BOWLES, SAMUEL (2004):
SOCIAL INTERACTIONS AND INSTITUTIONAL DESIGN

IN DERS.: MICROECONOMICS.
BEHAVIOUR, INSTITUTIONS AND EVOLUTION,
NEW YORK, U.A.O, CHAPTER 1, S.23-55.

The Roundtable Series in Behavioral Economics

The Roundtable Series in Behavioral Economics aims to advance research in the new interdisciplinary field of behavioral economics. Behavioral economics uses facts, models, and methods from neighboring sciences to establish descriptively accurate findings about human cognitive ability and social interaction and to explore the implications of these findings for economic behavior. The most fertile neighboring science in recent decades has been psychology, but sociology, anthropology, biology, and other fields can usefully influence economics as well. The Roundtable Series publishes books in economics that are deeply rooted in empirical findings or methods from one or more neighboring sciences and advance economics on its own terms—generating theoretical insights, making more accurate predictions of field phenomena, and suggesting better policy.

Colin Camerer and Ernst Fehr, editors

Behavioral Game Theory: Experiments in Strategic Interaction by Colin F. Camerer

Microeconomics: Behavior, Institutions, and Evolution by Samuel Bowles

Advances in Behavioral Economics, edited by Colin F. Camerer, George Loewenstein, and Matthew Rabin

The Behavioral Economics Roundtable

Henry Aaron
George Akerlof
Linda Babcock
Colin Camerer
Peter Diamond
Jon Elster
Ernst Fehr
Daniel Kahneman
David Laibson
George Loewenstein
Sendhil Mullainathan
Matthew Rabin
Thomas Schelling
Eldar Shafir
Robert Shiller
Cass Sunstein
Richard Thaler
Richard Zeckhauser

Microeconomics

BEHAVIOR, INSTITUTIONS, AND EVOLUTION

Samuel Bowles

RUSSELL SAGE FOUNDATION
NEW YORK
PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

CHAPTER ONE

Social Interactions and Institutional Design

Two neighbors may agree to drain a meadow, which they possess in common; because 'tis easy for them to know each others mind; and each must perceive, that the immediate consequence of his failing in his part, is the abandoning of the whole project. But 'tis very difficult and indeed impossible, that a thousand persons shou'd agree in any such action; it being difficult for them to concert so complicated a design, and still more difficult for them to execute it; while each seeks a pretext to free himself of the trouble and expense, and wou'd lay the whole burden on others.

—David Hume, *A Treatise of Human Nature, Volume II* (1739)

This is how men could imperceptibly acquire some crude idea of mutual commitments and the advantages to be had in fulfilling them. . . . Were it a matter of catching a deer, everyone was quite aware that he must faithfully keep to his post in order to achieve this purpose; but if a hare happened to pass within reach of one of them, no doubt he would have pursued it without giving it a second thought, and that, having obtained his prey he cared very little about causing his companions to miss theirs.

—Jean-Jacques Rousseau, *Discourse on the Origin and Foundations of Inequality among Men* (1755)

Getting the Rules Right

Like the overnight train that left me in an empty field some distance from the settlement, the process of economic development has for the most part bypassed the two hundred or so families that make up the village of Palanpur. They have remained poor, even by Indian standards: less than a third of the adults are literate, and most have endured the loss of a child to malnutrition or to illnesses that are long forgotten in other parts of the world. But for the occasional wristwatch, bicycle, or irrigation pump, Palanpur appears to be a timeless backwater, untouched by India's cutting edge software industry and booming agricultural regions.

Seeking to understand why, I approached a sharecropper and his three

The first epigraph is from Hume (1964:304), and the second from Rousseau (1987:62).

daughters weeding a small plot.[1] The conversation eventually turned to the fact that Palanpur farmers sow their winter crops several weeks after the date at which yields would be maximized. The farmers do not doubt that earlier planting would give them larger harvests, but no one, the farmer explained, is willing to be the first to plant, as the seeds on any lone plot would be quickly eaten by birds. I asked if a large group of farmers, perhaps relatives, had ever agreed to sow earlier, all planting on the same day to minimize the losses. "If we knew how to do that," he said, looking up from his hoe at me, "we would not be poor."

Planting on the right day, like successfully draining the meadow in Hume's example or preventing the unraveling of Rousseau's stag hunt, is a solution to a problem called a *social dilemma* or *coordination problem*. Thomas Hobbes and the other founders of European political philosophy, as well as the great classical economists from Adam Smith to John Stuart Mill, sought to discover the institutions that by addressing problems like these would be most conducive to human well-being. For them an over-arching question was: how can social interactions be structured so that people are free to choose their own actions while avoiding outcomes that none would have chosen? I call this the *classical constitutional conundrum*.

We now would say: they were interested in getting the rules right. A contemporary restatement of the conundrum would define "outcomes" as equilibria of a game specified by the structure of social interactions along with an account of how, given this institutional environment, individuals might come to act in such away that a particular outcome (perhaps one of many stable equilibria) might occur and persist over long periods. "Avoiding outcomes that none would have chosen" would be refined as the pursuit of a *Pareto-efficient* outcome, namely one for which no other feasible outcome would be preferred by at least one, and not less preferred by any.

I will make extensive use of the notion of Pareto efficiency, so a comment on its shortcomings is in order. As a basis for choice among allocations, the Pareto standard is at once too weak and too strong. It is too strong because in any practical application, large numbers of people will be involved, and it is almost always the case that a change in policy or institutions inflicts costs on some participants, even in the long run. This being the case, the Pareto standard has a strong status quo bias. The Pareto standard is too weak because it abstracts from other desiderata of an allocation. The most important of these is the principle that the distribution of benefits entailed by an allocation should be fair. Thus, the idea that good rules support Pareto-efficient equilibria hardly exhausts constitutional desiderata, but, subject to these two caveats, it is certainly among them. Unfortunately, including Pareto efficiency as a desideratum does not provide much guidance in making policy choices. There may be many reasons to prefer a Pareto-inefficient outcome over a Pareto-efficient one; all that is precluded is a preference for a particular outcome when some other feasible outcome is Pareto superior to that outcome. But few practical choices take this form: most policy alternatives cannot be Pareto ranked in this way.

The constitutional conundrum has broad contemporary relevance, including environmental protection on a global scale, the determination of work effort among members of a production team, the production and distribution of information, and the formation of the neighborhoods in which people live. The fact that since the emergence of capitalism, the aggregate effect of millions of individuals, each acting independently in pursuit of their own objectives, has been a long-term improvement in the material living conditions of most of those participating suggests that tolerably good solutions can be found to problems much more challenging than the Palanpur farmers' planting date, Hume's meadow, and Rousseau's stag hunt. How it comes about that large numbers of strangers with little or no concern for one another's well-being routinely act in mutually beneficial ways is one of the great puzzles of human society, and one that I will try to illuminate. But there is also unmistakable evidence of failures to solve modern day coordination problems: systematic overuse of some resources (the natural environment) and underutilization of others (human productive capacities), for example, and the enduring poverty of the people of Palanpur and villages like it around the world.

The reason why uncoordinated activities of individuals pursuing their own ends often produce outcomes that all would seek to avoid is that each person's actions affect the well-being of others and these effects are often not included in whatever optimizing process or rule of thumb results in the decisions made by self-interested actors. These unaccounted-for effects on others are sometimes called *externalities* or *spillovers*. Economists once treated these external effects as exceptional, the standard example being the one farmer's bees transporting pollen among a neighboring farmer's apple trees. But as the above examples suggest, they are ubiquitous in a modern economy.

The classical constitutional conundrum may be posed in this manner: what rules governing interactions among people would simultaneously facilitate the pursuit of their own ends, while inducing each to take adequate account of the effects of their actions on others? The first clause ("pursuit of their own ends") simply recognizes that any solution to coordination problems will be substantially decentralized, and none

[1] Lanjouw and Stern (1998) provide a detailed account of the economy and social structure of Palanpur.

that seek to simply override individual intentions is either workable or desirable. The key challenge is in the second clause: where a person's actions unavoidably affect the well-being of others, how can these effects be made sufficiently salient to influence the actor's behavior in appropriate ways?

If the "others" are our kin, or our neighbors, or friends, our concern for their well-being or our desire to avoid social sanction might induce us to take account of the effects of our actions on them. Reflecting this fact, an important response to the constitutional conundrum—one that long predates the classical economists—is that concern for the well-being of others should extend to all of those with whom one interacts, thus internalizing the effects of one's actions on others. With the increasing scope of markets over the last half-millennium, however, individuals have come to interact not with a few dozen, but with hundreds and indirectly with millions of strangers. And so, with the maturation of capitalism and growing influence of economic reasoning, the burden of good governance shifted from the task of cultivating civic virtue to the challenge of designing institutions that work tolerably well in its absence.

Modern day *implementation theory*, the *theory of mechanism design*, and *optimal contract theory* embody this tradition, asking what forms of contracts, property rights, or other social rules might achieve some desired aggregate social objective when that objective is shared by none of the participants. A prominent example is the Fundamental Theorem of Welfare Economics, which identifies the conditions under which well-defined property rights and competitive markets support Pareto-efficient competitive equilibria. The theorem thus provides a formalization of Adam Smith's argument that under the right institutional conditions, individuals pursuing their self-interest will be "led by an invisible hand" to implement socially desirable outcomes.

The problem of draining Hume's meadow or preventing Rousseau's stag hunt from unraveling are interesting precisely because—like almost all social interactions—they are situations for which the rather stringent axioms of the Fundamental Theorem do not apply. How difficult it might be to sustain the cooperation necessary for a socially beneficial outcome in these cases depends on the underlying structure of the interaction, namely, the beliefs and preferences of the individuals, the cause-and-effect relationships governing the translation of actions into outcomes, whether the interaction is episodic or ongoing, the number of people involved, and so on. The difficulty of solving the problem also depends on the information structure of the interaction—who knows what, when, and whether the information can be used to enforce contracts or governmental regulations.

All of these influences on the likely success or failure of the drainage, the hunt, or any other common project depend on the particular institutions governing the interactions among the participants. Markets, families, governments, communities, and other institutions relevant to an interaction influence the constraints and incentives as well as the information, norms, and other evaluative concerns of the participants in the interaction. An adequate analysis of coordination problems and their possible attenuation must illuminate how these institutions work. For this task the minimal representation of institutions in the Walrasian paradigm is substantially inferior to the more elaborate modeling of institutions made possible by game theory.

My main objective in this chapter is both to introduce some basic ideas of game theory and to use these ideas to provide a taxonomy of social interactions and their outcomes. I postpone until chapter 3 an in-depth consideration of individuals and their preferences. Of course, most institutions are not designed—or at least they do not function according to any blueprint—but I will delay treatment of institutions as *evolved* rather than *designed* until chapter 2. Questions of the stability of equilibria (or why we should be concerned with equilibria at all) will also be skirted in this chapter, as they are best handled once we have an explicit model of how things change in out-of-equilibrium situations, introduced in chapter 2. I begin with an example that illustrates the formal structure of the challenges raised by Hume and Rousseau.

Coordination and Conflict: An Example

Garrett Hardin (1968) famously described a group of herders overgrazing a pasture and driving it to ruin, coining the term *tragedy of the commons* and giving social science one of the most evocative metaphors since Smith's invisible hand. Indeed, Hardin called his tragedy a "rebuttal to the invisible hand." These two metaphors are powerful because they capture two essential but sharply contrasting social situations. When guided by an invisible hand, social interactions reconcile individual choice and socially desirable outcomes. By contrast, the *dramatis personae* of the commons tragedy pursue their private objectives to disastrous consequences for themselves and others.

Hardin chose the bucolic setting for his tragedy for concreteness only; the underlying problem applies to a wide class of situations in which individuals typically cannot or do not take account of the effects of their actions on the well-being of others. These include traffic congestion, payment of taxes and other contributions to common projects, the preservation of group reputations, team work, and many more.

An example will illuminate the structure of the problem, raising a large number of issues to be addressed in greater analytical detail in subsequent chapters. Consider two fishers, Jay and Eye, who share access to a lake and catch fish there which they consume. Fish are plentiful enough so that additional fishing always yields more fish to each of the two, but the more fish one catches, the fewer the other catches in an hour of fishing. Each of them decides how much time to spend fishing, selecting the amount that maximizes their own well-being. Suppose that this optimization process, when carried out separately and without any binding agreement between the two, leads each to fish eight hours a day and that the net benefits (no pun intended) of this activity are just sufficient to match the next best alternative for each (perhaps working for wages in the nearby town). Define the benefits flowing from this so-called *fallback option* (or reservation position) as $\underline{u} > 0$ for both fishers. They each know that if they both fished less, they could each be better off, their smaller catch being more than offset by their greater leisure. Assume that they study the matter and determine how they would fare if they both limited their hours to six (we'll assume that this is the only alternative to eight hours), or if one fished eight and the other six. They normalize their payoffs so that they assign a number 1 to the outcome of both fishing less, and zero to the one who fishes less while the other continues fishing more. Table 1.1 shows the relevant payoffs (according to convention, the row player's payoffs are listed first).

The tragedy of the fishers is a *prisoners' dilemma*. This is a situation in which for each individual there is an action that, if taken, yields higher payoffs than any of the other available actions independently of what the other does (the other actions are said to be *dominated*). But when all individuals act to maximize their payoffs by taking this action the outcome is worse for both than some other outcome they could have achieved by acting differently. Thus fishing for six hours is dominated because $\alpha > 0$ and $\underline{u} > 0$, and it is Pareto superior to eight hours because $\underline{u} < 1$.

It might seem a simple matter to determine that they should just agree that each will fish six instead of eight hours, but this is far from the case, for two reasons. The first is that they may have no way of enforcing an agreement, or even knowing if the agreement has been violated. While each may know how many hours the other has fished on a clear day, on a foggy day it may be impossible to know, and in any case each one's knowledge of how much the other fished may be insufficient to enforce an agreement judicially. This is the problem of *asymmetric* or *unverifiable* information, the former describing a situation in which what someone knows another does not, and the latter that in which what someone knows cannot be used in court.

TABLE 1.1
Tragedy of the Fishers: A Prisoners' Dilemma

	Eye	
Jay	*Fish 6 hours*	*Fish 8 hours*
Fish 6 hours	1, 1	0, 1 + α
Fish 8 hours	1 + α, 0	$\underline{u}$, $\underline{u}$

The second problem arises because the six-hours-a-day arrangement is an agreement *both* to fish less *and* implicitly to divide the benefits of fishing less in a particular way, namely, equally. But the fishers of course realize that they need not agree on six hours each. They could instead agree that Eye will fish eight hours and Jay four hours, or vice versa. The fishers have two problems, not one. The first, concerning *allocation*, is to determine how much fishing to do in total, namely, how to restrict the total hours of fishing, and the second, concerning *distribution*, is how to divide the benefits to fishing less, should they agree to do so.

Figure 1.1 illustrates the fishers' opportunities and predicament. In figure 1.1, as before, six and eight hours of fishing are the only alternative actions on a given day, but now Eye and Jay may adopt strategies whereby they fish eight hours one day and six the next, as well as other combinations over a period of time. Further, I assume that any allocation must be agreed to by both fishers.

The payoffs {1, 1} are feasible and implementable by the six hour rule, but more complex agreements can implement any point within the set *abcd*. For example, point *d* can be implemented simply by Eye agreeing to fish six hours every day, and Jay's fishing eight. While Eye would surely not agree to this (Eye does worse under this arrangement than if each fishes 8 hours), Jay might offer to fish six hours a fraction of the time equal to $\underline{u} + \varepsilon$ (ε is a small positive number) and eight the rest, while requiring Eye to fish six hours all the time, threatening to fish eight hours all the time if Eye refused. Eye might well accept, for Eye would then expect a net gain of one during ($\underline{u} + \varepsilon$) of the time and $\underline{u}$ the rest, the alternative being to get $\underline{u}$ all of the time, which would occur if Jay carried out the threat. Jay would then gain net benefits of one when they jointly fished six hours, which would happen ($\underline{u} + \varepsilon$) of the time, and (1 + α) the rest of the time when Jay fished eight hours and Eye only six. Jay's proposed contract is indicated by point *f* in figure 1.1. All the points along *cfd* can be achieved by a contract of the form above: Jay works six hours for a fraction of the time, β and eight hours the rest, while Eye works six hours all the time, giving the utilities $u_i = \beta$ and $u_j = \beta + (1 - \beta)(1 + \alpha)$. Of course Eye would reject contracts along *fd*.

If Jay is able to precommit to such an offer, Jay is the *first mover* and

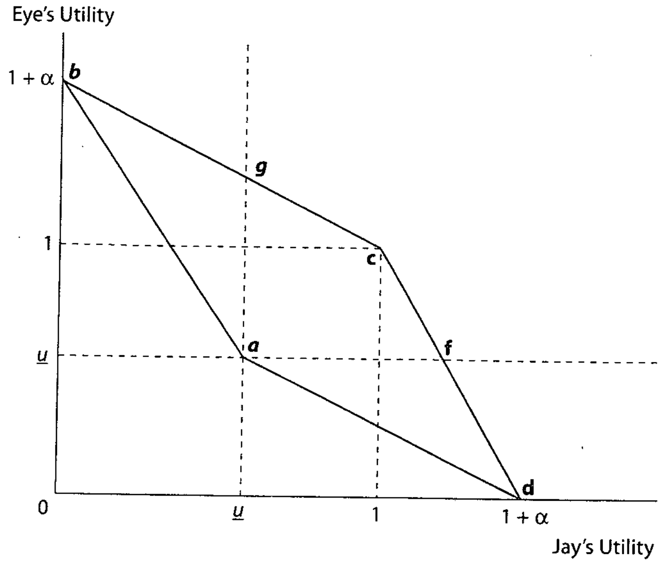

Figure 1.1 The tragedy of the fishers. Note at c both the fishers fish 6 hours while at a they both fish 8 hours.

has the *first mover advantage*. Of course, Eye might have made the identical offer to Jay. In this case the order of play (including who gets to be the first mover) makes a difference. A moment's reflection will confirm that there is not just one but rather an infinite number of agreements that are at once mutually beneficial (compared to the eight-hour rule) and efficient. An efficient agreement is one for which there exists no alternative that benefits at least one of the fishers without making the other worse off. These so called *Pareto-improving* (over the dominant strategy equilibrium outcome) and *Pareto-efficient* agreements are all the points along *fcg* in the figure (called the *Pareto frontier*.)

The fishers might have quickly agreed on the joint limitation to six hours if that were the only alternative to both fishing eight hours. But they might fail to agree once the range of possible agreements is enlarged; they may find that more options may be worse than fewer. This is because the indeterminate nature of the division of the benefits of fishing less raises the question of fairness and thus brings to bear considerations not captured in the game as described thus far. Eye, for example, might reject the disadvantageous "take-it-or-leave it" offer by Jay. But the same outcome might have been acceptable had it been arrived at in an impartial manner (by flipping a coin, for example), or had the benefits to fishing less been donated to a good cause rather than captured by Jay. If Eye and Jay cannot agree on a division, it may be that no agreement to restrict fishing is possible. But a third party, the government, might impose a seven hour limit on both fishers and then let them bargain to some more refined agreement if they are able. Or the fishers might come to adhere to an environmental norm inducing each independently to restrict his catch. The norm would imply a new payoff matrix in which the concern about environmental damage or the imposition of costs on the other fisher were taken into account.

It is just this type of indeterminacy that economic and other institutions address, answering such questions as who is positioned to make a take-it-or-leave-it offer, what other actions are available to the relevant parties, what information asymmetries or lack of verifiability bear on the problem (and, as a result, what agreements are enforceable by third parties), and what norms may affect the outcome of the conflict.

Real fishers, of course, are not acting out a tragic script, as Hardin supposed; nor are they prisoners of the dilemma they face. They are often resourceful in seeking solutions to the problem of overfishing. Turkish fishermen, for example, allocate fishing spots by lot and then rotate them. Information sharing among fishers discourages cheating, while governmental regulations supplement local social-network-based enforcement (Ostrom 1990).The extant rules regulating access to fishing are a small selection—from a much larger set of rules once tried—that have succeeded at least well enough to allow the communities using them to persist and not abandon their rules in favor of some other. As we will see, the persistence of rules does not require that they be efficient, only that they be reproduced over time. Nonetheless, we might expect a community of fishers who have hit on the ways of sustaining joint limitation to six hours to do better in competition with groups that overfish, and to be copied by other groups. We will return to the example of the fishers in chapter four to explore the analytics of how taxes, asymmetric power relations among actors, social norms, and other aspects of social interactions affect outcomes.

How might game theory illuminate the tragedy of the fishers and similar problems?

Games

Games are a way of modeling *strategic interactions*, that is, situations in which the consequences of individuals' actions depend on the actions taken by others, and this mutual interdependence is recognized by those involved. A *game* is a complete identification of the players, a list for each player of every course of action available to him (including actions contingent on the actions taken by others, or on chance events)—known as the *strategy set*—the payoffs associated with each *strategy profile* (combination of strategies), as well the order of play and who knows what, when. Players may be individuals or organizations such as firms, trade unions, political parties, or national states. In biological applications, subindividual entities such as cells or genes are also players.

Even this brief introduction reveals two great virtues of game theory as a contribution to the study of economic institutions and behavior (I will consider the drawbacks presently). First, few social interactions can be reduced to the interaction of an agent with a *given* environment (as is accomplished by the price-taking axiom and the other unrealistic assumptions of the Walrasian model). Most interactions have a strategic component, and game theory is designed to analyze the manner in which individual action is influenced by the fact that this interdependence is commonly recognized by one or more parties to an interaction. Second, the complete specification of a game requires detailed attention to the institutional environment in which the interaction takes place; outcomes often hinge on these details (for example, who takes the first move) in ways that would not be revealed in frameworks that suppress rather than highlight institutional detail. Game theory does not provide substantive insights any more than mathematics or any other language does. But it often provides a clear way of expressing insights originating elsewhere and for understanding the role of particular assumptions in a line of reasoning.

The "tragedy of the fishers" example above is a game, presented in what is called its *normal* (or strategic) *form*. This means that the time sequence of the actions taken by each player is not explicitly represented, the assumption being made that each player moves without knowing the move of the others. The *extensive form* of a game makes explicit the order of moves, and who knows what at each stage in the game. Moves made earlier in time need not be known by those making later moves, of course. An example of a game in extensive form is the representation of the experimental ultimatum game as a game tree in chapter 3. The extensive form conveys more information about the interaction in the sense that many extensive-form games may be represented by the same normal form game. When, as is common, the normal-form representation is used, this is because the additional information in the extensive form is thought to be irrelevant to how the game will be played.[2] As you will see in chapter 3, experimental subjects' behaviors appear to be quite sensitive to details that at first glance would not seem to affect the structure of the game (the name given to the game, for example, or the labeling of the players). Thus it is not a good idea to reduce an extensive form game to its normal form unless there is good reason to think that the temporal order of play will have no effect on the behaviors of the players.

The *outcome* of a game is a set of actions taken by the players (and the associated payoffs). Game outcomes cannot be deduced from game structures alone but require, in addition, a plausible *solution concept*, that is, a specification of how those involved might play. The relationship between games and their outcomes is far from settled, with sharply contrasting approaches. *Classical game theory* stresses sometimes quite demanding forward-looking cognitive evaluations by the players. By contrast, *evolutionary game theory* stresses rule-of-thumb behaviors that are updated by a backward-looking learning process, that is, in light of one's own or others' recent experience.

Two solution concepts are widely used in classical game theory: *dominance* and *Nash equilibrium*. Dominance purports to say what will *not* happen (and in some cases, by a process of elimination, is illuminating about what *will* happen). Dominance gives strong predictions of outcomes in such games as the prisoners' dilemma in which every player will choose some particular strategy irrespective of what the others do. (Games solvable by dominance are degenerate strategic interactions in that the action taken by each does not depend on the actions taken by others.) The idea behind the Nash equilibrium is that there may be one or more outcomes that no individual has any incentive to alter his strategy given the strategies adopted by all the others.

Both dominance and the Nash equilibrium are based on the notion of a *best response* strategy. A strategy may be an unconditional action (such as drive on the right), but it may also be a prescription for acting contingent on the prior actions of others or chance. "Fish six hours a day no matter what" is a strategy, as is "Fish today as many hours as the other fished yesterday" (called tit for tat). A firm's wage offer and promotion ladder contingent on worker performance is a strategy, as is an employee's choice of an effort level; a bank's interest rate, system of monitoring its clients, and method of handling their defaults is also a strategy; and so on. Thus a *strategy is a description of an action or actions to take under every situation that may be encountered in the game*. In addition to the *pure strategies* that make up the strategy set, an individual may adopt a *mixed strategy*, namely, a probability distri-

[2] Who moves first may affect behavior even if the second mover does not know what the first mover did. Examples are provided in Camerer and Weber (2003) and Rapoport (1997).

bution over some or all of the pure strategies in the set. For example, one could let a coin flip determine if one fished six or eight hours.[3]

Let there be n players indexed by $i = 1 \ldots n$, and a strategy set for each is called S_i. Suppose the j^{th} player selects a particular strategy $s \in S_j$. Let s_{-j} represent the strategies adopted by all other players (chosen from their strategy sets S_{-j}) and $\pi_j(s, s_{-j})$ the payoff to j under the strategy profile (s, s_{-j}). The payoff is j's evaluation of the outcome produced by the strategy profile (s, s_{-j}). Strategy s is j's *best response* to the strategies adopted by the others if no strategy available to j would result in higher payoffs for j. That is,

$$\pi_j(s, s_{-j}) \geq \pi_j(s', s_{-j}) \ \forall\, s' \in S_j,\ s' \neq s$$

which may be read: j's payoff to playing s against the given strategy profile of all others (s_{-j}) is not less than the payoff to playing any other strategy s' in j's strategy set against s_{-j}. A *strict best response* is a strategy for which the strict inequality holds for all s', while a *weak best response* is one for which the above expression holds as an inequality for at least one alternative strategy s'. A *weakly dominant strategy* is one for which no strategy yields a higher payoff regardless of the strategy choice of the others and that for some strategy profile yields higher payoffs. So s is weakly dominant if

$$\pi_j(s, s_{-j}) \geq \pi_j(s', s_{-j}) \ \forall\, s' \in S_j \text{ and } \forall\, s_{-j} \in S_{-j}$$

with the strict inequality holding for at least one strategy profile. A strategy is strictly dominant if no strategy weakly dominates it, that is, when the above inequality is strict in all cases. I reserve the terms "best response" and "dominance" (without the weak or strict modifier) for the stronger concept. If there exists a dominant strategy for each player, then the strategy profile in which all players adopts their dominant strategy is termed a *dominant strategy equilibrium*. Overfishing in the tragedy of the fishers is an example. Surprisingly, it may not always make sense to play a dominant strategy, but to see why, I will need to introduce another important solution concept—risk dominance—which I will do presently.

A *Nash equilibrium* is a strategy profile in which all players' strategies are best responses to the other strategies in the profile; if all of the best responses making up this strategy profile are unique (they include no weak best responses), then the Nash equilibrium is said to be strict. Because players have no reason to change their behaviors (the equilibrium is a mutual best response), it is said to be stationary, it is this characteristic that justifies calling it an equilibrium. This interpretation is based on the assumption that individuals will not *jointly* agree to alter their strategies. Responding to John Von Neumann's objection that people are not really all *that* uncooperative, John Nash (to whom we owe this and other contributions to game theory) once called it "the American way."

Finally, *iterated dominance* is a procedure by which a player may eliminate from consideration any of the *other* players' strategies that are strictly dominated (i.e., would not be advantageous to adopt in any strategy profile). Truncating the other players' strategy sets in this manner changes the structure of the game such that the game truncated by iterated dominance may have a Nash or dominant strategy equilibrium even though the complete game did not.

The Structure of Social Interactions

People interact in an endless variety of ways, but there are generic classes of interaction. Some game theoretic terminology will provide an insightful classification. The first distinction—between cooperative and noncooperative games—refers to the institutional structure governing the interaction. The second—between common interest and conflict games—refers to the extent to which the game's payoffs exhibit conflict or common interest among the players.

Cooperative and noncooperative games. Imagine an interaction for which it is the case that everything that both is affected by the actions of the players and is of concern to any of the players is subject to binding (meaning costlessly enforceable) agreement. This is termed a *cooperative* interaction (or a *cooperative game*; I use the terms game and interaction interchangeably, when appropriate). The term does not refer to the feelings of the parties about each other but simply to the institutional arrangements governing their interactions. As we will see, cooperative games may be highly conflictual: for example, the purchase of a house generally pits the interests of the buyer against the seller, but if a deal is struck, it is generally enforceable and its terms cover all of the aspects of the transfer that are of interest to the parties.

More commonly, however, something about the interaction is not subject to binding agreement. Such situations are modeled as *noncooperative games*. In some cases, part of an interaction may be addressed cooperatively, as when an employer and an employee bargain over a wage and working hours. Other aspects of the same interaction may be noncooperative because of the impossibility of writing or enforcing the

[3] While mixed strategies sometimes provide a handy modeling device (e.g., the monitoring and working example in chapter 8), for technical reasons they have been given much more attention by game theorists than is justified by any resulting illumination of human behavior.

relevant contracts. Examples include how hard the worker works or whether the employer will invest the resulting profits in this plant or elsewhere. As is the case with cooperative interactions, the parties to noncooperative interactions may have sharply conflicting interests, or share broadly common objectives; the term "noncooperative" refers simply to the fact that their interaction is not fully covered by a binding agreement. By the same token, many aspects of loving relationships among friends and family are noncooperative interactions, for example, the promise to do one's best to get a friend a job may be completely sincere, but it is not a binding agreement.

Common interest and conflict. Some interactions have the character of traffic patterns: traffic jams are a generally poor outcome, and managing to avoid them would benefit everyone. In other interactions, like settling on a price of a good to be exchanged or the division of a pie, more for one means less for the other. Many of the differences among scholars and policy makers grappling with questions of institutional design can be traced to whether they believe that the ills of society are the result of common interest problems like traffic jams or of conflicting interest problems like the division of a fixed pie. In one case, institutions may be represented as problem solvers and in the second as claim enforcers. But most institutions do both. Thus, it may be impossible to analyze the problem-solving and distributional aspects of institutions in isolation. It will be useful to have some language to differentiate between these classes of problems; to do this I will refer to the *common interest* and *conflict* aspects of an interaction, starting with pure cases.

A game in which the payoffs to only one of the strategy profiles is Pareto optimal and the payoffs associated with all strategy profiles can be Pareto ranked can be described as a *pure common interest game.*[4] What this means is that one outcome is better than all other outcomes for a least one participant and not worse for any participant, and there is a second best outcome that, while Pareto inferior to the first best outcome, is Pareto superior to all the rest, and so on. Thus, there is no outcome that any player would strictly prefer over an outcome preferred by any other player. As a result, conflict among the players is entirely absent.

Here is an example. A firm consists of an employer and an employee: If the firm succeeds, both get 1; if it fails, both get 0. The probability of success depends on actions taken (noncooperatively) by the two: the employer may invest in the firm or not, and the employee may work hard or not. If the employer invests and the worker works hard, the firm will surely succeed. In the opposite case the firm fails with certainty (table 1.2). If the employer invests and the worker does not work the firm succeeds with probability p_1, and in the opposite case the firm succeeds with probability $p_2 < p_1$. Suppose that both players choose the action that maximizes their expected payoffs, namely, the weighted sum of the payoffs occurring for each strategy chosen by the other(s), weighted by likelihood the player assigns to each of these events. It is easy to confirm that pure common interest games have a dominant strategy equilibrium, namely, the single Pareto-optimal outcome. (This is a game in which expected payoffs depend on a probabilistic outcome—the firm's success—which is influenced by the strategy profile adopted by the players. A realization of a stochastic process is sometimes referred to as *nature's move.*)

TABLE 1.2
Pure Common Interest Payoffs:
The Firm Survival Game

	Invest	*Do not*
Work	1, 1	p_2, p_2
Do not	p_1, p_1	0, 0

Note: the employer is the column player the worker is the row player: and $1 > p_1 > p_2 > 0$.

An interaction is termed a *pure conflict* game if all possible outcomes are Pareto optimal. An example is any zero sum game (meaning that for every strategy profile, the sum of the payoffs sum to zero). Pure conflict is illustrated by the set of strict Nash equilibria in the Division Game originally suggested by Schelling (1960). A dollar is to be divided between two individuals according to the following rules: without prior communication each player submits a claim of any amount, and if the claims sum to one or less the claims are met; otherwise, each gets zero. A portion of the payoff matrix for this game is as shown in table 1.3 (assuming that claims must be made in units of pennies). The off-diagonal strategy pairs are clearly not strict Nash equilibria (e.g. the lower right pair is a mutual weak best response and hence a nonstrict Nash equilibrium, as a claim of zero is also a best response to a claim of 100). The bold strategy pairs are the strict Nash equilibria of the game (there are 101 of them). Notice that each is Pareto optimal, so the outcomes

[4] The term "common interest game" has been used to refer to a payoff structure such that all players prefer a given outcome to any other (for example Aumann and Sorin 1989 and Vega-Redondo 1996); the definition here is stronger (hence the "pure") as it requires not only that a mutually preferred outcome exist but that all outcomes be Pareto rankable. Outcomes can be Pareto-ranked if the preference orderings of the outcomes—most to least preferred—of all the participants are such that if an individual prefers outcome A to outcome B, no individual prefers B to A.

TABLE 1.3
The Division Game

Claims	*0*	*1*	. . .	*99*	*100*
0	0,0	0,1		0,99	0,100
1	1,0	1,1		1,99	0,0
.					
.					
.					
99	99,0	99,1		0,0	0,0
100	100,0	0,0		0,0	0,0

making up the set of strict Nash equilibria of the Division Game describe a pure conflict interaction. The fact that all outcomes of pure conflict games are efficient in the Pareto sense does not mean that the rules defining the game are efficient; there may be other rules (that is, other ways of regulating the interaction given its underlying structure) that would yield outcomes that are Pareto superior to those defined by the pure conflict game. We will return to this.

Figure 1.2 depicts the payoffs in a generic two-person game in which each player has two strategies; hence, there are four strategy profiles and associated payoffs labeled **a** through **d**. For the pure conflict game, the payoffs are arrayed in a "northwest-to-southeast" direction (because each is a Pareto optimum, no outcome can lie to the "northeast" of any other), while in the pure common interest case they lie along a "southwest-to-northeast" axis, indicating that they can be Pareto ranked. The Firm Survival Game is an example of the class of pure common interest games in that the payoffs to the players are identical for each strategy profile (they share a "common fate") so the outcomes in figure 1.2 would be arrayed along a 45° ray from the origin. Similarly, a zero sum game is a strong form of a pure conflict game in which the payoffs would be arrayed along a line with a slope of -1.

Most social interactions are such that both common interest and conflict aspects are present. Driving on the right- or the left-hand side of the road is a matter of indifference to most people as long as others do the same. By contrast, while there are mutual gains to all people's speaking the same language, people are far from indifferent about *which* language they speak; thousands have died in wars on the subject. One of the reasons why the prisoners' dilemma has attracted so much attention is that it combines both common interest and conflict aspects.

Figure 1.1 (the tragedy of the fishers) illustrates both the conflict (northwest-to-southeast) and common interest (southwest-to-northeast) dimensions of the payoffs. A natural measure of the extent of the common interest as opposed to the conflict aspect of the payoff structure is available in *symmetric games* such as the tragedy of the fishers. (A symmetric game is one in which the payoff matrix for one player is the transpose of the payoff matrix of the other.) This measure, η, is given by the size of the improvement over the dominant strategy equilibrium made possible by cooperation $(1 - \underline{u})$, relative to the difference in payoffs when the two adopt different strategies, $1 + \alpha$:

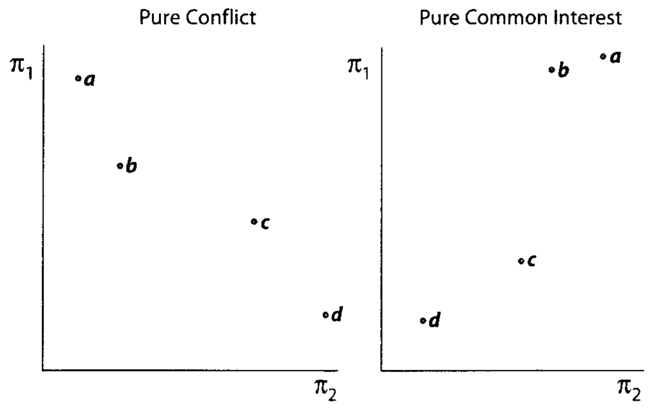

Figure 1.2 Pure conflict and pure common interest games. Note: the points **a, b, c,** and **d** indicate the payoffs to two players for each of four possible strategy profiles.

$$\eta = \frac{1-\underline{u}}{1+\alpha}.$$

For values of $\underline{u}$ and α such that the payoffs describe a prisoners' dilemma $\eta \in (0,1)$ with values approaching zero indicating virtually pure conflict, and approaching unity virtually pure common interest.

The cooperative–non-cooperative and conflict–common interest distinctions give us the typology of interactions presented in figure 1.3 with some examples for illustration. For example, the repayment of loans (analyzed in chapter 9) is a conflictual noncooperative interaction because repayment benefits the lender at a cost to the borrower, but the borrower's promise to do so is not enforceable (if the borrower has no funds). The evolution of individual property rights during the period of human history before the existence of states may have been at least initially a noncooperative common-interest interaction. By contrast, modern property rights are determined through cooperative interactions taking the form of enforceable restrictions on use and the like.

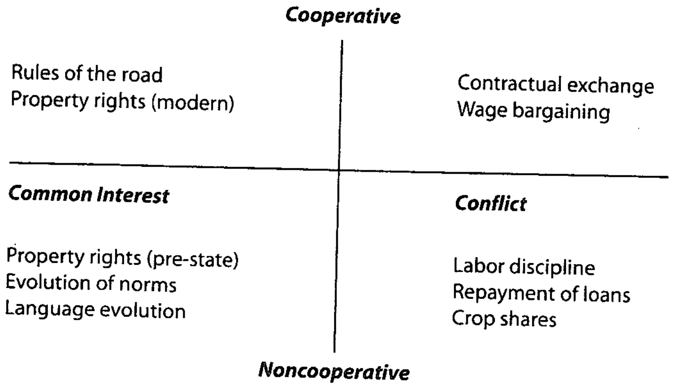

Figure 1.3 Aspects of social interactions. Note: it is not difficult to think of some property rights which should be placed on the conflict side of the graph; likewise some aspects of language evolution evolved by coercive imposition (that is, cooperatively) rather than non-cooperatively.

Another important aspect of social interactions is their temporal structure. An interaction may be repeated over many periods with the same players, possibly for a known number of periods or with a known probability of termination following each period. These are *repeated games*; nonrepeated games are often called *one-shot games*. Finally, many interactions resemble exchanges in which there is a single buyer and single seller; but in addition to these *dyadic*, or two-person, games there are many interactions involving large numbers, generically referred to as *n-person games*. Symmetric two person games with just two strategies are called *2 × 2 games*.

Coordination Failures

We now return to the constitutional conundrum, initially expressed as the challenge of ensuring that the pursuit of individual interests does not lead to "outcomes that none would have chosen." These undesirable outcomes are *coordination failures*, which are said to occur when the noncooperative interaction of two or more people leads to a result which is not Pareto optimal.[5] I refer to *coordination problems* as those situations in which coordination failures occur with significant likelihood. Familiar market failures such as those resulting from environmental externalities are a type of coordination failure, but the broader concept includes all types of noncooperative interaction, not simply those taking place in market interactions. Arms races and traffic jams are thus examples of coordination failures. An important class of coordination failures—state failures—arise when the equilibrium actions of governmental officials result in a Pareto-inferior outcome. I use the broader term *coordination failure* (rather than market failures) to draw attention to the fact that *all* institutional structures share with markets the tendency to implement Pareto-inefficient outcomes.

Table 1.4
The Invisible Hand Game

	Corn	*Tomatoes*
Corn	2, 4	4, 3
Tomatoes	5, 5	3, 2

Coordination failures may arise in out-of-equilibrium situations, but analytical attention has focused on equilibrium outcomes in which coordination failures arise in two cases. In the first, one or more Pareto-inferior outcomes may be Nash equilibria; in the second, there does not exist any Pareto-optimal outcome that *is* a Nash equilibrium. As a benchmark, consider a 2 × 2 game in which there exists a single Nash equilibrium and it is Pareto optimal, as in table 1.4. I call it the Invisible Hand Game because the self-interested actions of both actors yield an outcome that maximizes the well-being of each. (Namely, if Row grows tomatoes and Column grows corn, they each receive five, which is the best that either could do.) In this case, each not only pursues self-interested objectives but benefits from the fact that the other does as well. Row's choice of a strategy will depend on what he believes Column will do. Imagine that Rational Row notices that for Column, growing tomatoes is dominated, and therefore (using iterated-dominance reasoning) decides to grow tomatoes. But suppose that instead of pursuing his self-interest, Crazy Column flips a coin and as a result of the toss, grows tomatoes too. The example underlines that even if there is a unique Nash equilibrium, we still need to understand how players arrive at it, a topic to which we will return in chapter 2.

By contrast, in the Prisoners' Dilemma Game we have seen that a dominant strategy equilibrium exists and is Pareto inferior. A coordination failure results because the harm inflicted on the other by one's defection is not reflected in the payoffs of the defector, so neither prisoner takes adequate account of their actions' effects on the other.

[5] This is an inclusive definition of the term coordination failure, which is sometimes restricted to situations in which a Pareto-inferior equilibrium obtains when another (Pareto-superior) equilibrium exists. My definition includes cases in which no equilibrium exists.

Coordination failures arise for the same reason in the *Assurance Game*. But the game structure differs in an important way from the prisoners' dilemma: the Assurance Game payoff matrix is such that there exist more than one equilibrium, one or more of which may be Pareto inferior. (Games with this structure are sometimes called coordination games, but I will not use this term so as to avoid confusion with the terms "coordination failure" and "coordination problem" introduced above.) Thus, while a Pareto-optimal strategy profile may be the outcome of the game, it need not be. Examples include learning a language or a word processing system (its value depends on how many others have learned it), participating in a collective action such as a strike or a cartel (the expected benefits depend on the numbers participating), and the determination of employment in an entire economy (if all employers hire, the wages paid will support a level of aggregate demand justifying a high level of employment.) Other examples include the adoption of common standards (systems of weights and measures, academic credentials, computer operating systems, VHS as opposed to Betamax video technology), firms training skilled labor (if workers may move among firms, the private returns for a given firm offering training depend on the number of other firms engaging in training), and group reputations (if your trading community is known to be opportunistic, it may be a best response for you to behave opportunistically).

As these examples suggest, in Assurance Games, coordination failures occur because of generalized increasing returns or what is sometimes called *strategic complementarity*: individual payoffs are increasing in the number of people taking the same action. If I adopt the same word processing program as my colleagues, I confer benefits on them, but these benefits are not included in my decision process. (Compare this with the Invisible Hand Game above in which specialization is advantageous, so one persons' growing corn renders the other's payoff to growing corn lower.)

Because strategic complementarities may give rise to multiple equilibria, outcomes may be *path-dependent* in the sense that without knowing the recent history of a population it is impossible to say which equilibrium will obtain. In this case quite different outcomes are possible for two populations with identical preferences, technologies, and resources but with different histories. To see this, return to the farmers of Palanpur, whose crop yields would be higher if they all were to plant earlier in the year. But if a single farmer were to plant early, the seeds would be taken by the birds that would flock to his plot. Suppose there are just two farmers who interact noncooperatively for a single period with the payoffs in table 1.5. I'll assume that planting late gives a higher return if the other farmer planted early than if both plant late. The first planter gets all the predators, but if planting is simultaneous, predators are "shared" equally. While the mutual early planting equilibrium is clearly the only Pareto optimum, mutual late planting is also an equilibrium.

Table 1.5
Planting in Palanpur:
An Assurance Game

	Early	*Late*
Early	4, 4	0, 3
Late	3, 0	2, 2

The payoff matrix describes a poverty trap: identical individuals in identical settings may experience either an adequate living standard or deprivation, depending only on their histories. The planting in Palanpur problem is a special kind of assurance game in which there exist two or more *symmetrical pure strategy equilibria* (meaning that all players adopt the same pure strategy). These equilibria are called *conventions*, namely, mutual best response outcomes that are sustained by the fact that virtually all players believe that virtually all other players will best respond. (We return to the historical contingency of outcomes in chapter 2 where the analytical tools of population level dynamics are introduced.)

The games thus far introduced (plus a common children's game) allow an illustration of the sources of coordination failures listed in table 1.6. In the children's game, common around the world (English speakers call it "Rock Paper Scissors" and for others it is "Earwig Human Elephant") there is no Nash equilibrium in pure strategies.[6] Thus, no Pareto optimum is a Nash equilibrium, but because the game is zero sum (payoffs to each strategy profile sum to zero) all outcomes are Pareto optima. Because Pareto inferior outcomes cannot result, Rock Paper Scissors is not a coordination problem, even though there is no reasonable way to play the game (which is why it is fun to play).

The representation of different structures of social interaction as games has allowed a taxonomy of how coordination problems may arise. It also suggests a strategy for addressing the constitutional conun-

[6] Here is a variant of the game: on the count of three you and your partner each put forward either a flat palm (paper), a fist (rock), or two fingers in a V (scissors), with the rule that rock beats ("smashes") scissors, scissors beats ("cuts") paper, and paper beats ("covers") rock, the winner and loser gaining and losing a point each respectively. (A tie produces no score, but can result in mutual hilarity occasioned by rock fights, scissor wars, and paper coverups.) How the earwig beats the human is still a mystery to me; but then try explaining why paper beats rock. See Sato, Akiyama, and Farmer (2002).

TABLE 1.6
Sources of Coordination Failures

	P-inferior Nash exists	*No P-inferior Nash*
No P-optimum is Nash	Prisoners' dilemma	
A P-optimum is Nash	Assurance Game	Invisible hand

drum: if the likely outcome of the an interaction is Pareto inferior to some other feasible outcome, introduce policies or property rights that will change the game structure to make the second outcome more likely. An example follows.

The key difference between prisoners' dilemmas and Assurance Games is that in the former the undesirable outcome is the only Nash equilibrium, so the only way that any of the other outcomes can be supported is by a permanent intervention to change the payoffs or the rules of the game. In the assurance game, by contrast, a desirable outcome (mutual early planting, for example) is an equilibrium, so the challenge to governance is limited to the less challenging *how to get there* problem rather than also having to solve the more demanding *how to stay there* problem. In debates on the appropriate type (and duration) of government interventions in the economy, key differences among economists and others concern whether one believes that the underlying problem resembles a Prisoners' Dilemma Game or an Assurance Game. Interventions may be called for in both cases, but Assurance Game problems may sometimes be reasonably well addressed by one-time rather than permanent interventions. It is partly for this reason that a common approach to averting coordination failures is to devise policies or constitutions that transform the payoff matrix so as to convert a prisoners' dilemma into an Assurance Game by making the mutual cooperate outcome a Nash equilibrium. An interaction that is a prisoners' dilemma if played as a one-shot game, may be an Assurance Game with mutual cooperate a Nash equilibrium if played as a repeated game, as we will see in chapter 7.

But while a Pareto-optimal Nash equilibrium exists in an Assurance Game, that fact alone is not sufficient to guarantee a mutually beneficial solution; unsolved coordination failures arising from Assurance Game–like interactions are ubiquitous. An important reason is that one's decision about how to play depends on one's beliefs about how others will play, and the way that people cope with this indeterminancy may result in sub-optimal outcomes. The problem is illustrated in figure 1.4, in which the expected payoffs of planting late and early (π_l and π_e, respectively) are just linear functions of the payoffs in the Planting in Palanpur matrix above. Suppose you are the Row farmer in Palanpur and have no information on the likely play of the Column farmer, attributing equal likelihood to Column's two strategies. You will choose late planting because your expected payoffs are then 2½ (that is, ½(3) + ½(2)), while the expected payoff to early planting is 2. Even if the mutual early planting equilibrium were somehow to be attained, if you thought that the other might switch strategies by whim or by mistake, it might be difficult to sustain the early planting convention. To see why, imagine that the zeros in the figure were instead −100, namely, the payoff associated with the destruction of one's crop and as a result being without food.

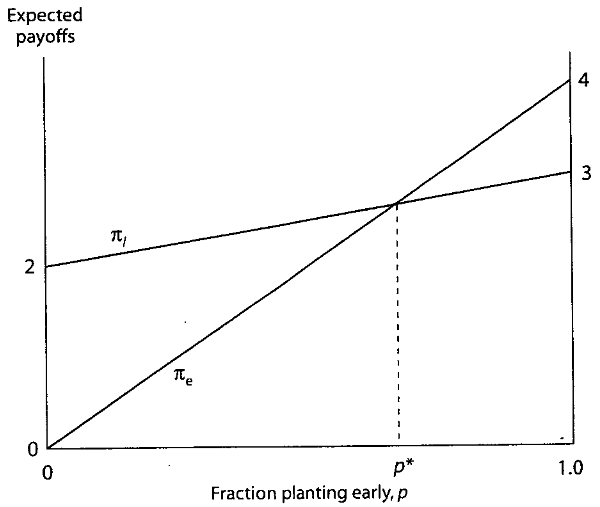

Figure 1.4 Planting late is risk dominant. Note: $p^* = ⅔$ so $\pi_l > \pi_e$ for $p =$ ½. The intercepts of the vertical axes are the payoffs in the payoff matrix on p.

As the underlying idea here will recur in the pages that follow, a few definitions (restricted to 2×2 games) will help. Call a convention in which both play strategy k, a k-equilibrium. The other is strategy k'. Define the *risk factor* of a k-equilibrium as the smallest probability p such that if one player believes that the other player is going to play k with probability greater than p (and k' with a probability less than $(1 - p)$) then k is the strict best response for the individual to make. The equilibrium with the lowest risk factor is the *risk dominant equilibrium*.

In the example above, the risk factor of the late planting equilibrium is ⅓, which is less than the risk factor of the early planting equilibrium (⅔). Late planting is termed Row's *risk dominant strategy*, namely, the strategy that maximizes the expected payoffs of a player who attributes equal probabilities to the strategies open to the other player. Because this is true for the column player as well, mutual late planting is the *risk dominant equilibrium*. Figure 1.4 illustrates these concepts. The fraction planting early is p, while π_l and π_e are the expected payoffs to planting late and early, respectively, conditional on one's belief about p. The early planting equilibrium is termed the *payoff dominant equilibrium*: An equilibrium is payoff dominant if it there is no other equilibrium which strictly Pareto dominates it. In our example, early planting is payoff dominant because the payoffs in this equilibrium exceed the payoffs for both players in the late planting equilibrium.

Notice that the farmers are assumed to maximize expected payoffs, which implies that they are risk neutral, so the fact that the risk dominant but Pareto-inferior equilibrium may obtain does not presume risk aversion on the part of the farmers. (Risk neutrality and risk aversion are discussed in chapters 3 and 9.) Note also that the coordination failure does not arise in this case due to a conflict of interest between the farmers, as it did in the prisoners' dilemma faced by the fishers. Each of the fishers prefers that he fish more and the other fish less. But both farmers prefer mutual early planting over any other outcome. Their failure to coordinate on the mutually desired outcome is the result of uncertainty about the actions to be taken by the others and not due to a conflict of interest. The prediction that the risk-dominant equilibrium will be favored over the payoff-dominant equilibrium is strongly supported by the actual play of experimental subjects in games capturing the logic of the planting in Palanpur problem (Van Huyck, Battalio, and Beil 1990). We will see (in chapter 12) that risk dominant equilibria may persist over long periods even when a payoff dominant equilibrium exists.

Thus even if a policy intervention succeeded in converting a Prisoners' Dilemma Game to an Assurance Game, the desired Pareto-optimal outcome may not result. A more ambitious objective is to convert the underlying social interaction from a prisoners' dilemma to an Invisible Hand Game. To see how this might work, consider a generic prisoners' dilemma with the payoffs a, b, c, and d in table 1.7. (Ignore the payoffs in bold type for the moment.) The interaction is a prisoners' dilemma if $a > \mathrm{b} > c > d$ and $a + d < 2b$, the second requirement expressing the fact that the expected payoff of both Row and Column is greater if they cooperate than if one were to defect and the other cooperate, with the assignment of the two roles being decided by chance. Suppose Row and Column decided to embrace "cooperate" as the norm and to adopt a liability rule according to which anyone violating the norm must compensate those whose payoffs are reduced as a result of the violation, with compensation sufficient to exactly offset the losses (we will postpone the important question of the enforcement of the new property rights). Thus if Row defects on Column, Row initially gets a as before but then must compensate Column for the costs his defection has inflicted, that is, compensation sufficient to give Column a payoff of b (which would have occurred had the norm not been violated). If they both defect, they both gain c but then must compensate the other by a transfer of $b - c$. The transformed payoff matrix for Row's payoffs is thus given by the bold entries in the figure below.

Table 1.7
Implementing a Desired Outcome by Transforming Property Rights

	Column	
Row	*Cooperate*	*Defect*
Cooperate	b, b $\boldsymbol{b, b}$	d, a $\boldsymbol{d + (b - d), a - (b - d)}$
Defect	a, d $\boldsymbol{a - (b - d), d + (b - d)}$	c, c $\boldsymbol{c, c}$

Transformed payoffs are in bold.

Did the improved property rights succeed? Because $\boldsymbol{a - b + d < b}$ by the definition of a prisoners' dilemma, cooperate is a best response to cooperate and mutual cooperation is a Nash equilibrium. Cooperate is also a best response to defect (because $b > c$), so cooperate is the dominant strategy and mutual cooperation is the dominant strategy equilibrium. Thus a redefinition of property rights (to take account of liability for damages) implements a social optimum by inducing each to take account of the effect of his actions on the other. The property rights redefinition transformed the game from a mixed conflict and common interest game to a pure common interest game. However, as we will see in subsequent chapters, most coordination failures do not allow such simple solutions. The reason is that the identification of the defection and the assessment of the relevant damages requires information that either is not available to the relevant parties or is not useable in a court of law or any other feasible body charged with enforcement of the relevant rights.

Games and Institutions

Do games illuminate institutions? *Institutions* (as I use the term) are *the laws, informal rules, and conventions that give a durable structure to* *social interactions among the members of a population.* Conformity to the behaviors prescribed by institutions may be secured by a combination of centrally deployed coercion (laws), social sanction (informal rules), and mutual expectations (conventions) that make conformity a best response for virtually all members of the relevant group. Institutions influence who meets whom, to do what tasks, with what possible courses of action, and with what consequences of actions jointly taken. It is clear from this definition that an institution may be formally represented as a game. The labor market institutions explored in chapters 8 and 10 are modeled in this way: the relevant institutions define what the employer may do (vary the wage as first mover, terminate the job) and may not do (physically punish the employee), and similarly for the worker (vary the level of work effort) with the payoffs to the two depending on the strategy profile. These labor market and firm-level institutions are modeled as games. Institutional innovations such as minimum wages or regulations governing terminations may be considered as ways of altering the strategy sets, payoffs, information structure, and players such that the equilibrium of the game may be displaced.

But to understand why institutions might change, it will sometimes be insightful to represent an institution not as a game but rather as the equilibrium of an underlying game. Because institutions are persistent rather than ephemeral it is natural to represent them as stable equilibria of an underlying game in which the strategy set encompasses a wide range of possible actions (whip the shirking worker, refuse to hand over the goods produced to the employer) that are not observed in the institutional set up described above but could be part of some other equilibrium strategy profile. Thus, to continue the employer-employee example, the expectation that the employer and not the employee will have possession of the goods produced is a mutual best response, that is, an *outcome* of some game (or, more likely, games), presumably one in which the players include not only employers and employees, but also police and judicial officials and many others. When a particular set of mutual best responses is virtually universal in a population over an extended period of time, it constitutes one or more institutions.

In chapters 2 and 11 through 13, I will model property rights, crop shares, rules governing resource sharing, and the like as equilibria, and study the manner in which these equilibria may evolve in response to chance events, collective action by those affected, and exogenously induced changes in the structure of the relevant underlying games. In chapter 2, I model the process of racial segregation of a residential neighborhood to illustrate how an institution (segregated residences) can be understood as the equilibrium of a game.

There is no inconsistency and little risk of confusion in representing institutions both as games and also as equilibria of an underlying game. Which is appropriate will depend on the analytical problem at hand. If we are interested in understanding why the poor are credit constrained (chapter 9), modeling the lender-borrower relationship as a game will be adequate (and asking about the origins of limited liability and the other underlying property rights is a distraction). On the other hand, if we want to know why limited liability exists, we would model this aspect of property rights as the outcome of an underlying game. Similarly, if we wanted to know why primogeniture is less common in Africa than in Asia, we would need to model rules of inheritance as conventions, that is, as equilibria of Assurance Games.

Table 1.8
Rousseau's Stag Hunt

	Hunt stag	*Hunt hare*
Hunt stag	½ Stag	0
Hunt hare	1 hare	1 hare

Note: the entries are Row's outcomes; payoffs can be calculated using the fact that one-third of a stag is worth one hare.

The term "institution" is sometimes also used to refer to such individual entities as a particular firm, a trade union, or a central bank; but to avoid confusion I will call these entities *organizations*. One may also treat organizations as if they were individual players in a game; this may be insightful as long as one has reason to think that the entity does indeed act as a unit; treating the firm as a single person may make more sense than applying the same logic to "the working class."

Rousseau's stag hunt illustrates the relationship between games and institutions. Suppose you observe a group of hunters, who hunt for hare, though there are stag in the forests around them. You wonder why they do not hunt stag, and consult the Stag Hunt Game (table 1.8) seeking an explanation. Assume there are two hunters, who decide, independently and without knowledge of the other's choices, either to hunt for stag (capturing one and consuming it equally if they both hunt stag, and otherwise capturing—and hence consuming—nothing) or to hunt for hare (bagging one hare and consuming it, independently of what the other does). For the moment, we assume that the hunters do not expect to meet again. Finally, each hunter values a third of a stag as much as one hare. The hunting technology (not the payoffs) is summarized in table 1.8. The game captures important aspects of the relevant institutions, for example, that they do not decide jointly what to hunt (or to be more precise, they have no means of binding themselves to abide by any decision they might make), that if both participate in the stag hunt, the kill will be shared equally, and that even if one hunts hare, rendering the other's stag hunting fruitless, one may consume the hare without sharing. This exemplifies using a game to describe an institution, along with the relevant technologically given cause and effect relationships.

By itself, however, the game is not very illuminating. Given the payoffs, both mutual hare hunting and a joint stag hunt are conventions (it is an Assurance Game), so without knowing anything about the beliefs of the hunters about the likely actions of the other we would not be able to predict whether the hare or the stag would be in jeopardy. Imagine, now, that the interaction is ongoing, and that in the previous period both hunted hare (for whatever reason); one of the hunters considers hunting stag this period instead. For this to be in the interest of the hunter (considering only this period's payoffs), she would have to expect that the other would do the same, attaching a likelihood of at least two-thirds to this occurrence. In making this assessment she would need to know something of the history of this group of hunters, and in particular, past outcomes of the game, possibly including complex outcomes such as joint stag hunting on weekends or solitary hare hunting on weekdays. If the undecided hunter has no such clues to go on and therefore attaches equal likelihood to the other hunter's two actions, she will hunt hare, for it is transparent that while mutual stag hunting is the payoff dominant equilibrium, hare hunting is risk dominant. Thus mutual expectations (whether arising from historical experience or from any other source) are as much a part of explaining why it is hare rather than stag that they hunt as is the assumption that they have no way of subjecting one another to binding agreements.

Notice, also, that some aspects of the game taken as exogenously given in the above account may be explained as the result of other institutions, that is, as the equilibria of underlying games. The practice of allowing the hare hunter to consume his catch even if the other has nothing, or dividing the stag equally may (as we will see) be modeled as outcomes of an underlying game in which these particular property rights are an equilibrium and in which other property rights (share the hare, for example, or, the stag goes to the one whose arrow felled it) could have obtained.

While game theory illuminates many important aspects of institutions and economic behavior, there are serious gaps in our current knowledge. First, while much of the use of game theory in the social sciences concern 2 × 2 games of the type introduced here, the relevant numbers involved in many social interactions are much greater and the strategy sets far more complicated. The analysis of n-person games or games

with large strategy sets lacks the simplicity, tractability, and transparency of the above games. The 2 × 2 games introduced thus far are best considered metaphors for much more complex problems, often pointing to important aspects of interactions but falling far short of an adequate analysis. Steps towards realism need not come at too high price in tractability, however. Two-person interactions are often embedded in interactions of much larger populations, as in the population level analysis of the Hawk Dove Game presented in chapter 2, the exchange games in chapter 7, and the conventions studied in chapters 11 through 13. And it is often possible to model a complex set of interactions as a series of separable two-person or larger interactions. When we turn to the analysis of the firm, for example, it will be analyzed using a two-person interaction between employer and employee, a separate two-person interaction between the firm and a lending institution, and a large-n interaction on competitive goods markets.

But many of the decentralized solutions to coordination problems based on such things as game repetition and reputation (presented in chapter 7) have far wider applicability to two-person (or very small n) interactions than to the large-n interactions that characterize many of the coordination problems of interest. The exaggerated emphasis on two-person games (due in part to their pedagogical value) that are amenable to solution in a repeated game framework may have contributed to the view that coordination failures are exceptional rather than generic aspects of social interactions.[7]

The fact that game theory has made less progress with noncooperative n-person interactions than with either cooperative or two-person games is hardly a criticism of the approach, for it arises because game theory addresses intrinsically complex aspects of human interaction that are abstracted from in other approaches. What makes the analysis of interactions among many individuals intractable is the assumption that they act strategically rather than taking the others' actions as given. Where one can abstract from strategic action—as in competitive markets for goods governed by complete contracts and in which only equilibrium trades take place, namely, the paradigmatic Walrasian case—much of the analysis is reduced to a single individual interacting with a

[7] Pedagogy, not realism, must also explain why so much attention has been given to symmetric games. The games that real people play are *asymmetric* in the sense that players often come with (or acquire) labels that assign to them different strategy sets and payoffs: men and women, insiders and outsiders, employers and employees, typically interact asymmetrically. Asymmetrical games are common in game theoretic models of labor markets, credit markets, and other situations in which institutions allocate individuals to distinct structural positions (borrower, lender) with different strategy sets. These models appear in chapter 2 and in chapters 5 through 10.

given set of prices, technological blueprints, and constraints. But, as we will see, there are many important interactions—labor markets, credit markets, markets for information and for goods of variable quality—for which this particular way to achieve tractability is not insightful.

Second, the main solution concepts of classical game theory—dominance (direct, iterated, and risk) and Nash equilibrium—are intended to supply the standard of reasonable ways that the game would be played. But they are not entirely adequate as a guide to what will happen. Other than the prisoners' dilemma, few games have dominant strategy (or iterated dominance) equilibria, and many (pure strategy) games do not even have Nash equilibria. Iterated dominance may not be robust as a solution concept because it is a reasonable way to play only if the other players have the same understanding of the game and its payoffs, are using the same solution concept, and are not prone to make errors (the common knowledge and common rationality assumptions.)

The Nash concept is more robust: if we are concerned with the explanation of durable (as opposed to ephemeral) phenomena, it is natural to look at outcomes for which it is true that no one with the ability to alter the outcome through his actions alone has an interest in doing so. Thus, we can say that a Nash equilibrium is an outcome at which there are no endogenous sources of change (this is an adequate definition of any *equilibrium*). By confining our attention to stable Nash equilibria the concept is made considerably more useful. But as a guide to outcomes, even under the assumptions of common rationality and common knowledge, the stable Nash equilibrium is incomplete in two ways. First, we need to know how reasonable play would lead to a Nash equilibrium and why it might be stable. This requires attention to what the players do in out-of-equilibrium situations. In some cases, there is little reason to think that reasonable play would lead to the Nash equilibrium. If you doubt this, try to explain why one would expect players in the Rock Paper Scissors Game to play the mixed-strategy Nash equilibrium for that game (that is, play each with probability ⅓, the only Nash equilibrium). Second, many games have many Nash equilibria, so the Nash concept alone cannot predict outcomes; information about initial conditions plus an analysis of out-of-equilibrium behavior are required to understand which of many Nash equilibria will obtain. Thus, historical contingency and dynamics (including learning) are necessary complements to the Nash concept.

The problem of indeterminancy arising from the multiplicity of equilibria has been addressed in different ways by classical game theory and evolutionary game theory. Classical game theory has sought to narrow the set of possible outcomes through restrictions on the behaviors of the players based on ever stronger notions of rationality. These additional

restrictions, called *refinements*, preclude equilibria involving strategies which include *noncredible threats* (i.e., those that would not be best responses ex post should they fail to be effective), or are not robust to small deviations from best-response play ("trembles") or payoffs, or that are supported by beliefs that fail to make appropriate use of all the available information (e.g., that do not make use of backward induction or iterated dominance).

By contrast, evolutionary and behavioral game theory addresses the above limitations by relaxing the common knowledge and common rationality assumptions and by using empirically (mostly experimentally) grounded assumptions about how real people interact. Evolutionary game theory, for example, typically assumes that individuals have limited information about the consequences of their actions, and that they update their beliefs by trial-and-error methods using local knowledge based on their own and others recent past experience. In contrast to the highly intelligent and forward-looking players in classical game theory, the subjects of evolutionary game theory are "intellectually challenged" and backward looking. Because there is little evidence that individuals are capable of (or predisposed to) conducting the quite demanding cognitive operations routinely assumed by classical game theory, I will proceed (in chapters 2 and 3) to develop a set of assumptions more in line with empirical knowledge. A second reason for rejecting the classical approach is that it is a mistake to think that indeterminacy among equilibria can be settled by game theory itself, without reference to the particular history of the players. Embracing rather than seeking to skirt the fact that social outcomes will be influenced by the recent past—that history matters—attests to a necessary insufficiency of theory, not its weakness.

A third concern about game theory as the foundation of the analysis of economic institutions and behavior is its narrow scope. Society is not well-modeled as a single game, or one with an unchanging structure. An approach to games that would be adequate to understanding society would have to take account of the following characteristics. Games are *overlapping*: people regularly participate in many distinct types of social interaction ranging from firms, to markets, to families, to citizen-state relationships, neighborhood associations, sports teams, and so on. Credit markets are often linked to labor and land markets, for example, and loan agreements that would be infeasible in a credit market taken in isolation may be possible when the borrower is also the employee of the lender, or the renter of his land, and in both cases subject to eviction should default occur. The overlapping character of games is also important because the structure of one game teaches the players lessons and imparts direction to cultural evolution, affecting not only how they play the game in subsequent periods but how they play the other games they are involved in. Citizens endowed with well-defined individual liberties and democratic rights in their relationship to their government may, for example, seek to invoke these in the workplace. Games, in other words, are *constitutive* of the players' preferences. Furthermore, not only the players evolve; the rules do as well. The games are thus *recursive* in the sense that among the outcomes of some games are changes in the rules of this or other games. In the pages that follow, I will introduce *overlapping* and *asymmetric* games in the analysis of firms, credit markets, employment relationships, and class structure. *Constitutive* and *recursive* games will be used to analyze the coevolution of preferences and institutions.

Conclusion

Why, then, do the farmers of Palanpur remain poor, planting late and bearing the costs of the other coordination failures that appear to limit their economic opportunities? Why do meadows go undrained and stags roam the forest unmolested? The long term persistence of Pareto-inferior outcomes is a puzzle of immense intellectual challenge and practical importance.

A number of possible impediments to solving coordination problems have been mentioned thus far (I will return to them in subsequent chapters). Coordination failures that are readily avoided among two individuals may pose insurmountable obstacles if a hundred or a thousand individuals are interacting, as Hume pointed out in his comment on the difficulty of securing the drainage of the meadow. The underlying interaction may be such that the dominant strategy is noncooperation (as in the prisoners' dilemma). Because of nonverifiable information or for other reasons, there may be no way of transforming the relevant game to remove this obstacle. The changes in the rules of the game necessary to avert a particular coordination failure may be resisted due to the open endedness of institutions and the losses some players might as result fear due to the effect of institutional changes on some *other* game. Even if a payoff dominant equilibrium exists, it may not obtain because some other equilibrium is risk dominant and there is no way of coordinating expectations. If, as is often the case, an acceptable division of the gains from cooperation cannot be assured, those involved may prefer noncooperation to cooperation. Finally, where the degree of common interest is small (as opposed to conflict), the gains to mutual cooperation may be insufficient to justify the risk or the cost of securing conditions to implement cooperation.

It was once widely thought that governmental intervention could readily attenuate the most serious coordination failures. But few would now share Hume's optimism, expressed in the sentence immediately following the passage quoted in the epigraph: "Political society [meaning a government] easily remedies . . . these inconveniences" (Hume 1967: 304). "There are persons," Hume wrote, "whom we call . . . our governors and rulers, who have no interest in any act of injustice . . . and have an immediate interest . . . in the upholding of society" (pp. 302–3). Among the reasons for our modern skepticism that "political society easily remedies these inconveniences" is the realization that institutions and policies are not simply instruments ready to be deployed by Hume's well-meaning public servants. Rather, they are the products of evolution as well as design and are themselves subject to the same kinds of coordination failures introduced above.

So far I have identified a number of Pareto-inferior outcomes as Nash equilibria. Understanding the underlying coordination failures, the impediments to their solution, and how they might be overcome requires an understanding of why individuals take the actions that implement and sustain inefficient Nash equilibria over long periods. To answer these questions we need to understand how both individual behaviors and social institutions evolve over time. In chapter 2 we introduce the tools of evolutionary modeling to address these issues.