had gone through the so-called game theory revolution and was busy rewriting graduate-level economics textbooks. Graduate programs put a large premium on abstract formal modeling and accompanying mathematical techniques. This created dismay among students, who had other reasons for pursuing an academic career in economics.

David Colander and Arjo Klamer captured this mood in a *Journal of Economic Perspectives* article titled "The Making of an Economist," which they later expanded into a book.[2] During interviews with students in top graduate programs, they observed that their interlocutors didn't like the outsized role of economic theory and mathematical technique in their curriculum:

> As to the things they liked least, the majority of comments focused on the heavy load of mathematics and theory and a lack of relevance of the material they were learning.

Still, the students understood the culture they were immersed in:

They are convinced that formal modeling is important to success, but are not convinced that the formal models provide deep insight into or reflect a solid understanding of the economic institutions being modeled. Believing this, they want to be trained in what the profession values. Thus we find that students who believe they are not being taught the most complicated theory feel deprived and unhappy because they worry about the ability to compete.

The sense that "real economists" are being oppressed by a subculture that fetishizes formal modeling and mathematical pizzazz keeps resurfacing from time to time. Here is Thomas Piketti's memorable quote:[3]

The discipline of economics has yet to get over its childish passion for mathematics and for purely theoretical and often highly ideological speculation, at the expense of historical research and collaboration with the other social sciences.

Occasionally, the expression of this sentiment carries political overtones. Paul Krugman's famous 2009 *New York Times* article "How Did Economists Get It So Wrong?" associated it with political conservatism and a strong belief in the postulates of rational choice and competitive markets:[4]

The economics profession went astray because economists, as a group, mistook beauty, clad in impressive-looking mathematics, for truth. . . . As memories of the Depression faded, economists fell back in love with the old, idealized vision of an economy in which rational individuals interact in perfect markets, this time gussied up with fancy equations. . . . The central cause of the profession's failure was the desire for an all-encompassing, intellectually elegant approach that also gave economists a chance to show off their mathematical prowess.

Krugman's beef was with macroeconomic rather than microeconomic theory (which is what most academic economists associate with the term "economic theory"), but the resentment is similar: a culture in love with "fancy equations" derails the discipline from its right path. It is significant that Krugman lumps "theory loving" with belief in rationality and markets (and implicitly, with right-wing politics). He's not the only one performing this trick (Kay 2012), and I'm not the only one who noticed (see Michael Woodford's [2011] response to Kay's article).

These gripes about the unwarranted dominance of theory in economics have become less frequent over the years. Once the game theory revolution was complete and the textbooks were rewritten, economic theory reached a stage of consolidation and gradually reassumed its traditionally marginal position in the professional landscape. At the

same time, the status of empirical work in economics has risen dramatically. Increased computing power, proliferating data sets, and greater confidence in their methods have made empirical economists happier about the state of affairs. They have developed a sense that the discipline is moving in the right direction and becoming more scientific. When David Colander wrote a sequel to *The Making of an Economist* in 2007, he was pleased to report that twenty years after the original Colander-Klamer interviews, the students at top graduate programs were at ease with the more modest role of theory in their education.[5]

Indeed, the balance of power between theorists and "real economists" has shifted. A popular narrative has emerged: once upon a time, data was scarce, and so we had to base economic analysis on theoretical arguments, but now there is plenty of data and we know how to deal with it, and so the theorists can return to the back seat, where they belong; the inmates no longer need to run the asylum.

A parallel trend, which may or may not be related, is the increasing career premium for publishing papers in what my longtime collaborator Kfir Eliaz calls the "high five" journals.[6] This trend has become so strong that people now refer to it as the "curse" or "tyranny" of the "top five."[7] Since members of this mighty fist orient themselves as "general readership" journals, authors are expected to address the "general reader," who is—needless to say—not a theorist. This further shifts the balance of power. Theorists can no longer settle for satisfying each other; they are busy pleasing members of other fields.

This attitude is a one-way street: labor economists probably don't have theorists in mind when submitting their work to the top-five journals, whereas theorists are expected to put themselves in the labor economists' shoes. The eminent theorist Debraj Ray, until recently a coeditor at the *American Economic Review*, once told me that his editorial decisions on theory papers are guided by what he called the "Mark Gertler test"—namely, whether he can successfully pitch the paper to his NYU colleague, the leading macroeconomist Mark Gertler. I replied that I wonder whether Gertler would apply a "Debraj Ray test" if he handled a macroeconomics paper as an *AER* editor.

The Applied Dimension

Theorists' anxiety about their place in the broader economics community is nothing new. I remember that, in 2000, Kfir Eliaz and I went to

Bilbao for the first World Congress of the Game Theory Society. I had recently finished my PhD; Kfir was about to finish his. We surveyed the colleagues who swarmed the large conference halls and played the silly game "economist or modeler": the task was to classify every senior theorist we saw into one of the two categories, "real economist" or "mere modeler" (the two of us clearly belonged to the latter).

Yet, the pressure on theorists to define themselves vis-à-vis applied economists and seek their affirmation has only grown stronger over the last two decades. For a recent demonstration, we need look no further than the 2020 economics Nobel Prize that went to Paul Milgrom and Robert Wilson. As any theorist would agree, these are two highly deserving laureates who made several landmark contributions to economic theory. And yet, a huge portion of the background information provided by the prize committee was devoted to the laureates' *practical* work on auction design at the service of governments or private companies.[8] The message was not lost on commentators. Tyler Cowen (2020) wrote in his blog,[9]

> The bottom line? If you are a theorist, Stockholm is telling you to build up some practical applications. . . . The selections themselves are clearly deserving and have been "in play" for many years in the Nobel discussions. But again, we see the committee drawing clear and distinct lines.

The pressure to be practically useful is arguably the most powerful force that acts on contemporary economic theorists. In the course of this book, we will have many opportunities to see the pull of this "applied dimension" at work.

The Aesthetic Dimension

Another dimension represents a view of economic theory that emphasizes "artistic" or "aesthetic" values—particularly the tickle that we get when encountering a good *story*, dressed in the language of a formal economic model. Here is what Robert Lucas had to say in 1988, in a beautiful commencement address to University of Chicago students, which was later published under the title "What Economists Do" (and it is significant for our story that Lucas was a chief villain in the narrative that Krugman's 2009 journalistic piece concocted):[10]

> Economists have an image of practicality and worldliness not shared by physicists and poets. Some economists have earned this image. Others—myself and many of my colleagues here at Chicago—have not. I'm not sure whether you

will take this as a confession or a boast, but we are basically story-tellers, creators of make believe economic systems. . . . In any case, that is what economists do. We are storytellers, operating much of the time in worlds of make believe. We do not find that the realm of imagination and ideas is an alternative to, or a retreat from, practical reality. On the contrary, it is the only way we have found to think seriously about reality.

I don't know if Lucas felt this way later in his life, but I know that Ariel Rubinstein does. In various lectures and essays, such as his Econometric Society presidential address or popular-science-ish book, appropriately titled *Economic Fables*,[11] Rubinstein presented the unadulterated view of economic models as stories. According to him, our response to a successful economic model is like the response to a good fable. It is not a scientific response but an "artistic" one. It is a recognition that the model offers an abstract representation of reality that we find edifying in a way that we cannot or will not subject to a properly scientific test.

Ticking Boxes

The culture of economic theory can be viewed as an intricate maneuver between the applied and the aesthetic, the "scientific" and the "artistic." A theorist's professional identity has a lot to do with how she locates herself in the space defined by the applied and aesthetic dimensions.

Of course, the theorists' value system is not two-dimensional; they use additional criteria to guide their own work and evaluate the work of their peers. One criterion is technical brilliance. Above-average aptitude for math is a key part of many theorists' self-worth: Krugman got that one right! Theorists' sense of mathematical superiority offers partial compensation for their sense of inferiority on the "usefulness" dimension. As the latter became more acute, theorists felt a need to double down on the former. Over the last two decades, economic theory has become outwardly more technically demanding.

Another criterion is conceptual innovation, the mission of broadening the scope of what formal models can say about economic behavior. In the revolutionary 1970s and 1980s, when economic theory exerted its "oppressive" power over the rest of the economics profession, expanding the language of economics was a shared core mission among theorists. Even in today's postrevolutionary phase, our culture still rewards theorists for pushing economics' conceptual envelope (although demand for this kind of work appears to be weaker now).

These four coordinates—the applied, the aesthetic, the technical, and the conceptual—have always shaped the professional culture of economic theory. Changes in our culture amount to changes in the relative weights that we assign to them, but also in our expectations as to *how many* of these dimensions a single piece of economic theory should occupy. My impression is that, over the years, this number has gone up, especially when it comes to "high five" publications. Yet, ticking multiple boxes with a single paper—offering a conceptual innovation *and* demonstrating it with a convincing "economic application," or writing a thought-provoking story that *also* shines with flashy mathematical technique—is a devilishly difficult feat. It may be a fool's errand, but many theorists still try, fueled by the increasing pressure to score top-five publications. This tendency is another key factor that defines the contemporary culture of economic theory.

Structure . . .

This book is a series of explorations into how theorists deal with the pressures that shape our professional culture, especially the tension between "applied" and "aesthetic" values.

Chapters 2, 3, and 4 are devoted to the interplay between "pure" and "applied" approaches to economic theory. Chapter 2 explores the fine line that separates the applied from the paradoxical, using the theory of "global games" as a test case. Chapter 3 continues this theme, highlighting various rhetorical and stylistic devices that economic theorists use to escape paradox and lend an "applied" veneer to their models. Chapter 4 shifts attention from individual papers to entire subfields. Using behavioral economics as a test case, it explores how subfields "choose" to orient themselves in the pure-applied spectrum.

Chapters 5, 6, and 7 are a series of reflections on various aspects of the current culture of economic theory: the "rationalizing" mode of explanation that is so popular in economics, the growing dimensions of theory papers and the resulting practice of relegating material to "supplementary appendices," and the norms that govern our evaluation of incremental modeling innovations.

In chapters 8 and 9 I get more personal and use my own work to illustrate two themes: the emerging culture of "market design" at the expense of the older competitive-equilibrium culture, and the "artistic" nature of economic models as stories. I conclude in chapter 10 with brief thoughts about the future of economic theory.

. . . and Style

The style of this book's essays seems to be new in economics. Economists have used the essay form before, but usually to talk about methodology or to support a position in a debate between schools of thought. The essays in this book, by contrast, are not about core methodologies or philosophies of economic theory. Instead, they address the style of its delivery, the rhetorical gambits its practitioners employ, and the ancillary modeling choices they make, as well as the norms that shape audiences' response to these rhetorical and stylistic moves. This is why I classify the essays as "cultural criticisms." I should qualify this label by saying that I have no expertise in the academic disciplines that are usually associated with this term and make no attempt to establish links to those disciplines. I am an expert economic theorist but an amateur cultural critic.

The manner in which I execute my cultural criticisms is not methodical, but allusive and impressionistic; the claims and judgments I make along the way are informed, but also subjective. Yet, the book is not all fluff: my discussions of style and rhetoric are grounded in concrete models from the literature, such as the e-mail game, Bayesian persuasion, or rational inattention. While the selection of these examples is subjective and reflects my own experience, their description is as precise and self-contained as possible while striving for minimal notation and math. This mixture of precise (yet accessible) exposition of formal models and impressionistic verbal discussion is, as far as I can tell, a novelty in economics. It hopefully makes the book a valuable companion to "proper" texts in microeconomic-theory courses. At any rate, approaching the text as if it is meant to be fully objective and tightly argued can lead to misunderstandings.

In an attempt to preempt some of the misunderstandings that my style can generate, I wish to alert the reader to two features of this style. First, when an essay in this book highlights a rhetorical effect in some modern economic-theory classic, the reader might infer that I am suggesting the authors *deliberately* engineered the effect. That would be what literary critics call an *intentional fallacy*—namely, a tendency to overattribute literary effects to authorial intent.[12] Therefore, I ask the reader to resist this instinctive response: I am merely proposing an interpretation of the paper's reception by our profession, whether its authors intended it or not.

A second possible reaction to my "cultural" take on economic theory is that it reflects some kind of disrespect for its scholarly value. That

would be a false impression that has less to do with my attitude to economic theory and more to do with the "cultural criticism" mode itself. For example, I make liberal use of scare quotes; that will not be sarcasm but a useful distancing device that enables me to dissociate terms from their conventional interpretations.

The suggestion that successful pieces of economic theory make their impact partly through rhetorical devices and calibration of audiences' stylistic expectations does not diminish from their status. In this sense, I am in agreement with McCloskey (1985), possibly the most well-known foray into the role of rhetoric in economics. I am less sure that this agreement extends to our basic attitudes to economic theory. When I first read *The Rhetoric of Economics*, it felt like yet another grudging response to theorists' 1980s oppressive reign (and a very well-written one). This is definitely not going to be the case here. Unlike McCloskey, I am a theorist. Accordingly, my "cultural criticism" of economic theory is an affectionate one. The bewildering professional norms that govern what "works" and "doesn't work" in the world of economic theory can be a source of frustration, but they also fascinate me. Economic theory's elusive mixture of "scientific" and "artistic" elements is probably what attracted me to it in the first place. I don't think I would have been drawn to the field if it had been too far on either side of the art-science spectrum. Maybe the mixture will change in the future, in which case it is likely to attract a different type of scholars. Maybe it is already changing.

2 The Paradox around the Corner

Coordinated Attack

Imagine a scene from ancient times. Two armies—call them A and B—face a common enemy. The enemy is camping in a valley and therefore vulnerable to an attack from the surrounding hills. There is a snag, however. Three snags, actually. First, the attack must be coordinated: neither army is big enough to overcome the enemy on its own. Second, even a coordinated attack can be successful only if enemy forces are depleted to begin with. An unsuccessful attack—whether because it is uncoordinated or because the enemy is strong—is deadly and humiliating; no army general would want to launch an attack unless he is sufficiently certain it will be successful. Which brings us to the third and final snag: *only army A* has a vantage point that enables it to observe the size of enemy forces.

To a modern reader, this doesn't sound like much of a predicament. When army A's general learns from his watchmen that the enemy is feeble, all he has to do is pick up a secure phone and call his counterpart at army B, and they can coordinate the attack. But remember, these are ancient times. No phones. The two parties must rely on a different communication protocol. Army A sends a messenger on a camel. The messenger must climb down the hill, ride through the valley, and climb up to army B's location.

It's a somewhat dangerous ride. There is a small chance that the messenger will be spotted and executed by a gang of robbers. If the messenger makes it to army B's camp, conveys the good news, and fixes the time of the attack, he turns back and rides all the way back to army A's camp, facing the same risk of getting caught. If he reaches it, he informs army A that he has conveyed the good news to army B. But the protocol is not over: the messenger saddles up and makes yet another

trip to army B's camp, in order to let army B know that army A knows that he broke the good news to army B.

And so, our camel-riding messenger keeps traveling back and forth between the two camps. Each time he crosses the valley, there is a small chance he will be captured by the robbers, and the communication will be broken. However, if this chance is very small, the communication protocol is the best simulation of modern, simultaneous communication that the ancient technology can offer. With very high probability, the messenger will make a large number of trips, thus assuring army A that army B knows that army A knows that army B knows . . . that army A knows that conditions are ripe for a successful attack, where the length of this chain of iterated knowledge is arbitrarily high. Eventually, the messenger will be caught and therefore the communication chain will be finite. Our army generals will never attain what game theorists call "common knowledge"—namely, an *infinite* chain of iterated knowledge. But they can get awfully close. (As with any made-up story like this, the reader is expected to ignore certain unrealistic features, such as that, by the time the messenger completes more than a couple of rides, it will be too late for an attack.)

And here's the question. Suppose the messenger never came back from his first voyage to army B's camp. Will the general of army A order an attack? How would the answer change if the messenger came back from the first voyage but not from the second? And what if the messenger managed to complete forty-nine trips before his eventual demise?

The E-mail Game

Fast-forward to our present day. The scenario I have described is known in the computer science literature as the "coordinated attack problem."[1] It is a parable that was meant to illustrate the difficulty of attaining a coherent state of knowledge in a distributed computing system.

But the computer scientists did not address our *behavioral* question: How will the army general make the strategic decision whether to attack, given this imperfect communication protocol? Addressing this question requires us to describe the situation in a way that will capture both its informational intricacies and their implications for the generals' behavior. In other words, we may want to write it down as a *game*.

In 1989, Ariel Rubinstein published a paper that did precisely that.[2] The first thing his paper did was to remove the anecdotal aspect of the game and replace it with an abstract, storyless 2×2 game, which does,

however, fit the coordinated attack story. The next thing he did was to modernize the communication method. In the 1980s, electronic mail was a shiny new technology for academics, and messages that failed to arrive at their destination were not unheard of.

Rubinstein described the following communication protocol, in which e-mails replaced the human camel-riding messenger. A priori, the enemy is weak with probability p, which is below ½ but arbitrarily close. When army A's general learns that the enemy is weak—and *only* then—his computer sends an *automatic* message to army B's computer. When this message arrives at its destination, army B's computer sends an automatic confirmation message to army A's computer, which in return sends an automatic confirmation message to army B's computer. This orgy of confirmation e-mails continues until one of the messages fails to reach its destination. Each message has an independent failure probability of q. Therefore, conditional on the enemy being weak, the probability that the communication stops after a total of K messages is $q(1-q)^{K-1}$. At the end of this process, the computer screen of each army general displays the total number of messages that his computer *sent*. This number encodes the general's state of knowledge.

For example, when army A's general sees the number 2 on his screen, this means that his computer sent the original message and another confirmation message but did not receive confirmation for the latter. Thus, army A's general knows that the enemy is weak; he knows that army B knows that it is weak; but he does not know whether army B's general knows that he (army A's general) knows that army B knows that the enemy is weak. This is because he does not know whether the failure to receive confirmation of his second message was due to failure of his last outgoing e-mail or failure of the subsequent incoming confirmation e-mail from army B's computer.

A larger number on a player's screen thus represents a higher level of iterated knowledge. As with the ancient messenger story, the e-mail communication protocol stops after finitely many rounds with probability one. Therefore, the two generals will never reach the infinite chain of iterated knowledge that defines common knowledge. However, if q is small, they are likely to reach a high level of iterated knowledge.

Having described the game's information structure, let us write down its payoffs, which reflect the coordinated attack story. Suppose that, when an army does not attack, it gets a payoff of 0 for sure. In other words, not attacking is a safe action. In contrast, attacking is a risky action: it yields a gain of 1 if the attack is successful and a loss of 1 if

	Attack	Don't attack
Attack	x, x	–1, 0
Don't attack	0, –1	0, 0

Figure 2.1
Payoffs in the e-mail game.

the attack is unsuccessful. Recall that the attack is successful if and only if the enemy is weak and the other army attacks as well. This payoff structure can be encapsulated by the 2×2 payoff matrix in figure 2.1 (the value of x is 1 when the enemy is weak and –1 when it is strong).

The numbers have been cooked so that if an army general is clueless about whether an attack is going to be successful (by clueless I mean that the chances are fifty-fifty), he will be indifferent between attacking and abstaining because the expected payoff from attacking will be

$$0.5 \cdot 1 + 0.5 \cdot (-1) = 0$$

Nash Equilibrium

In the e-mail game, a strategy for a player is a function that assigns one of the two actions for each number on his computer screen. Rubinstein conventionally applied the solution concept of Nash equilibrium to this game. In Nash equilibrium, each player's strategy always prescribes an action that maximizes the player's expected payoff given his information, taking the other player's strategy as given.

In the common-knowledge benchmark—that is, the case of $q = 0$, in which the e-mail communication never breaks down and players' chain of iterated knowledge is infinite—each of the 2×2 payoff matrices that fit $x = 1$ and $x = -1$ can be analyzed in isolation. When the enemy is strong, there is a unique Nash equilibrium, in which neither army attacks. Indeed, attacking is manifestly a strictly dominated action: it yields a fixed payoff of –1, whereas not attacking yields a fixed payoff of 0. When the enemy is weak, there are two "pure" Nash equilibria: in one equilibrium, neither army attacks; in the other, both attack. The latter is a good equilibrium, as it gives both players a payoff of 1, whereas the bad equilibrium gives them both a payoff of 0.

But what about the e-mail game—that is, the case of $q > 0$? Here comes a surprise. Rubinstein showed that no matter how small q is, the e-mail

game has a *unique* Nash equilibrium, in which neither player attacks—regardless of the number on his computer screen.

The proof is by mathematical induction on the cumulative number m of messages that are sent before the communication breaks down. When m is an even number, we will examine the behavior of army A; when m is odd, we will examine the behavior of army B.

Let's start with $m = 0$. This corresponds to army A learning that the enemy is strong (and therefore his computer doesn't send any message). We saw that, in this case, attacking is strictly dominated, hence army A will not attack.

How about $m = 1$? This corresponds to army A sending a message that goes astray: army B is not receiving any message. But the general of army B doesn't know whether this is because the enemy is strong or because the enemy is weak, but the first e-mail from army A failed. In other words, army B cannot distinguish between $m = 1$ and $m = 0$. Using Bayes' rule, the conditional probability that $m = 1$ is

$$\frac{pq}{pq + 1 - p}$$

(I remind the reader that p is the prior probability of a weak enemy, and q is the probability that a message goes astray.) Since $p < ½$, this conditional probability is less than ½. If $m = 0$—that is, the enemy is strong—attacking is unsuccessful by assumption. Therefore, the probability that army B's attack will be successful given that army B receives no message is below the breakeven point of ½. The upshot is that regardless of what army B believes about A's behavior, it will not attack when $m = 1$.

Now comes the masterstroke. Suppose we proved the claim for all integers up to some $m > 0$. That is, we proved that both armies choose not to attack when the cumulative number of sent messages is at most m. Now suppose that the cumulative number of sent messages is $m + 1$, and consider the player who didn't receive the last message. This player doesn't know whether the total number of sent messages was m or $m + 1$. In other words, he knows that the last message his computer sent either failed or reached its destination and the confirmation message failed. By the inductive argument, the opponent doesn't attack in the former scenario. What is the probability of that scenario? That is, given that an army didn't receive confirmation for its last outgoing message, what is the probability that the message failed?

A cute Bayesian calculation will give us the answer. The probability the outgoing message failed is q. The probability that the outgoing

message arrived and the ingoing confirmation message failed is $(1-q)\cdot q$. The total probability that the player didn't receive confirmation for the last message he sent is the sum of these two probabilities. Bayes' rule tells us that conditional on this event, the probability that the outgoing message failed is

$$\frac{q}{q+(1-q)q}$$

This number is greater than ½. Therefore, regardless of what the player thinks about how the opponent will behave in case he did receive the player's last message, the probability of a successful attack is less than ½. Therefore, the player will prefer not to attack. We have thus proved the claim for $m+1$, which—by the logic of mathematical induction—means that we have proved it, full stop.

Note that in this proof, for large values of m, there is no uncertainty as to whether the situation is ripe for a successful attack: both players know that the enemy is weak. The proof makes it clear that the result is all about the *strategic* uncertainty due to each player's uncertainty about his opponent's *high-order* knowledge. It is a minor uncertainty in the sense that the player does not know whether that level is K or $K-1$, where K can be arbitrarily large. The constant, independent failure rate per message implies that $K-1$ is more likely than K; and the inductive argument implies that in the more likely case of $K-1$, the opponent doesn't attack.

The inductive reasoning is more than a mathematical proof technique. It has a deeper behavioral meaning: the outcome is driven by iterated elimination of strictly dominated strategies. Each round of the proof corresponds to a stage in this iterative procedure. The argument that army A won't attack when $m=0$ corresponds to deleting all strategies in which he attacks when the enemy is strong. The argument that army B won't attack when $m=1$ corresponds to deleting all strategies in which the army attacks when it sees zero on its computer screen. The argument that army A won't attack when $m=2$ corresponds to deleting all strategies in which the army attacks when it sees the number *1* on its screen. And so forth. This solution concept is weaker than Nash equilibrium: in a general finite game, the set of outcomes that survive the procedure contains the set of Nash equilibria. In the e-mail game, the two coincide because a unique outcome survives the procedure.

Paradox

How should we interpret the stark result? Rubinstein makes it clear that he doesn't treat the Nash equilibrium outcome in the e-mail game as a plausible prediction. First, he puts the term *prediction* under scare quotes. Second, he refers to the result explicitly as *paradoxical* and compares it to other well-known vignettes of game theory, like the Chain Store or Centipede Games—both examples of how inductive reasoning leads to a behaviorally implausible prediction.[3] While the term "paradox" is philosophically deep and multifaceted, I use it here the way I believe most game theorists do in this context: simply to characterize a theoretical prediction that powerfully clashes with our intuition about what actual behavior would look like.

Indeed, the e-mail game is written as a thought experiment that we can easily run in our head. Would we attack if we saw a high number on our computer screen? Most of us would. In fact, there is a sense in which the communication protocol makes coordinated attack a focal point. A high number on one's computer screen, when one knows that the opponent also saw a high number, is an implicit invitation to coordinate on the efficient outcome (attacking when both know the enemy is weak). There is a clash between this intuition and the game-theoretic "prediction." Refutation of this prediction in the thought experiment has been confirmed by actual lab experiments.[4]

How do we respond to this paradox? One obvious response is that it is an empirical refutation of standard game-theoretic methods. My experience from teaching this example is subtler: the students' response seems more "artistic" or "aesthetic." It is in fact a marvelous *joke*. Indeed, when I explain the inductive argument, many students begin smiling. I deliberately play it for laughs by conjuring up the image of the camel-riding messenger. That poor messenger, riding back and forth on his camel toward his inevitable demise. No matter how many rounds he manages to complete, he will never assuage the generals' fear that their army will be the only one launching an attack. Funny, in a sadistic sort of way. A bit like watching someone slip on a banana peel.

The source of this humor is that the e-mail game highlights a serious and real concern: that successful coordination in many important situations is hampered by strategic uncertainty due to incomplete high-order knowledge. The relentless logic of iterated elimination of dominated strategies takes this realistic phenomenon to an absurd extreme. This is what makes it funny: the over-the-top execution of a basically sound

logic. But the absurd humor doesn't mean the exercise has been empty entertainment. After seeing the example, we understand something—namely, the role of high-order beliefs in coordination problems—better than before.

The "artistic" response to the e-mail game doesn't require us to know its broader context, the evolution of game theory, and its role in economics. A "scientific" response does. The e-mail game was a watershed in the history of game theory. It showed the crucial role of common knowledge for strategic interactions that contain an element of a coordination problem. It was the first example to demonstrate that even an apparently small incomplete-information perturbation of a common-knowledge environment can dramatically change the game-theoretic analysis. Preoccupation with robustness to common-knowledge assumptions was in the air. Around the same time, Robert Wilson issued his famous "Wilson critique," which cautioned against mechanism-design exercises that rely on common-knowledge assumptions.[5]

Even more than that, the e-mail game is the first example in the economics literature that I am aware of that demonstrated the behavioral implications of high-order beliefs in situations of incomplete information. The 1970s were the heyday of "information economics," showing that asymmetric information can have dramatic effects on economic interactions, but the examples that economists thought about in the 1970s and 1980s involved only "first-order" asymmetric information: one player knew something, another player didn't. In the e-mail game, players may both know that the situation is ripe for an attack, but coordination will be thwarted because of a small asymmetry in their high-order information.

All these heady considerations were latent in Rubinstein's 1989 paper. But the immediate experience of reading or teaching the paper is simply that it is funny—the best piece of high humor in modern economic theory that I am aware of.

Global Games

In 1993, Hans Carlsson and Eric van Damme published a wonderful paper that offered a general treatment of a class of games like the e-mail game.[6] These games have a coordination component that is captured by some parameter. In a complete-information version, as a result of this coordination effect, there are multiple Nash equilibria for some parameter values, but there are strictly dominant actions for other

parameter values. We perturb the game by introducing uncertainty regarding this parameter, such that there can never be common knowledge of its true value.

Carlsson and van Damme referred to this class of games as "global games." The "global" aspect of the game is the influence of certain regions of the space of parameter values on players' behavior in very distant regions, due to strategic reasoning.

While Rubinstein analyzed a specific example, Carlsson and van Damme offered a general analysis of global games. Nevertheless, they did make use of a leading example. The payoff function is given by figure 2.2, which is a tiny variant on figure 2.1.[7]

Now perform two additional changes. First, while in the e-mail game x takes two possible values, suppose now that x can take any real value in the interval [–2, 2]. Second, players' information regarding the value of x follows a different protocol. Player 1 observes a signal $t_1 = x + e_1$, and player 2 observes a signal $t_2 = x + e_2$, where e_1 and e_2 are independent random variables that are uniformly distributed on the interval $[-\varepsilon, \varepsilon]$, where $\varepsilon > 0$ should be viewed as a small number. That is, each player doesn't get to see the number x with absolute precision. Instead, he gets to see x with some noise. The smaller ε, the smaller the noise. The limit $\varepsilon \to 0$ corresponds to "almost common knowledge," in much the same way that a large number on players' computer screens captured "almost common knowledge" in the e-mail game. These are two different notions of "almost." Each of them makes sense in terms of its underlying information technology.

Carlsson and van Damme showed that the game has an essentially unique Nash equilibrium, in which each player attacks when he receives a signal above ½ and refrains from attacking when he receives a signal below ½.[8] This is remarkable. Even if $x = 0.49$ and ε is extremely small, such that players observe x with arbitrarily high precision, they will almost surely coordinate on a suboptimal outcome.

	Attack	Don't attack
Attack	x, x	$x - 1, 0$
Don't attack	$0, x - 1$	$0, 0$

Figure 2.2
Payoffs in the Carlsson–van Damme game.

Carlsson and van Damme's result highlights a feature that was only latent in the e-mail game, and that is the role of *risk dominance*. An action is risk dominant if it maximizes the player's expected payoff against a uniform belief over the other player's actions. In the payoff function given by figure 2.2, attacking is risk dominant when $x > ½$ and not attacking is risk dominant when $x < ½$. Thus, when players' signals are arbitrarily precise, Nash equilibrium selects the risk-dominant action.

Like robustness to common knowledge, the notion of risk dominance was also "in the air" when Carlsson and van Damme performed their exercise. John Harsanyi and Reinhard Selten had introduced the concept in a recent book.[9] Evolutionary game theorists showed how risk-dominant actions are selected by evolutionary dynamics in which players "learn" to play coordination games.[10]

The proof of Carlsson and van Damme's result, like Rubinstein's, is based on iterative elimination of strictly dominated strategies. In the first step, we consider negative values of a player's signal t. For such values of t, the expectation of x conditional on t is below zero, such that attacking is strictly dominated. Thus, players will not attack when they see a negative signal. But now consider the case of a small, positive signal. The player believes that, in expectation, x will be equal to t, such that coordinated attack would bring a small benefit. However, when t is close to zero, the probability that the other player received a negative signal is close to ½. Therefore, the probability that the other player attacks cannot be significantly greater than ½. Because the expectation of x conditional on t is small, the expected gain from a coordinated attack is small compared with the cost of a solo attack. Therefore, the player will prefer not to attack. Thus, in the second round of the iterative procedure, we eliminate strategies that prescribe attacking to small positive values of t. In the third round, we eliminate strategies that prescribe attacking to slightly higher values of t. And in the following rounds, we keep gobbling up regions of t up to ½, such that after infinitely many rounds, we eliminate all strategies that prescribe attacking to signals below ½.

The case of signals above ½ is a mirror image. In the first round, we eliminate strategies that prescribe not attacking to signals above *1*. In subsequent rounds, we eliminate strategies that prescribe not attacking to lower signals, and after infinitely many rounds, we eliminate all strategies that prescribe not attacking to signals above ½. This leaves us with a strategy of attacking when $t > ½$ and not attacking when $t < ½$ as the essentially unique outcome of successive elimination of strictly dominated strategies.

Paradox? What Paradox?

Although the structure of players' incomplete information is different in the Rubinstein and Carlsson–van Damme games, they both lead to a unique Nash equilibrium that is obtained by iterative elimination of dominated strategies, featuring similar strategic reasoning. One might therefore expect Carlsson and van Damme to treat their result as "paradoxical," just as Rubinstein did. Yet Carlsson and van Damme very emphatically deny that their result is paradoxical. Instead, they claim that it is a *useful* result that resolves the indeterminacy of the coordination game under common knowledge. Recall that when the value of x is commonly known (which corresponds to $\varepsilon = 0$ in their example), there are two "pure" Nash equilibria when x is between 0 and 1: coordinated attack and coordinated failure to attack. The latter is inferior to the coordinated attack outcome, but as far as Nash equilibrium is concerned, it is an equally valid prediction.

Unlike Rubinstein, Carlsson and van Damme talk about prediction without scare quotes. They regard Nash equilibrium as a recipe for predicting outcomes in games—and note that the recipe is only partially satisfactory because of its indeterminacy when x is between 0 and 1. They subject the game to a realistic perturbation, such that players do not observe x with complete precision—who can ever observe anything with complete precision?—et voilà! The same recipe delivers a crisp, unique prediction that seems to make sense: players coordinate on the risk-dominant action.

For Carlsson and van Damme, there is no paradox: the unique equilibrium is merely a consequence of applying the same conventional solution concept to a tiny variant on the original game; and moreover, this variant is more realistic than the original game because it relaxes the far-fetched assumption that players observe the state of nature with absolute precision.

Thus, while Rubinstein's and Carlsson and van Damme's examples are very similar, their surrounding rhetoric couldn't be more different. Rubinstein invites his readers to mock his "prediction" and explicitly frames it as paradoxical, whereas Carlsson and van Damme invite the reader to think of the result as bringing us closer to a realistic and valuable prediction. Consequently, they call on their readers to go out and seek areas of economic activity that exhibit indeterminacies due to coordination effects and impose a similar incomplete-information perturbation in order to get unique predictions.

This call was heeded. Morris and Shin (1998) was an influential model of currency attacks, based on the idea that speculators' incentive to attack a currency depends on their beliefs about economic fundamentals and other speculators' behavior, in a way that resembles the coordinated attack problem. Goldstein and Pauzner (2005) revisited the well-known Diamond-Dybvig model of bank runs. This is a scenario in which an individual depositor's decision whether to withdraw his money from the bank depends on his assessment of the bank's solvency as well as his belief regarding other depositors' behavior. There are many more examples; this is not the place for a serious list. Morris and Shin's (2003) review article would be a good starting point for interested readers. Because these models are written in the applied-theory mode, their assumptions are meant to approximate a concrete economic environment. This means that they do not always fall neatly into the rigid global game framework, and some analytical work is needed to bridge this gap. But the main thrust of these works emanates from the Carlsson–van Damme example.

Between the Absurd and the Applied

How can two examples that are so similar give rise to such different responses? Both examples introduce small incomplete-information perturbations into the same underlying game. Although the perturbations are different, they lead to the same prediction: the risk-dominant action is taken as the consequence of iterated elimination of strictly dominated strategies. The proof method is basically the same. How could the same result lend itself to a "paradoxical" or an "applied" pitch at the authors' pleasure?

I can think of a few explanations. First, explicit intentions matter. Rubinstein announces his result as a paradox, while Carlsson and van Damme announce theirs as a prediction without scare quotes. The authors essentially tell their readers how to think about their results, and readers usually do as they are told.

Going into details, the "states of nature" in the two examples are different. In Rubinstein's example, the state is binary, whereas in Carlsson and van Damme's it is continuous. Continuous variables tend to convey a "realistic" impression, whereas binary variables are often used for pedagogical or "merely illustrative" purposes. The enemy's strength is not *really* binary; there are many degrees of strength. Therefore, an example that describes it as a continuous variable announces itself as

more "descriptive" than an example that describes it as a binary variable.

Furthermore, the players' noise structure has an "applied" connotation in Carlsson and van Damme's example. The typical reader has seen countless examples of applied-economics exercises in which decision-makers observe a real-valued economic variable with additive noise. Usually the noise in such works is normally distributed, rather than uniformly distributed as in Carlsson and van Damme's example. And indeed, when Morris and Shin present their version of the example in their 2003 review, they use normal noise distributions. This lends an air of "applied economics" to the exercise. In contrast, the elaborate e-mail protocol in Rubinstein's example has been constructed for the specific purpose of this example. No "applied-economics" paper has ever used anything like it.

Viewing this from outside the economics culture, a reader might think this is getting things backward. Rubinstein's protocol describes a concrete mechanism for generating asymmetric information, based on an actual technology. And everyone has had experience with messages that fail to reach their destination! In contrast, the additive noise specification is obviously a mathematical abstraction. Rubinstein's protocol is more tangible and, in this sense, more realistic than Carlsson and van Damme's abstract specification. Nevertheless, the conventions of economic theory condition us to treat the former as "artificial" and the latter as "realistic."

These factors may explain why we are primed to view Carlsson and van Damme's game in "applied" terms. But why don't we think of the result itself as absurd, given that it has the same underlying reasoning as Rubinstein's? Morris (2002) grappled with this question. He claimed that players' equilibrium strategy in Carlsson and van Damme's example can be described as a heuristic of responding to a "Laplacian" belief that the opponent is equally likely to play the two actions. In other words, it is natural and simple, and doesn't require sophisticated strategic reasoning. But so is the equilibrium strategy in the e-mail game! What can be simpler than playing the same action regardless of one's information?

In my opinion, there are two reasons for our tendency not to be "outraged" by Carlsson and van Damme's prediction. First, in their example, a player's signal t plays a double role: (1) it gives him information about the value of x, which determines the value of a successful attack; (2) it measures the player's layer of mutual belief that efficient coordination

is possible. The latter role mirrors the number on the player's computer screen in the e-mail game, but this role is masked by the first role. The e-mail game throws players' degree of mutual knowledge in the reader's face; Carlsson and van Damme's example conceals it behind a payoff-relevant detail.

Second, consider our instinctive assessment of the difference between a few key numbers—the cutoff value $t = ½$ that determines whether players attack, and the values 0 and 1 of x at which attacking becomes a dominant or dominated action. The difference between ½ and 0 doesn't seem large because it is on the game's payoff scale. Therefore, it doesn't surprise us that players might demand a "cushion" that protects them against the risk of a miscoordinated attack. In fact, the appropriate unit of measurement for gauging the difference is ε, which quantifies the precision of players' signals. When ε is infinitesimal, a signal $t = 0.4$, say, is "infinitely larger" than $x = 0$ in these terms, and therefore the model effectively predicts that players demand an *infinitely large* safety cushion in order to coordinate with their opponent. This pitch sounds more paradoxical, doesn't it? Thus, while Rubinstein's framing of the information structure invites us to regard a huge number on the computer screen as an invitation to be supremely confident that the opponent realizes that coordinated attack will be successful, Carlsson and van Damme's framing obscures this—the difference between 0.4 and 0 looks small, not like the arbitrarily large multiple of ε that it is.

We see that small stylistic and rhetorical differences can make all the difference between viewing a stark result as a credible, useful prediction or as a funny paradox. Such is the distance between the applied and the absurd in economic theory.

Holdups and Ultimatums

Global games are not an isolated example of this fine line. Here is another example, which is a key building block in the modern theory of the firm. It played a crucial role in the development of the theory of incomplete contracts.[11] Imagine a worker who is about to enter a venture with a firm. Before doing so, she decides whether to make an investment in firm-specific human capital. The cost of this investment is c, where $0 < c < 1$. Prior to the investment, the value of the output she can produce for the firm is 1. After the investment, it jumps to 2. Because the gain from the investment outweighs the cost, investing is the economically efficient thing to do.

If the two parties can sign an advance contract saying, "If the agent makes the investment, she commits to produce X for the firm and receive W in return," they can bargain ex-ante over the value of W. Conventional bargaining models with complete information predict immediate agreement on some value W. The efficient outcome will prevail.

But now suppose that such contracts are infeasible. The product X is impossible to define before it has been developed, and a contract that doesn't specify exactly what X is cannot be enforced by the courts. The ability to describe X arises only *after* the worker has made her investment. Only at that stage can the two parties bargain over the division of surplus. A typical telling of this story doesn't specify the bargaining process and instead assumes that the worker's share in the surplus is some $\lambda < 1$.

But what is the divided surplus? By the time the two parties enter the bargaining, whatever investment the worker has made is a *sunk cost*. Therefore, her rational calculation will ignore it. The relevant surplus for the bargaining process is 1 if the worker did not make a prior investment, and 2 if she did. Given that her share in the surplus is λ, the worker's benefit from making the investment is $\lambda \cdot (2-1) = \lambda$. If $\lambda < c$, the worker will not make the investment, and the efficient outcome will not prevail.

This is the holdup problem: when parties cannot write advance contracts, their incentive to make efficiency-enhancing investments is dampened because they anticipate that the future bargaining process will treat these investments as irrelevant sunk costs.

Where is the lurking paradox in this story? Let's look at the bargaining process. Consider the extreme case of $\lambda = 0$, where the holdup problem is at its worst. This value of λ means that the firm has all the bargaining power in its relationship with the worker. In conventional game-theoretic models of bargaining, this extreme bargaining power can derive only from the assumption that the firm makes all the offers. In the simplest case, the firm makes a single take-it-or-leave-it offer to the worker.

But, of course, this bargaining protocol is known as the *Ultimatum Game*. A proposer offers a division of some amount of money. The responder says yes or no. If he rejects the offer, no one gets anything. A huge experimental literature, starting with the seminal paper by Güth, Shmittberger, and Schwarze's (1982), documents people's behavior in this take-it-or-leave-it bargaining game. The experiments are usually run over small stakes, although enterprising experimentalists have been

able to run them over reasonably large stakes—for example, by running NSF-funded experiments in poorer countries.[12] The robust finding is that the proposer makes an offer that is substantially far from claiming the entire surplus for himself. The modal offer in low-stakes experiments is a fifty-fifty split of the surplus. In the rare occasions that an offer gets dangerously close to the standard prediction, the responder usually rejects it.

Like a few other classic experiments in the history of behavioral economics, this one didn't really have to be performed. Our intuition about it is so robust that we could carry it entirely in our head as a thought experiment, like the e-mail game. As Colin Camerer quipped, only economists find the Ultimatum Game surprising.[13] Indeed, in the early days following the Ultimatum Game, economists proposed various outlandish explanations for this experimental finding. When the dust settled, I think that there was one clear winner, having to do with perceptions of *fairness*. The selection of the two parties into the proposer-responder roles is arbitrary. As a result, the responder doesn't think that the proposer's first-mover advantage entitles him to a disproportionate share of the surplus, and therefore resents the proposer when he behaves as if he *is* entitled. Might doesn't make right. The responder is willing to give up money to express this resentment. Anticipating this sentiment, the proposer is reluctant to antagonize the responder with an unfair offer.

One strand in the voluminous experimental literature explored what can affect the responder's fairness judgments. For example, suppose the identity of the responder is not random, but selected according to a prior trivia quiz. In this case, the proposer *did something* to get the first-mover advantage, and therefore it is more acceptable if he exploits it. Offers in this variant on the Ultimatum Game are somewhat more favorable to the proposer than in the bare-bones version.[14]

But now let us return to the holdup problem with $\lambda = 0$. Not only is the bargaining process following the worker's investment equivalent to the Ultimatum Game, but the parties' behavior prior to the bargaining phase also intensifies the *responder's* sense of entitlement. We can imagine her fuming (expletives deleted): "I made this sacrifice, learning new skills and acquiring new technologies, losing sleep and risking a divorce, and now you're telling me that I should disregard it because it's a *sunk cost*?! So that you can enjoy all the benefits of my investment?!" In other words, the protocol of the holdup game doesn't mitigate the fairness considerations that the Ultimatum Game has revealed; on the contrary, it makes them more prominent. An astute employer

will recognize it and make a generous offer to the worker. From this point of view, the sunk cost actually strengthens the worker's bargaining position because it lends credibility to her threat to burn all bridges if she doesn't get her fair share. It's the exact opposite of the usual sunk-cost story. (Of course, when stakes are large, we shouldn't expect a fifty-fifty split, but an allocation that lies somewhere between this benchmark and the standard, proposer-take-all prediction.)

The paradox that lurks underneath the holdup problem is that its standard economic argument runs against the fairness-based interpretation of the Ultimatum Game. Why are we willing to look the other way and pretend that the Ultimatum Game argument is irrelevant to the holdup problem? Somehow, we have managed to compartmentalize our knowledge. Yes, we know that the Ultimatum Game is one of the most robust and frequently run experiments in the history of experimental economics, and we realize that it will upset the classical argument in the holdup problem, upon which such an important literature has been erected. But we seem to have this tacit agreement not to mix these two pieces of knowledge.

One can argue that economists use experimentally refuted theories all the time. For example, we regularly use expected utility theory despite classic experimental refutations like the Allais paradox.[15] The analogy is not accurate. When we apply expected utility theory, we usually don't rely on the specific configuration of Allais's experiment. In contrast, the holdup problem is a specific argument about the role of sunk costs in bilateral bargaining, which runs against the insights we learned from the Ultimatum Game.

This example illustrates yet another variety of the phenomenon that this chapter has examined. Here it is a matter of our willingness, or lack thereof, to approach an economic application from a slightly different perspective that would link it to a different body of literature within economics (in this case, experimental economics) and absorb the lessons this literature might teach us. If we do look at this other literature, the application suddenly becomes "paradoxical."

A Tight Space

This is the condition of economic theory: paradox can always be just around the corner. Move a bit away from it, and you have a triviality. Move a bit toward it, and you have a result no one can trust. The space in which you can use the tools of economic theory to say something that

is not trivial and has some credibility is tight. Rhetoric, stylistic tricks, and arbitrary conventions can determine whether you land in the area of paradox or away from that cliff.

In his Econometric Society presidential address, Rubinstein (2006) referred to the "dilemma of absurd conclusions"—namely, the fact that any economic model can be twisted and extended to the point where it will deliver paradoxical results. What this chapter has shown us is how apparently minor and superficial details of the model's delivery and its surrounding rhetoric can bounce us back and forth between the absurd and the applied.

The reader may think that, by making such a claim, I am diminishing economic theory. I don't think I am. That the serious and the grotesque can be very close is a fact of life. If living in the post–November 2016 world has taught us anything, it is that sometimes, ridiculous things should be taken very seriously (as if we hadn't known this already). That economic theory can accommodate this irony is a measure of its ability to portray an essential aspect of human interactions.

Index