EXPLAINING SOCIAL BEHAVIOR

More Nuts and Bolts for the Social Sciences

JON ELSTER
COLLÈGE DE FRANCE

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
32 Avenue of the Americas, New York, NY 10013-2473, USA
www.cambridge.org
Information on this title: www.cambridge.org/9780521771795

First published 2007

Printed in the United States of America

A catalog record for this publication is available from the British Library.

Library of Congress Cataloging in Publication Data
Elster, Jon, 1940–
Explaining social behavior : more nuts and bolts
for the social sciences
/ Jon Elster
p. cm.
Expanded and rev. ed. of: Nuts and bolts for the social sciences, 1989.
Includes bibliographical references and index.
ISBN-13: 978-0-521-77179-5 (hardback)
ISBN-13: 978-0-521-77744-5 (pbk.)
1. Social sciences – Methodology. 2. Social interaction. I. Elster, Jon,
1940 – Nuts and bolts for the social sciences. II. Title.
H61.E434 2007
302 – dc22 2006022194

ISBN 978-0-521-77179-5 hardback
ISBN 978-0-521-77744-5 paperback

For Jonathan and Joanna

Contents

PREFACE

This book began as a revision of a book I published in 1989, *Nuts and Bolts for the Social Sciences*. It ended up as a quite different and more ambitious kind of book. It covers a much greater variety of topics, in considerably more detail, and in a different spirit. Although nine chapters have the same headings as chapters in the earlier book, only Chapter 9 and Chapter 24 remain substantially the same.

Although comprehensive in scope, the book is not a treatise. It is both less and more than that. It is an elementary, informal, and personal presentation of ideas that have, I believe, considerable potential for illuminating social behavior. I use plenty of examples, many of them anecdotal or literary, others drawn from more systematic studies. The very occasional use of algebra does not go beyond high school level. At the same time, the book has a methodological and philosophical slant not usual in introductory-level presentations. There is an effort to place the social sciences within the sciences more generally – the natural sciences as well as the humanities. There is also an effort to make the reader keep constantly in mind how general principles of scientific explanation constrain the construction of theories with explanatory pretensions.

The style of the bibliographical notes to each chapter reflects the rise of the Internet, in particular of Wikipedia, Google.com, and Scholar .Google.com. Since readers can find most relevant references in a matter of minutes, I have omitted sources for many of the statements and findings in the text. Instead I try to point readers to important sourcebooks, to some modern classics, to books and articles that are the sources of claims that might be harder to track down on the Internet, and to authors from whom I have taken so much that not mentioning them would justify a pun on my name (*Elster* in German means magpie).

Although the main text contains few references to contemporary scholars, I refer extensively to Aristotle, Seneca, Montaigne, La Rochefoucauld, Samuel Johnson, H.C. Andersen, Stendhal, Tocqueville, Proust, and other classical writers who remain literally inexhaustible

Chapter 19

STRATEGIC INTERACTION

❀ ❀ ❀

Strategic Interaction with Simultaneous Choices

The invention of *game theory* may come to be seen as the most important single advance of the social sciences in the twentieth century. The value of the theory is partly explanatory, but mainly conceptual. In some cases it allows us to explain behavior that previously appeared as puzzling. More important, it illuminates the structure of social interaction. Once you see the world through the lenses of game theory – or "the theory of interdependent decisions," as it might better be called – nothing looks quite the same again.

I first consider games in which agents make simultaneous decisions. The goal is to understand whether and how n agents or *players* may achieve an unenforced coordination of their *strategies*. Often, we shall look at the special case of n = 2. The players may be able to communicate with each other, but not to enter into binding agreements. To any n-tuple of strategies, one chosen by each agent, there corresponds an *outcome*. Each agent ranks the possible outcomes according to his or her *preference order*. When needed, we shall assume that the conditions for representing preferences as cardinal utilities are satisfied (Chapter 11). The *reward structure* is the function that to any n-tuple of strategies assigns an n-tuple of utilities. Although the word "reward" may suggest a monetary outcome, the word will be used to refer to psychological outcomes (utilities and ultimately preferences). When, as is often the case, the monetary or material reward structure and the psychological reward structure diverge, only the latter is relevant.

As briefly mentioned in the last chapter, an agent may have a strategy that is *dominant* in the sense that regardless of what others do, it yields a better outcome for her than what she would get if she chose any other strategy. Her *outcome* may depend on what others do, but her *choice* does not. In other cases, there is genuine interdependence of choices. If others

drive on the left side of the road, my best response is to drive left too; if they drive on the right, my best response is to drive right.

An *equilibrium* is an n-tuple of strategies with the property that no player can, by deviating from his equilibrium strategy, unilaterally bring about an outcome that he strictly prefers to the equilibrium outcome. Equivalently, in equilibrium the strategy chosen by each player is a best response to the strategies chosen by the others, in the weak sense that he can do *no better* than choosing his equilibrium strategy if others choose theirs. The strategy need not, however, be optimal in the strong sense that he would do *worse* by deviating unilaterally. In the general case, a game may have several equilibria. We shall see some examples shortly. Assume, however, that there is only one equilibrium. Assume moreover that the reward structure and the rationality of all players are common knowledge.[1] Under these assumptions, we can predict that all agents will choose their equilibrium strategy, since it is the only one that is based on rational beliefs about what others will do.

Some games with a unique equilibrium turn upon the existence of dominant strategies. The phrase "turn upon the existence of dominant strategies" can mean one of two things, illustrated in panels A and B of Figure 19.1.[2] In an accident involving two cars, both are harmed. In an accident involving a pedestrian and a car, only the former is harmed. Car-car accidents occur if at least one driver is careless. If both are careless, the outcome is worse. Car-pedestrian accidents occur only if both are careless. Taking due care is costly. From these premises, it follows that in the car-car case, taking care is the dominant strategy for each driver. In the car-pedestrian case, no-care is dominant for the driver.

[1] A fact is common knowledge if all know it, all know that all others know it, all know that all others know that all others know it, and so on. To avoid reliance on the phrase "and so on," which suggests an infinite sequence of beliefs, the idea may also be stated as follows: there is no n such that the fact is common knowledge up to level n in the sequence but not at level n + 1. For a simple illustration, common knowledge may be realized in a classroom. When the teacher tells a fact to the students, they all know it, know that others know it, and so on.

[2] By convention, the first number in each cell represents the payoff for the "row player" who chooses between the top and bottom strategies, and the second the payoff for the "column player" who chooses between the left and right strategies. Depending on the context, the payoffs may be cardinal utilities, ordinal utilities, money, or anything else that the players may be assumed to maximize. In Figure 19.1, payoffs may be seen as standing for ordinal utilities, reflecting preferences over outcomes. Here and later, equilibria are circled.

(A)
Car

	Due care	No care
Due care	(5, 5)	2, 3
No care	3, 2	1, 1

(B)
Pedestrian

	Due care	No care
Due care	0, 2	0, 3
No care	(1, 2)	1, 1

FIGURE 19.1

The pedestrian has no dominant strategy, since due care is the best response to no care and no care the best response to due care. Since he knows that the driver has no-care as a dominant strategy and, being rational, will choose it, the pedestrian will nevertheless choose due care.[3]

Games in which all players have dominant strategies are quite common and empirically important, as we shall see. Theoretically, they are somewhat trivial, except when they are repeated over time. Games in which some players have dominant strategies that can induce clear-cut choices in others are less common but also important. They have stronger informational requirements, however, since in our example the pedestrian needs to know the possible outcomes for the driver as well as for himself, whereas the two drivers only need to know their own outcomes. Often, we can impute dominant strategies to others without much trouble. We do not usually, for instance, look both ways before crossing a one-way street because we assume that the fear of drivers of being liable for an accident will make them obey the one-way rule.

[3] According to some legal analyses, an important function of tort law is to use the system of fines and damages to change the reward matrix so that the emerging equilibrium has some desirable property (efficiency or fairness).

A special class of games has *coordination equilibria*, often called "conventions," in which each player not only has no incentive to deviate unilaterally, but also would prefer that nobody else does so. In an equilibrium in which everybody drives on the right side of the road, an accident might occur if I deviate *or* if anyone else does. In this case, the equilibrium is not unique, since driving on the left side has the same properties.[4] Often it does not matter what we do as long as we all do the same thing. The meanings of words are arbitrary, but once they are fixed, they become conventions. In other cases, it does matter what we do, but it is more important that we all do the same thing. I return to some examples shortly.

Two Duopoly Examples

Some games have unique equilibria that do not turn upon the existence of dominant strategies. Duopoly behavior is an example (see Figure 19.2).

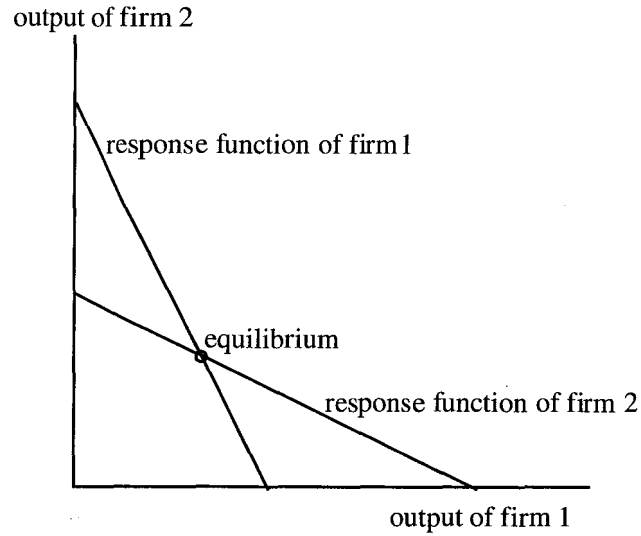

FIGURE 19.2

[4] Although the nonuniqueness does not follow from the formal definition, this seems to be a general feature of real-life coordination games.

When two firms dominate a market, lower production by one firm will induce higher prices and an expansion of production by the other firm. In other words, each firm has a "best response" schedule that tells it how much to produce as a function of the output of the other firm. In equilibrium, the output of each firm is a best response to the output of the other. This statement does not imply that they could not do better. If they formed a cartel and restricted their production to below-equilibrium levels, both would earn greater profits. Yet these collectively optimal levels of production are not best responses to each other. The firms are, in fact, facing a Prisoner's Dilemma (defined in Figure 19.3).

For another case of duopoly, consider two ice cream vendors on a beach, trying to find the best location for their stalls on the assumption that customers (assumed to be evenly distributed across the shoreline) will go to the closer stall. There is no dominant strategy. If one of them puts up a stall some distance left of the middle of the beach, the best response of the other is to position himself immediately to the right, to which the best response of the first is to move right again, and so on, until their stalls are beside each other at the middle of the beach. This unique equilibrium is obviously not the best for the customers in the aggregate. For them, the best outcome is one in which each stall is positioned halfway between the middle and one end of the beach. Although this outcome is just as good for the sellers as the equilibrium outcome, these positions are not best responses to each other. This model has also been applied to explain the tendency for political parties (in a two-party system) to move toward the middle of the political spectrum.

Suppose, however, that when both stalls are at the middle customers close to the ends abstain from buying ice cream because it would melt by the time they walked back. If no customer is willing to walk more than half the length of the beach, one-quarter to get to the stall and one-quarter to get back, the optimal consumer outcome is also the unique equilibrium since neither has an inventive to relocate. Suppose the beach is 1,000 meters long. If the seller at 750 meters moves his stall to 700, he will lose the 50 customers between 950 and 1,000 who are not willing to walk more than 500 yards and gain the 25 customers between 475 and

500 to whom his stall is now closer than the other – a net loss. A similar argument might also explain why political parties never converge fully to the middle, since extremists at either end might prefer to abstain rather than vote for a centrist party. In addition, as I noted at the end of Chapter 17, it is simply not plausible to view vote maximization as the only aim of political parties.

Some Frequently Occurring Games

A few simple interaction structures, with payoffs as in Figure 19.3, occur very often in a great variety of contexts.[5] C and D stand for "cooperation" and "defection." In the Telephone Game the column player is the one who first called. In the Focal Point Game, A and B can be any pair of actions such that both players would prefer to coordinate on either than not to coordinate but are indifferent between the two ways of coordinating.

The games illuminate the structure of the two central issues of social interaction – *cooperation* and *coordination*. In a society with no cooperation for mutual benefit, life would be "solitary, poor, nasty, brutish, and short" (Hobbes). That it would be *predictably* bad is a meager consolation. In a society where people were unable to coordinate their behavior, unintended consequences would abound and life would be like "a tale told by an idiot, full of sound and fury, signifying nothing" (*Macbeth*). Both cooperation and coordination sometimes succeed, but often fail abysmally. Game theory can illuminate the successes as well as the failures.

The Prisoner's Dilemma (PD), the Stag Hunt, and Chicken involve in one way or another the choice between cooperation and defection (noncooperation). The Prisoner's Dilemma is so called because the following story was used to illustrate it in an early discussion. Each of two prisoners, who have been involved in the same crime but are now in separate cells, is told that if he informs on the other but she does not

[5] Although stated here as two-person games, they easily generalize to the case of many agents. An n + 1-person version of the Prisoner's Dilemma, for instance, is illustrated in Figure 24.2.

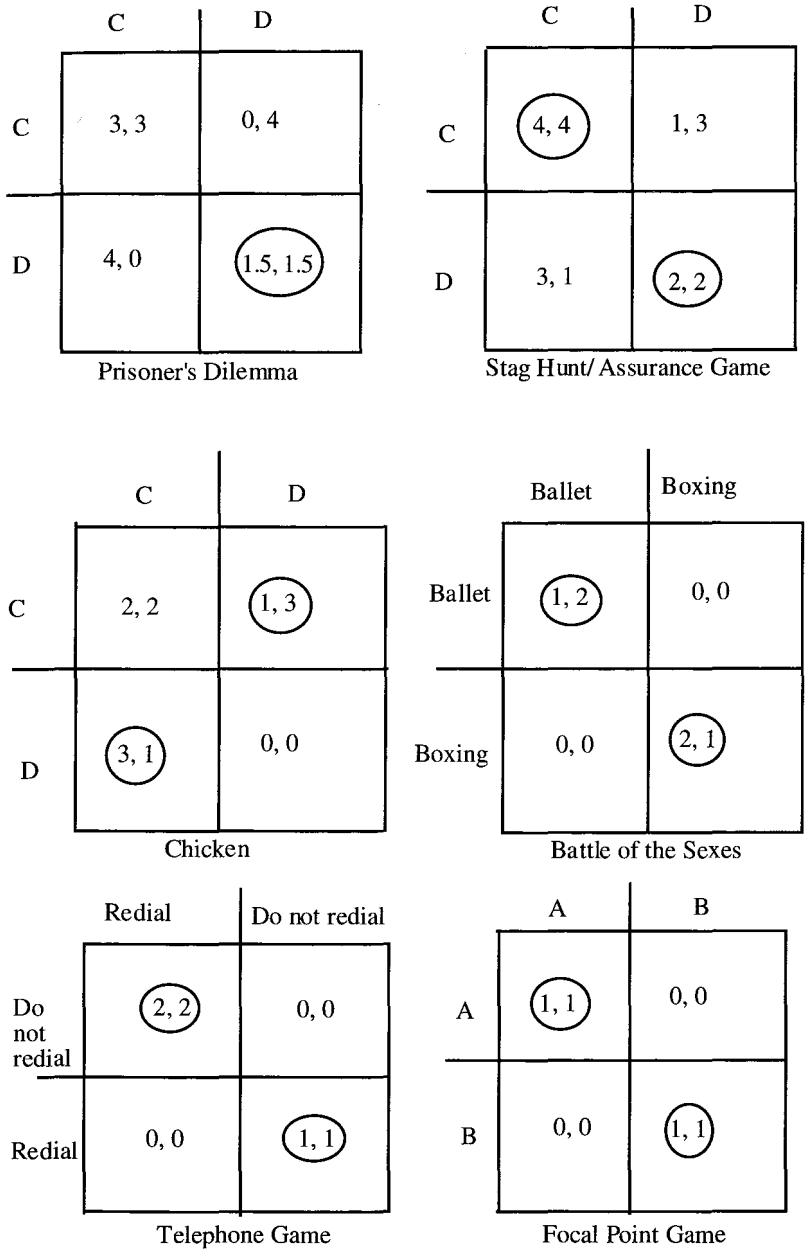

FIGURE 19.3

inform on him, he will go free and she will go to prison for ten years; if neither informs on the other, both will go to prison for one year; and if both inform on each other, both will go to prison for five years.[6] Under these circumstances, informing is a dominant strategy, although both would be better off if neither informed. The outcome is generated by a combination of the "free-rider temptation" (going free) and the "fear of being suckered" (getting ten years).

The negative externalities discussed in the last chapter can also be viewed as many-person PDs. Some other examples follow. For each worker (assuming selfish motivations) it is better to be nonunionized than to join a union, even when it is better for all if all join and gain higher pay. For each firm in a cartel it is better to break out and produce a high volume to exploit the high prices caused by the output restrictions of the other firms, but when all do that, prices fall to the competitive level; profit maximization by each firm undermines the maximization of joint profits. The Organization of Petroleum Exporting Countries (OPEC) cartel is vulnerable in the same way. Other examples are situations in which everybody has to run as fast as he can to stay in the same place, such as the arms race between the United States and the former Soviet Union, political advertising, or students writing papers for a teacher who "grades on the curve."

The idea of the Stag Hunt is often imputed to Jean-Jacques Rousseau, although his language was somewhat opaque.[7] In more stylized form, it involves two hunters who can choose between hunting stag (C) and a hare (D). Each can catch a hare by himself, but the joint effort of both is necessary (and sufficient) to catch a stag. Half a stag is worth more than a hare. It takes more time and effort to catch hares when both are trying because the noises the hunters make scare them away. As in the Prisoner's

[6] The payoffs for the Prisoner's Dilemma in Figure 19.3 might seem artificial. For the present purposes, all that matters is the (ordinal) ranking of the outcomes. Later, the payoffs will be reinterpreted as monetary rewards.

[7] "If a deer was to be taken, every one saw that, in order to succeed, he must abide faithfully by his post: but if a hare happened to come within the reach of any one of them, it is not to be doubted that he pursued it without scruple, and, having seized his prey, cared very little, if by so doing he caused his companions to miss theirs." This could be read as saying that pursuing hares is a dominant strategy.

Dilemma, there is a risk of being a sucker, hunting for stag while the other goes for a hare. There is no free-rider temptation, however. The game has two equilibria, in the upper left-hand and lower right-hand cells.

Although the first equilibrium is clearly better, it may not be realized. To see why this might happen we can drop the assumption that the payoff structure is common knowledge and allow the agents to have mistaken beliefs about the payoff structure of other agents. Actions taken on these beliefs will form an *equilibrium in a weak sense* if, for each agent, the actions taken by the others confirm his beliefs about them. Assume, for instance, that in a Stag Hunt each agent falsely believes the others to have PD preferences. Given that belief, the rational action is to defect, thus confirming the belief of the others that *he* has PD preferences. This society might end up with high levels of tax evasion and corruption. I return to such cases of "pluralistic ignorance" in Chapter 23. In another society, where people correctly believe others to have Stag Hunt preferences, a good equilibrium will emerge in which people pay their taxes and do not offer or take bribes. "Cultures of corruption" might be a belief-dependent, not a motivation-dependent, phenomenon.

International control of infectious diseases can have the structure of a Stag Hunt. If only one country fails to take the appropriate measures, others will not be able to protect themselves.[8] For another example, consider counterterrorist measures. If only one of two nations invests in such measures, it benefits the other as well as itself. If the costs exceed the benefits to itself, it will not invest unilaterally. Yet if both invest, the ability to pool information may lead to a greater security level for each than it could achieve by exploiting the investment of the other.

In these examples, the payoff structure arises from the causal nature of the situation. In the Stag Hunt and the disease control case, the "threshold technology" implies that individual efforts are pointless. In the counterterrorism case, the underlying cause is something like economies of scale: ten units of effort have more than twice the effect of five units. In other cases, the payoff structure is due to the fact that the

[8] This is a huge simplification, made simply for the sake of illustration.

agents care for other things than their own material rewards. In such cases, it is more common to refer to the game as an Assurance Game (AG). Even if the material payoff structure is that of a PD, each individual may be willing to cooperate if he is *assured* that others will. The desire to be fair, or the reluctance to be a free rider, may overcome the temptation to exploit the cooperation of others. Alternatively, altruistic preferences may transform a PD into an AG.

Let us interpret the payoffs in the PD in Figure 19.3 as monetary rewards and assume that each person's utility equals his monetary reward plus half the monetary reward of the other. In that case, the utility payoff will be as in Figure 19.4 – an AG. The PD may also be transformed into an AG by a third mechanism, if an outside party attaches a penalty to the choice of the noncooperative strategy D. If we again interpret the payoffs in the PD in Figure 19.3 as monetary rewards, *and* assume that this is all the agents care about, deducting 1.25 from the reward to defection will turn it into an AG. A labor union might, for instance, impose formal or informal sanctions on nonunionized workers. Finally, one might transform a PD into an AG by rewarding cooperation, for example, by offering a bonus or bribe of 1.25 to cooperators. Promises of reward have to be respected, however, whereas a threat does not have to be carried out if it works. If the free-rider payoff is very high, the benefits from cooperation may not be large enough to fund the bribes.[9] In some cases, though, rewards are used. Workers who join a union may benefit not only from higher wages, which usually accrue equally to nonunionized workers, but also from pension plans and cheap vacations offered only to members.

The game of Chicken is named after a teenage ritual from the 1955 movie *Rebel Without a Cause*. Los Angeles teenagers drive stolen cars to a cliff and play a game in which two boys simultaneously drive their cars off the edge of the cliff, stopping at the last possible moment. The boy who stops first is "chicken" and loses. In another variant, two cars drive toward each other and the one who swerves first is

[9] Whether one uses punishments or rewards, the costs of establishing the system and monitoring the agents also have to be funded by the gains from cooperation. In practice, this can easily make such arrangements impossible or wasteful.

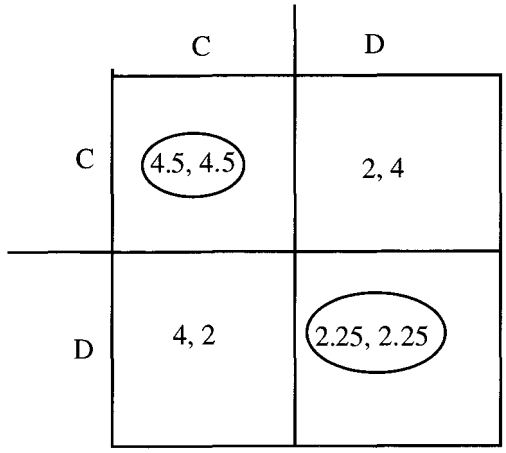

FIGURE 19.4

"chicken." In each of the two equilibria, each agent does the opposite of the other. Even with common knowledge of the payoff structure and of the rationality of the agent, we cannot predict which of the equilibria (if any) will be chosen. From the point of view of rational choice, the situation is *indeterminate*. In the second ("swerve") version of the game, a player might try to break the indeterminacy by (visibly) blindfolding himself, thus inducing the other to swerve. Yet this creates the same predicament with the two options being "blindfolding" and "not blindfolding" rather than "swerving" and "not swerving."[10] It is a deeply frustrating situation.

On one understanding of the arms race, it has the structure of Chicken. The Cuban missile crisis is often cited as a case in which the two superpowers were locked in a Chicken-like confrontation and the USSR "blinked first." Another example is that of two farmers who use the same irrigation system for their fields. The system can be adequately maintained by one person, but both farmers gain equal benefit from it. If one farmer does not do his share of maintenance, it may still be in the other farmer's interest to do so. The Kitty Genovese case can

[10] Similarly, the "solution" to the Prisoner's Dilemma that consists of each person's promising to cooperate merely recreates the PD with the choices being "keeping the promise" and "reneging."

also be seen in this perspective, if we assume that each neighbor would prefer to intervene if and only if nobody else did.

Turning now to questions of *coordination*, consider first the Battle of the Sexes. The stereotype behind the story is the following. A man and his wife want to go out for the evening. They have decided to go either to a ballet or to a boxing match after work and to settle the final choice over the telephone. His phone breaks down, however, so they have to decide by tacit coordination. They have a common interest in being together, but divergent interests about where to go. As does the game of Chicken, this game has two equilibria, coordinating on the ballet or on the boxing match. And as in that game, there is no way common knowledge of the payoff structure and of rationality will tell the couple where to meet. Once again, the situation is indeterminate.

Games of this kind arise when coordination can take many forms, all of which are better for all agents than no coordination at all, but each of which is preferred by some agents to the others.[11] In social and political life, this seems to be the rule rather than the exception. All citizens may prefer any political constitution (within a certain range of possible regimes) to no constitution at all, because long-term stability is important in enabling them to plan ahead. When the law is fixed and hard to change, one can regulate one's behavior according to it. Yet each interest group may prefer a specific constitution in the range over the others: creditors lobby for a ban on paper money in the constitution, each political party favors the electoral system that will favor it, those with a strong candidate for the presidency want that office to be strong, and so on.

Multiple coordination equilibria also arise when different societies initially develop different standards of weight, length, or volume and later discover the potential benefits from a common solution. Continental Europe and the Anglo-Saxon world retain separate standards in these areas. Unlike the case of multiple constitutional solutions, the obstacle to agreement is not permanent divergence of interest, but short-term transition costs. The choice of standard might also, however, be a game of

[11] As we shall see later (Chapter 25), this question of dividing the benefits from cooperation can also be studied within *bargaining theory*, a more specialized branch of game theory.

Chicken. Assume, implausibly, that the standard is written into the constitution as an entrenched clause (immune to amendment). Each country will then have an incentive to commit itself before the other does.

The Telephone Game is defined by the need for a rule to tell the parties what to do when a phone conversation is accidentally interrupted. There are two coordination equilibria: the redialing is done by the person who made the call in the first place or by the person who received it. Either rule is better than having both redial or neither. Yet in this case, unlike the Battle of the Sexes, one equilibrium is better for both than the other. It is more efficient to have the caller do the redialing, since he is more likely to know which number to call. Rational, fully informed agents will converge on the superior coordination equilibrium. This statement ignores, however, the cost of redialing. If the cost is large, the game becomes a Battle of the Sexes.

Consider finally the Focal Point Game, which can be illustrated by a variant of the Battle of the Sexes. The spouses have agreed to watch a movie that is playing both in movie theater A and movie theater B but have postponed the choice of venue. We assume that neither is closer or otherwise more convenient than the other. As in the Battle of the Sexes, information, rationality, and common knowledge by themselves will not tell them where to go. There might, however, be a psychological cue in the situation that will serve as a "focal point" for coordination. If the couple had their first date in theater A, this might make them converge to that location. In this case, the cue is a purely private event. In other cases, cues might be shared by a large population. Among New Yorkers, for instance, folklore says that if you get separated from your companion you meet at noon under the main clock at Grand Central Station. And even when there is no folklore, many people would still go to the railway station, since in many cities the railway station is the most important building of which there is only one.[12] Its uniqueness renders it attractive as a focal point. Noontime has the same property.[13]

[12] In New York City, those ignorant of the folklore would not go to Grand Central Station, since the presence of Penn Station makes it nonunique. Instead, they might coordinate on the Empire State Building.

[13] Although midnight, too, is a focal point, it is inferior to noontime because of the inconvenience.

This focal point effect is easily demonstrated in experiments. If you ask all members of a group to write down a positive integer (whole number) on a piece of paper and tell them that they will get a reward if all write down the same number, they invariably converge on 1. There is a unique smallest integer, but no unique largest one. In other contexts, 0 may emerge as the unique focal point. In debates during the cold war whether the United States might use tactical nuclear weapons without triggering an escalation into full-blown nuclear war, various ideas were suggested for a "bright line" that would allow limited use. In the end, it was decided that *no use* was the only focal point.

Pascal made a similar observation about the importance of custom: "Why do we follow old laws and old opinions? Because they are better? No, but they are unique, and remove the sources of diversity." Elsewhere he wrote:

> The most unreasonable things in the world become the most reasonable because men are so unbalanced. What could be less reasonable than to choose as ruler of a state the oldest son of a queen? We do not choose as captain of a ship the most highly born of those aboard. Such a law would be ridiculous and unjust, but because men are, and always will be, as they are, it becomes reasonable and just, for who else could be chosen? The most virtuous and able man? That sets us straight away at daggers drawn, with everyone claiming to be most virtuous and able. Let us then attach this qualification to something incontrovertible. He is the king's eldest son: that is quite clear, there is no argument about it. Reason cannot do any better, because civil war is the greatest of evils.

This reasoning can actually influence the choice of a king when there are several pretenders to a throne. In the choice of king in the French Restoration, Talleyrand successfully argued that the legitimate heir of the last king of France was the unique focal point that could prevent divisive conflicts. As he wrote in his memoirs, "An *imposed* King would be the result of force or intrigue; either would be insufficient. To establish a durable system that will be accepted without opposition, one must act on a principle." Later, Marx argued that the Republic of 1848 owed its existence to the fact that it was the second-best option for each of the two

branches of the royal family. Tocqueville made a similar observation to explain the stability of the rule of Napoleon III. Democracy, too, can be seen as a focal-point solution. When there are many competing qualitative grounds on which people can claim superiority – wisdom, wealth, virtue, birth – the quantitative solution of majority rule acquires unique salience. Former colonial countries in which tribes speak different languages may choose the language of the colonizer for official purposes. Litigating parties easily converge on a proposal that is everybody's second-best option.

In 1989, the reburial of Imre Nagy provided a focal point for 250,000 people to march in the streets of Budapest to signal their disaffection with the regime. As in the previous examples, the focal point allowed cooperation through coordination. In conflictual situations, focal points can have quite different effects. In the Crimean War the French general Pélissier decided to stage the second attack on Sebastopol on June 18, 1855, because he wanted to please Napoleon III by gaining a victory on the anniversary of the Battle of Waterloo. As this date and its importance to the French were common knowledge, the Russians were able to anticipate and defeat him.

One lesson from this survey is that a given real-world situation can be modeled as several different games, depending on additional assumptions. The arms race has been modeled as a PD, as Chicken, and as an AG. Joining the labor union may be a PD or an AG. Redialing has been seen as a Battle of the Sexes or as a Telephone Game. Coordination of weights and measures could be a game of Chicken or Battle of the Sexes. The fine grain of interaction structures may not be immediately visible. By forcing us to be explicit about the nature of the interaction, game theory can reveal unsuspected subtleties or perversities.

Sequential Games

Let me turn more briefly to games in which agents make *sequential decisions* (I discuss such games at greater length in the next chapter) and begin by a simple example that demonstrates the power of game

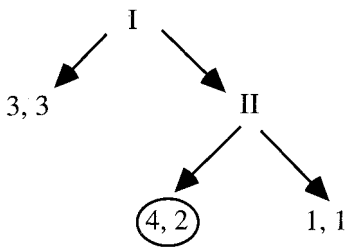

FIGURE 19.5

theory to clarify interaction structures that were only dimly understood earlier.[14]

In Figure 19.5, two armies are confronting each other at the border of their countries. General I can either retreat, leaving the status quo (3, 3) in place, or invade. If he invades, General II can either fight, with outcome (1, 1), or concede a contested piece of territory with outcome (4, 2). Before I makes his decision, II may be able to communicate an intention to fight if attacked, hoping to induce I to choose (3, 3) rather than (1, 1). However, *this threat is not credible.* I knows that once he invades, it will be in II's interest to concede rather than to fight. The unique equilibrium outcome is (4, 2). This equilibrium concept is not the static "best-response" concept we have been discussing so far. Rather, it is a dynamic concept that begins with the later stages of the game and works back to the earlier ones. (The technical term is "backward induction.") First, we ask what it would be rational for II to do if I invaded. The answer, "Concede," leads to the outcome (4, 2). I's choice, therefore, is between a course of action leading to (3,3) and one leading to (4,2). Being rational, he chooses the latter.

As Thucydides observed in *The Peloponnesian War*, promises also have to be credible for the other side to base its behavior on them.

> Oaths made in support of any reconciliation had only momentary validity, as they were made by each side only in the absence of any other source of strength to get out of an impasse; but whoever found the opposition off-guard at a given moment and seized the first opportunity for a bold strike,

[14] I retain the assumption that rationality and information are common knowledge.

> enjoyed a revenge sweeter for having exploited good faith than winning in an open fight. . . . For no word was reliable enough, nor any oath formidable enough, to bring about reconciliation, and all who found themselves in a superior position, figuring that security could not even be hoped for, made provisions to avoid injury rather than allow themselves to trust anyone.

The person who received the promise, in other words, should ask himself whether it would be rational for the promiser to keep his word. Allowing for communication in the Trust Game (Chapter 15), for instance, the second player might try to induce the first to make a large transfer by promising to make a large back transfer. If there is nothing to hold him to his word, the promise is not credible. In *Democracy in America*, Tocqueville comments sarcastically on a letter of the secretary of war to the Cherokees, in which he "states that they must abandon hope of retaining the territory they presently occupy, but he offers them the same positive assurance once they have crossed the Mississippi, *as if he will then have the power he now lacks.*" Economic reform in China has been vulnerable to a similar problem. When the government introduced market reforms in agriculture it promised the farmers fifteen-year leases on the land to give them an incentive to improve it. Since there is no way of holding an autocratic government to its promise, many farmers disbelieved it and used the profits for consumption instead. An autocratic government is *unable to make itself unable* to interfere.

The notion of credibility is central in the "second-generation" game theory that began around 1975. (The first generation began around 1945.) Once we take the idea seriously, we are led to ask how agents might *invest in credibility* to lend efficacy to their threats and promises. There are several mechanisms. One is by *reputation-building*, for instance by investing in a reputation for being somewhat or occasionally irrational. Thus it has been reported that President Nixon, encouraged by Henry Kissinger, deliberately cultivated an erratic style to make the Soviets believe he might act against the American interest if they provoked him. Also, people might carry out threats when it is not in their interest to do so in order to build a reputation for toughness that will make others believe their threats on later occasions.

Another mechanism is *precommitment*, discussed in Chapter 13. There, precommitment was viewed as a second-best rational response to the agent's proclivity to behave irrationally. In the strategic context, precommitment can be fully rational. In the game depicted in Figure 19.5, General II might build a "Doomsday machine" that would automatically launch a nuclear attack on the other country in the case of invasion. If both the existence of this machine and the fact that its operation is outside the control of country II are common knowledge, it would deter the invasion. Alternatively, II might use the strategy of "burning his bridges," that is, of cutting off any possibility of retreat. Again, General I would be deterred if he knew that General II has no alternative to fighting if invaded.

In some cases, both parties may try to use precommitment to get an edge over the other. In labor-management bargaining, strike threats and boycott threats may not be credible. The management knows that as the workers have mortgages to pay and families to support they cannot afford to go on strike very long. The labor union knows that as the firm has delivery contracts to fulfill, it cannot afford to have production come to a halt. To enhance the credibility of their threats, the union might invest in a strike fund (perhaps jointly with other unions) and the management might invest in large inventories. Alternatively, the negotiators on each side might state their minimal demands and maximal offers publicly, thus making sure they will incur high reputation costs if they concede. Such a "precommitment game" might be either a PD or a game of Chicken, depending on the structure of the subsequent game.

«««

Bibliographical Note

The best elementary introduction to game theory is A. Dixit and S. Skeath, *Games of Strategy*, 2nd ed. (New York: Norton, 2004). Among more advanced treatments, I suggest F. Vega-Redondo, *Economics and the*

Theory of Games (Cambridge University Press, 2003). An encyclopedic survey with many applications is R. Aumann and S. Hart, *Handbook of Game Theory with Economic Applications*, vols. 1–3 (Amsterdam: North-Holland, 1992, 1994, 2002). Applications to specific topics are found in J. D. Morrow, *Game Theory for Political Scientists* (Princeton, NJ: Princeton University Press, 1994), and in D. Baird, H. Gertner, and R. Picker, *Game Theory and the Law* (Cambridge, MA: Harvard University Press, 1994). A classic study of conventions is D. Lewis, *Convention* (Cambridge, MA: Harvard University Press, 1969). It is largely inspired by another classic, T. Schelling, *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960), in which the idea of focal points was first expounded. Schelling's work also provided the intuitive foundation for the "second generation" of game theory, formally developed in R. Selten, "Re-examination of the perfectness concept for equilibrium points in extensive games," *International Journal of Game Theory* 4 (1975), 25–55. For various precommitment techniques in political games, see J. Fearon, "Domestic political audiences and the escalation of international disputes," *American Political Science Review* 88 (1994), 577–92. For their use in wage bargaining, see my *The Cement of Society* (Cambridge University Press, 1989).

Chapter 20

Games and Behavior

❀ ❀ ❀

Intentions and Consequences

The conceptual structure of game theory is illuminating. Does it also help us *explain behavior*? Consider the game-theoretic rationale for burning one's bridges or one's ships. This behavior could be undertaken for the strategic reasons set out in the last chapter, but also for others. The *Oxford English Dictionary* gives the following quotation from E. R. Burroughs, *Tarzan of the Apes*: "Because she had been afraid she might succumb to the pleas of this giant, she had burned her bridges behind her." This is not a piece of strategic reasoning. Rather, the woman in question seems to fear that she might yield to entreaties if she did not make it impossible for herself to do so. Even in the military sphere, such nonstrategic rationales might be as important as the strategic ones. A commander might burn his bridges lest fear of the enemy make his soldiers take flight. He might want to prevent *himself* from deserting, if he is afraid that he might give in to weakness of will. Commander A might burn his bridges or ships to signal to enemy commander B that B cannot count on A's troops' running away. This was apparently the reasoning of Cortes when, telling his sailors (credibly but not truthfully) that his fleet was not seaworthy, he burned all his ships but one. (Also, by burning the ships he could add the sailors to his infantry.) To differentiate among these various explanations, we need to determine the intentions of the agents. Actual benefits of bridge burning are neither necessary nor sufficient to establish an explanation in terms of expected benefits (Chapter 3).

Although game theory explains behavior by appealing to the intentions of the actors to bring about certain consequences, it can also account for situations in which some of the actors do not care about the consequences. Consider for instance the interaction between the European Union and the new entrants from Eastern Europe. The old member states might be tempted to impose conditions for entry that