4 Game Theory

- **Overview**
- **Whither the Euro?**
- **Some Games and Some Formalism**
- **Game Theory's Payoffs**
- **Game Theory and Rationality**
- **Game Theory as Explanatory Theory**
- **Study Questions**
- **Suggested Readings**

Overview

In this chapter I will look at some pressing philosophical issues concerning game theory in both its descriptive as well as its prescriptive modes. After some motivation and preliminaries, I will introduce the game-theoretic formalism along with famous games that have proved to be of philosophical interest. I will then look at some arguments in the debate whether game theory is an empirically more or less empty formalism or a substantive theory and finally whether it is a normative theory of rational choice.

Whither the Euro?

When this chapter was written (December 2011), the euro still existed as the official currency of the eurozone of 17 of the 27 member states of the EU. Here is a prediction: when the chapter is being read, the euro will no longer exist in that form. It might continue to exist as a common currency for a core EU, but if so, the financial architecture behind it will be quite different.

One of the most important issues dividing the EU today is whether to introduce so-called European bonds or "Stability bonds," sovereign debt backed collectively by all eurozone member states, in order to save the common currency. So far sovereign debt within the eurozone has been

denominated in euros but *guaranteed* only by the issuing country. Is it a good idea to change that and to introduce a system of joint liability?

Any economic argument makes idealizing assumptions (see Chapter 7), so let us begin by caricaturing the actual situation. There are two blocs of eurozone states, the "North" and the "South." Within each bloc, national policies are completely homogeneous so that we can treat each bloc as a single economic agent. Both North and South have a menu of fiscal policy options. Simplifying—caricaturing—again, let us assume that there are only two available fiscal policies: being "frugal" or being "profligate." A "frugal" policy is one that encourages saving, that reduces the fiscal deficit or even pays back government debt. A "profligate" policy is one that encourages consumption, that does not mind running budget deficits and that will pile up government debt. If both blocs are profligate, total eurozone government debts will soon run out of control, interest rates will surge and, on occasion, interventions by the European Central Bank (ECB) in the form of direct purchases of government bonds will enable continued spending. These interventions will fuel inflation, which will rise to 10 or more percent per year. The currency will be devaluated relative to the US dollar and the yuan. Let us call this scenario "currency erosion." Neither bloc particularly likes it.

What about other scenarios? Let us suppose that both blocs are strong enough to shoulder the other's debt as long as at least one of them is frugal. That is, if either North or South but not both are frugal, sovereign debt will rise at best modestly, inflation will be contained and the euro remain strong relative to other currencies. Whichever bloc is profligate will of course consume much more than the frugal one; and employees in the frugal bloc will work, to some extent, in order to subsidize the other bloc's consumption.

What policies do North and South prefer? Let us begin with the stereotype. The North with its Protestant and principled ethics prefers to be frugal no matter what—for frugal is what one ought to be. The preferences of the Catholic and utilitarian South are more complicated. It likes best the North to finance its lavish consumption: i.e., itself to be profligate when the North is frugal. With joint liability it is free to act in this way—the North guarantees its government debt, after all. We can assign this situation an arbitrary number, let's say 5. The South likes least the counterfactual situation in which the South works in order to finance lavish consumption in the North—this is unheard of. Assign this situation the number 0. Being frugal when the North is frugal is intermediate, perhaps 3. Currency erosion is liked much less but still preferred to being a sucker to the North, so let's give it a 1.

In the beginning, the North will be frugal and the South profligate. But the North, though principled, is certainly not pea-brained and will soon learn that it could profit from the Eurobond system if it started consuming and had others pay for it. They of course will not like the currency to erode, but will like even less continuing to be a sucker to the South. Being frugal alongside the other is still an option they like, though not quite as much as

having others paying for an opulent lifestyle. In other words, the North's preferences are now a mirror image of those of the South.

If we now call the numbers (which are arbitrary as long as they preserve the ordering of policy options) "payoffs," North and South "players," the policy options "strategies" and arrange the whole shenanigans in a matrix such that player 1's strategies appear in the rows and player 2's strategies in the columns, we have what economists call a "game" (see Figure 4.1).

How do we make a prediction about what players will do in a game? Focus on the North's actions first. Suppose—counterfactually—that the South plays "frugal." If the North is frugal itself, it will end up with a payoff of 3. If, by contrast, it plays "profligate," it will end up with 5. So it will play "profligate." This is precisely what the numbers mean: a higher payoff just means that that strategy will be played given the chance, no mistakes are being made etc. If the South plays "profligate" the North will end up with 0 when it plays "frugal" but 1 when it plays "profligate." In that situation, too, it will play "profligate." That is, no matter what the South does, the North will play "profligate"—despite its initial Kantian ethics. Since the South's preferences are a mirror image of those of the North, the exact same reasoning applies to the South, so they will end up both playing "profligate," the currency will erode and the union dissolve. The bottom line is: perhaps it's not such a good idea to have countries with different interests guarantee one another's debt.

Game theory is the branch of economics which concerns itself with strategic situations such as this. The game I mockingly called the "Eurobonds game" is in fact a version of the Prisoner's Dilemma, one of the most important and most widely discussed games analyzed by the theory. To some, Game theory has been the main driver of theoretical progress in economics in the last 30 or so years (Kincaid and Ross 2009). To others, it is a mere "branch of applied mathematics" (Rosenberg 1992: 248). It is thus controversial whether game theory is a substantive empirical theory that allows making predictions and explains phenomena of interest. What appears to be less controversial is the understanding of game theory as the theory of rational choice in situations of strategic interaction (this is certainly how the founders of game theory saw it; see von Neumann and Morgenstern 1944).

		South	
		Frugal	Profligate
North	Frugal	(3, 3)	(0, 5)
	Profligate	(5, 0)	(1, 1)

Figure 4.1 The Eurobonds Game

But appearances deceive. While most philosophers and economists agree that there *is* a prescriptive reading of the theory, it is by no means clear what this reading amounts to. We can illustrate this in a simple and intuitive way using our Eurobonds Game. If the two players play the strategies they most prefer, they end up in a situation they both dislike—currency erosion. Given the players know this, to many observers playing "profligate" does not seem so rational after all, especially when there seems to be a Pareto-superior set of strategies available ("frugal," "frugal"). Economists tend to respond that the Pareto-superior set of strategies is only seemingly available. This is because if one of the players did play that strategy, the other would have a massive incentive *not* to do so himself, which is predicted by the first player, who will therefore also stick with playing "profligate." But who is to say that *this* reasoning is compelling?

The set of strategies ("profligate," "profligate") is called a "Nash equilibrium." As we will see below, the justification for why rational agents ought to play Nash-equilibrium strategies is very thin. Moreover, the Prisoner's Dilemma is a very unusual game in that it has a unique Nash equilibrium. Most games have many such equilibria, and there are no good principles to guide players in selecting a specific Nash equilibrium. It is therefore often not quite clear what "the" rational course of action in a given strategic situation amounts to.

Some Games and Some Formalism

The folk of economics are *economic agents*. Economic agents have *preferences*, which means that they can order the options available to them. When the value of the different options an agent faces depends on what other agents do (that is, when they are in strategic situations), the agents are said to play a *game*, and they are referred to as *players*. Players in games select among *strategies*. A strategy is a sequence of actions that tells a player what to do in response to other players' possible strategies. There are two main forms to represent a game: strategic (or normal) and extensive. The *strategic form* is a matrix, as in Figure 4.1. The *extensive form* resembles a tree, as in Figure 4.2, where the same game is represented in this alternative mode. When I was first taught game theory, our teacher referred to extensive forms as Baselitz-style trees, for obvious reasons.

In strategic form, the players' strategies are arranged in the *rows* and *columns* of the matrix, and there is a convention to list player 1's strategies in the rows, and player 2's strategies in the columns. (Games with more than two players exist of course, but we will not consider them in this chapter.) In extensive form, players' decision points are represented by *nodes*, and their strategies by *branches* (or arrows) emanating from the nodes. In either form, the *outcome* of the game is a set of *payoffs* to the players, and payoffs are ordinal utilities assigned to players (or cardinal utilities when mixed strategies are allowed; mixed strategies will be introduced below).

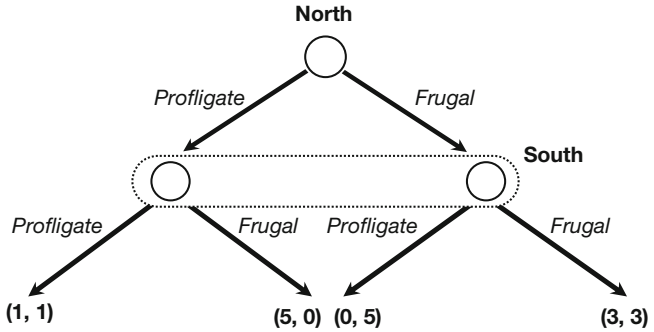

Figure 4.2 The Eurobonds Game in Extensive Form

To solve a game, one has to find its equilibrium or set of equilibria. The most important equilibrium concept is that of the *Nash equilibrium*. A Nash equilibrium is a set of strategies such that no player can improve her payoff by changing her strategy, given the strategies of the other players. One way to find a Nash equilibrium in a strategic-form game is by eliminating *strictly dominated strategies*. A strategy is said to be strictly dominated whenever it is inferior to all other strategies regardless of what the other player does. Thus, in the matrix of Figure 4.1 playing "frugal" is a strictly dominated strategy for player 1, North, because playing "profligate" is superior independently of whether player 2, South, plays "frugal" or "profligate." We can therefore eliminate the top row. The same reasoning applies to player 2, so here we can eliminate the left column. There remains a single set of strategies ("profligate," "profligate"), which constitutes the unique Nash equilibrium of the game.

But as mentioned above, the Prisoner's Dilemma is an unusual game. Consider a second game philosophers enjoy discussing, the Stag-Hunt Game (Figure 4.3).

The story associated with it stems from Jean-Jacques Rousseau's *Discourse on Inequality* in which he describes the advantages of social cooperation:

		Hunter 2	
		Stag	Hare
Hunter 1	Stag	(3, 3)	(0, 2)
	Hare	(2, 0)	(1, 1)

Figure 4.3 The Stag-Hunt Game

> Was a deer to be taken? Everyone saw that to succeed he must faithfully stand to his post; but suppose a hare to have slipped by within reach of any one of them, it is not to be doubted that he pursued it without scruple, and when he had seized his prey never reproached himself with having made his companions miss theirs.
>
> (Rousseau 2002 [1755]: 116)

This passage is often interpreted as describing the following situation (cf. Skyrms 2004). Two individuals go out on a hunt. Each can individually choose to hunt a stag or hunt a hare. Each player must choose an action without knowing the choice of the other. If an individual hunts a stag, he must have the cooperation of his partner in order to succeed. An individual can get a hare by himself, but a hare is worth less than a stag.

The first thing to notice about the Stag-Hunt Game is that there are no dominated strategies. "Hare" is the inferior strategy only when the other hunter plays "stag," and "stag" is inferior when the other plays "hare." We can nevertheless find Nash equilibria, namely ("stag," "stag") and ("hare," "hare"): when one hunter plays "stag," the other hunter cannot improve his payoff by playing "hare"; the same is true for "hare."

There is an obvious sense in which ("stag," "stag") is a "better" solution than ("hare," "hare"): both players receive a higher payoff, which means that both players prefer that outcome. But the Nash equilibrium concept by itself does not allow us to choose among different equilibria. To do so, various "refinements" of the concept have been proposed which we will consider below.

The main difference between strategic- and extensive-form games is that the latter allows the representation of sequential moves. Suppose, counter to fact, that in the Eurobonds Game the North moves first and plays "frugal." If the South knows this, it will play "profligate" and the outcome will be (0, 5). If the North plays "profligate" and the South knows this, it will also play "profligate," resulting in (1, 1). The North will anticipate the South's move and therefore play "profligate." Figure 4.2 shows this sequence of moves. However, in the original set-up, the players were assumed to move simultaneously or, more accurately, ignorant of the other player's move. In extensive games such assumptions about knowledge can be represented by the device of an *information set*. The dotted lines around the South's nodes indicate that the South does not know whether it is at the left or at the right node.

In other sequential games the first mover's actions are known. Consider Figure 4.4, called the Ultimatum Game. In an Ultimatum Game (another philosophers' favorite), player 1 (the "proposer") divides a cake and player 2 (the "responder") responds by either accepting the division, in which case both players receive the amount allocated by player 1, or rejecting it, in which case both players receive zero.

(Figure 4.4 is simplified; more accurate would be a representation in which there are as many branches from the proposer node as there are ways

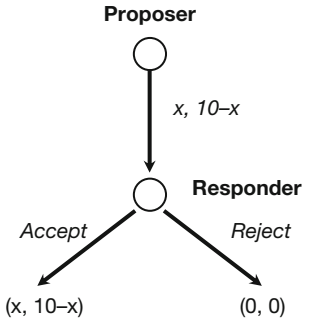

Figure 4.4 The Ultimatum Game

to divide the cake; for instance, one for 10–0, one for 9–1, and so on.) In this game, the responder will accept any offer greater than zero, which will be predicted by the proposer, who therefore offers the minimum amount possible. Sequential games are solved by what is called *backward induction*. Backward induction reasons from outcomes backwards to earlier decision problems. I will say more about the method below. In this simple game, there are only two stages. The responder makes the final (second) move, accepting or rejecting the offer. He will prefer to accept any non-zero offer. Knowing this, the proposer decides about the size of the offer. Maximizing, he will offer the smallest amount the responder will still accept, i.e., the smallest possible non-zero offer.

There are two final distinctions I want to introduce. The first is the distinction between *one-shot* and *repeated* games. *Finitely* repeated games are analyzed in exactly the same manner as one-shot (sequential) games, by backward induction. Suppose the Eurobonds Game of Figure 4.2 was repeated 20 times. At the very last stage, South would still do better by playing "profligate," which would be predicted by the North. The North would at the second-but-last stage also play "profligate," which in turn would be predicted by the South. This reasoning continues all the way to stage one.

Infinitely repeated games (or games in which agents do not know the number of rounds that are being played) are really a different matter. Here "always play "profligate" is no longer a dominant strategy (Aumann 1959). Rather, other strategies have higher payoffs. One such strategy is known as "tit-for-tat." "Tit-for-tat" means that the player starts by playing a "nice" strategy (in this case, "frugal") and then observes what the other player does. If the other player's move is also "nice," he continues doing the same. If the other player plays selfishly (in this case, "profligate"), the first player retaliates by doing so himself now and in future games (Rapoport and Chammah 1965).

Finally, there is a distinction between *pure* and *mixed* strategies. So far we have considered only pure strategies, by which deterministic strategies are meant: an agent makes up his mind and plays the chosen strategy. In the

mixed case, a strategy is assigned a probability by the player, who then lets a random device determine which way he chooses. When there are only two options, one can think of the agent's decision as one about how to load a coin (and let the coin toss determine which strategy is in fact chosen).

A simple mixed strategy can be illustrated by the children's game Rock–Paper–Scissors. As can easily be seen in Figure 4.5, there is no Nash equilibrium in pure strategies. If player 1 plays "rock," player 2's best response is "paper"; player 1's best response to that is "scissors," to which player 2 best responds by playing "rock" and so on. If the players had a random device (such as a three-sided coin), how would they "load" it? Suppose player 1 assigns probabilities 90%/5%/5% to the three options. After a while, player 2 would learn these probabilities from player 1's actions, respond by playing "paper" most of the time and therefore win most of the time. The only way to avoid this is by assigning equal probabilities to the three strategies.

More generally speaking, a mixed-strategy Nash equilibrium is found by assigning probabilities to the different strategies such as to make the other players indifferent to what strategy one plays. In a symmetric game such as Rock–Paper–Scissors this is easy to see, but asymmetric games can be solved analogously (Varian 2010: 540).

This ends our brief survey of game-theoretic techniques. No doubt, this was a most rudimentary introduction. As we will see below (and in other chapters), it is enough to appreciate some of the most pressing philosophical problems of game theory. Before delving into these problems, we have to examine the nature of the utilities involved in game theory in slightly more detail.

Game Theory's Payoffs

Economists sometimes use "preference" and "choice" interchangeably. That is, they identify preferences with choices. Given that (ordinal) utility is nothing but an index of preference, one can say: "In the standard approach,

		Player 2		
		Rock	Paper	Scissors
Player 1	Rock	(0, 0)	(–1, 1)	(1, –1)
	Paper	(1, –1)	(0, 0)	(–1, 1)
	Scissors	(–1, 1)	(1, –1)	(0, 0)

Figure 4.5 Rock–Paper–Scissors

the terms 'utility maximization' and 'choice' are synonymous" (Gul and Pesendorfer 2008: 7). This standard approach is the theory of revealed preferences.

We saw in Chapter 3 that revealed-preference theory is untenable as a general theory of preferences for decision theory. Here we will make some additional observations concerning the use of "preference" in game theory. A theory which identified preference with actual choices is a non-starter for game theory applications, because it would be impossible to write down the structure of most games. When both agents are rational in the Eurobonds Game (Figure 4.2), the top left of the diagram will never be reached. And yet, we assign utilities to these outcomes. Or think of Rock–Paper–Scissors. If we identified preferences with choices all we could infer from people playing the game is that they prefer whatever move they make to the other two. Observing them using each strategy a third of the time would lead us to believe that they are indifferent between the three moves. But of course, they are not. Utility cannot mean "preferences as revealed in actual choices."

Some economists therefore say that preferences are identical to *hypothetical* choices:

> In game theory, we are usually interested in deducing how rational people will play games by observing their behavior when making decisions in one-person decision problems. In the Prisoner's Dilemma, we therefore begin by asking what decision Adam *would make* if he *knew* in advance that Eve had chosen dove.
>
> If Adam *would* choose hawk, we would write a larger payoff in the bottom-left cell of his payoff matrix than in the top-left cell. These payoffs may be identified with Adam's utilities for the outcomes (dove, hawk) and (dove, dove), but notice that our story makes it nonsense to say that Adam chooses the former because its utility is greater. The reverse is true. We made the utility of (dove, hawk) greater than the utility of (dove, dove) because we were told that Adam would choose the former. In opting for (dove, hawk) when (dove, dove) is available, we say that Adam reveals a preference for (dove, hawk), which we indicate by assigning it a larger utility than (dove, dove).
>
> (Binmore 2007: 13–14; emphasis added)

Rational-choice theory, accordingly, is a theory of consistent behavior. What we learn about a person's preferences licenses inferences about what that person will do in a similar situation. Unfortunately, this won't do for game theory, either. One problem is that players have preferences over outcomes they are never in the position to choose (cf. Hausman 2012: 34). In an ultimatum game, the proposer clearly prefers the respondent to accept as long as he offers a positive amount (as indicated by his payoff $x > 0$). But he is never in the position to choose between these outcomes, that choice is up to player 2. Of course, one could ask player 1 what he would do if he had to

choose between the outcome (x, accept) and (x, reject). But to do so would be the same as to ask him about his preference between the two situations. The *choice* is up to player 2.

Another problem with identifying preferences and hypothetical choices is that people could not make mistakes. If preferences just were choices (actual or hypothetical) people could not fail to maximize their utility. On the one hand, it is implausible that people never make mistakes in their choices, and therefore a theory that makes it conceptually impossible to make mistakes should be rejected. On the other hand, some of the difficulties economists debate regarding equilibrium selection presuppose that people can make mistakes (see below). Hence, to the extent that we want to be able to make sense of these debates, the revealed-preference theory cannot be used in the context of game theory.

The upshot is that the utilities assigned to outcomes in the games described in this chapter are an index of preference satisfaction, and "preference" here refers to some mental state. All we need is that people can mentally rank the available outcomes. Higher utilities indicate a higher mental rank.

As young teenagers at school we enjoyed writing up what we then called "love lists": if you were a boy you'd rank every girl in class from "like the most" to "like the least," and vice versa for girls. Of course, we boys were never in the position to choose. And yet, we were all more than eager to construct love lists. For the intents and purposes of this chapter, this is what is meant by having preferences.

Game Theory and Rationality

Among other things, game theory purports to answer the question: How does one act rationally in strategic situations? In this section we will see that game theory is not always completely clear about how to answer that question. We will first ask whether game theory's main solution concept, the Nash equilibrium, is philosophically well founded. After that we will look at a number of its refinements.

Is It Always Rational to Play a Nash Equilibrium in One-Shot Games?

Do agents have compelling reason to always play Nash equilibria and to only play Nash equilibria in one-shot games? Are there perhaps considerations that might lead a rational agent to make an out-of-equilibrium move? In this section I will examine a number of arguments that have been made in defense of the Nash equilibrium as a solution concept for one-shot games. More complex games will be considered below.

The most frequently heard defense of the Nash equilibrium is that it constitutes a self-enforcing agreement (e.g., Hargreaves Heap *et al*. 1992: 101). Suppose North and South could meet before playing the Eurobonds

Game. They look at the structure of the game and observe that both could profit by playing "frugal." They therefore agree to do so. In our actual case this was called the "Euro-Plus Pact," which was adopted in 2011 in order to pave the way for Eurobonds. What are the chances of the players sticking to their agreement? In the structure of Figure 4.1, nil. Unless there is external enforcement—which there is neither in the Eurobonds Game nor in reality—the incentives are strong not to honor the agreement. To promise to stick to a certain course of action in such a situation amounts to no more than "cheap talk."

By contrast, if the players agreed beforehand to play "profligate" instead, neither would have an incentive to deviate. "Thus, the Nash equilibrium can be seen as the only sustainable outcome of rational negotiations in the absence of externally enforceable agreements" (Hargreaves Heap *et al*. 1992: 102). Are all and only Nash equilibria self-enforcing? It appears not. Consider the game in Figure 4.6 (Risse 2000: 366).

Here ("bottom," "right") is the only Nash equilibrium. But both players have incentives to deviate from it, provided they believe that the other does so too. Why wouldn't rational players believe that other rational players do something that is better for them?

Another example is the Stag-Hunt Game of Figure 4.3. Here ("stag," "stag") is a Nash equilibrium but a fragile one. Suppose the hunters meet beforehand and agree on the Pareto-superior outcome. To stick to the agreement requires a lot of trust. What if hunter 1 suspects hunter 2 of not honoring the agreement (because hunter 1 knows that hunter 2 is very suspicious, say, and therefore might not trust hunter 1 either, or simply because he may make a mistake)? He might then decide to play "hare," which secures him a payoff of at least 1. The same reasoning applies to hunter 2, of course. Thus, while ("stag," "stag") is the Pareto-superior outcome ("hare," "hare") is the less risky outcome. Less risky outcomes can be preferable in situation where one cannot be so sure that others are trusting (even if one is trusting oneself) or when people are prone to make mistakes.

		Player 2		
		Left	Centre	Right
Player 1	Top	(4, 6)	(5, 4)	(0, 0)
	Middle	(5, 7)	(4, 8)	(0, 0)
	Bottom	(0, 0)	(0, 0)	(1, 1)

Figure 4.6 A Non-Self-Enforcing Nash Equilibrium

Risk considerations also motivate the next game (Figure 4.7) which shows that a self-enforcing agreement does not have to be a Nash equilibrium (Risse 2000: 368).

If the players have a chance to negotiate before playing, ("bottom," "right") suggests itself as point of agreement. Of course it is true that both players have incentives to deviate from this outcome. But if they did as they preferred, they would risk ending up with nothing, as long as they believe that the other might also not honor the agreement.

Lesson: the self-enforcing agreement argument cannot be used as a justification of the Nash equilibrium because it is neither the case that all Nash equilibria are self-enforcing agreements nor that all self-agreements have to be Nash equilibria.

Another influential defense of the concept is that playing a Nash equilibrium is required by the players' rationality and common knowledge thereof. Consider the Figure 4.8 (Hargreaves Heap and Varoufakis 2004: game 2.9).

Here every strategy is a best reply to some strategy played by the opponent. If player 1 believed that player 2 would play "right," he will play "bottom." Why would player 1 believe that player 2 will play "right"? Perhaps because he believes that player 2 expects him to play "top," and "right" is an optimal

		Player 2	
		Left	Right
Player 1	Top	(0, 0)	(4, 2)
	Bottom	(2, 4)	(3, 3)

Figure 4.7 A Self-Enforcing Non-Nash Equilibrium

		Player 2		
		Left	Centre	Right
Player 1	Top	(3, 2)	(0, 0)	(2, 3)
	Middle	(0, 0)	(1, 1)	(0, 0)
	Bottom	(2, 3)	(0, 0)	(3, 2)

Figure 4.8 A Rationalizable Strategy

response to that. The problem with this kind of reasoning is that it cannot apply to everyone sharing the same beliefs: if the agents actually played a set of strategies such as ("bottom," "right"), at least one player's expectations would be frustrated. The only strategy set that avoids regretting one's beliefs or actions is the Nash-equilibrium strategy ("middle," "center").

The problem with this defense is that it is only plausible when there is a unique rational way for each player to play the game. But this is not always the case. Figure 4.9 shows that considerations of riskiness could sometimes trump the Nash equilibrium (Hargreaves Heap and Varoufakis 2004: game 2.11).

In this game ("top," "left") is the unique Nash equilibrium in pure strategies. Will rational players play it? If player 1 plays "top" he risks ending up in the ("top," "right") cell, which he prefers least because player 2 is indifferent between "left" and "right." The same reasoning applies to player 2's playing "left." By contrast, playing ("bottom," "right") means that both players end up with payoff 1, no matter what the other player does. Why would a rational agent not play a non-equilibrium strategy *guaranteeing* him a payoff which he can reach playing the Nash equilibrium only by assuming a great risk?

Another problem with this defense is that the idea that there is a specific kind of recommendation of what to do in strategic situations is somewhat incoherent. As Isaac Levi points out, an agent who wants to use the principles of rational choice critically cannot predict that he will act rationally (Levi 1997). But arguably, game theory portrays agents as deliberating about what they are going to do given the preference structure of the game and relevant beliefs. Game theory, so the objection goes, cannot assume that players know that they are rational, and therefore *a fortiori* it cannot assume that they know that the other players are rational.

		Player 2		
		Left	Centre	Right
Player 1	Top	(1, 1)	(2, 0)	(–2, 1)
	Middle	(0, 2)	(1, 1)	(2, 1)
	Bottom	(1, –2)	(1, 2)	(1, 1)

Figure 4.9 Risk-Dominant Non-Nash Equilibria

Refinements of the Nash Equilibrium

We could end our discussion of "game theory and rationality" right here because the Nash equilibrium is *the* central solution concept of the theory. But the above criticisms pertained only to one-shot games, and there are at least potential defenses of the Nash equilibrium as a result of rational learning (Kalai and Lehrer 1993) or evolution (Binmore 1987). I will not consider these defenses here and instead take a look at another fundamental problem of game theory as theory of rationality: there are almost always multiple Nash equilibria in a game, and game theory does not provide good advice about equilibrium selection. Thus, even if the Nash equilibrium were defensible from the point of view of rationality, no defense would gain much ground because "solve the game by finding the Nash equilibria" underdetermines what rational players should do in most games.

A class of games with multiple equilibria in which players mutually gain by performing the same action is called "coordination games." The structure of the simplest one is shown in Figure 4.10.

Suppose two drivers (or horse riders) have to coordinate whether to veer left or right in order to avoid a collision. Doing either would be fine as long as both do the same. How, in the absence of a government that can enforce rules, do rational agents decide which way to go?

Thomas Schelling proposed that most situations of this kind have "some focal point for each person's expectation of what the other expects him to expect to be expected to do" (Schelling 1960: 57). Back in the day when people traveled on horseback, keeping left may have been a focal point because most people are right-handed. By keeping left, horsemen could hold the reins in their left hand and keep their right hand free to greet the oncoming rider or pull out their sword (Parkinson 1957).

In this particular story, right-handed riders would probably have a *preference* for keeping on the left so that the payoffs in the ("left," "left") equilibrium should be higher than the payoffs in the alternative equilibrium. But in other situations the choice of equilibrium is driven by expectations alone. An example Schelling uses is this: "Name 'heads' or 'tails.' If you and

		Player 2	
		Left	Right
Player 1	Top	(1, 1)	(0, 0)
	Bottom	(0, 0)	(1, 1)

Figure 4.10 A Simple Coordination Game

your partner name the same, you both win a prize" (Schelling 1960: 56). Arguably, there is nothing intrinsically more valuable about "heads" than about tails; there is a reason to *expect* people to choose heads, though, and heads is therefore a focal point, because it is customarily mentioned first.

Focal points are a reasonable way to choose among equilibria, based on expectations of what people might do because of the existence of certain customs, habits or conventions. But it is not clear how to model focal points formally, which is why the theory remains rather undeveloped to this day. If considerations regarding focal points affect the payoffs (as they should at least sometimes), the focal-point strategy is also the Pareto-dominant strategy. This is another refinement of the Nash equilibrium: if there are multiple Nash equilibria, choose the Pareto-dominant outcome.

Pareto dominance sometimes conflicts, however, with another idea we have already encountered: risk dominance. The Stag-Hunt Game (Figure 4.3) shows as much. ("stag," "stag") is the Pareto-dominant outcome. If, however, for whatever reason, the other player chooses "hare," the hunter choosing "stag" ends up in the least preferred state. Playing "stag" is thus risky. By playing "hare," by contrast, the hunter can only gain: if the other hunter makes a mistake and plays "stag," the first ends up with a payoff that is *higher* than the equilibrium payoff.

Playing the risk-dominant rather than Pareto-dominant strategy has sometimes been defended on the basis of evolutionary considerations (Kandori *et al*. 1993; Young 1993). But is it always or even most of the time rational to play the risk-dominant strategy? To give up risk also means to give up the extra profit, and surely it is not always in one's best interest to do so. At any rate, civilization is to a large degree built on trust, and without mutual trust little economic interaction and development would be possible. To trust (the other not to make a mistake in this case) means to play "stag," despite being the risk-dominated strategy. Rationality considerations by themselves cannot decide between these two refinements, at least not generally.

To introduce another series of refinements, consider the Battle of the Sexes Game (Figure 4.11).

		Player 2	
		Bananarama	Schoenberg
Player 1	Bananarama	(4, 2)	(0, 0)
	Schoenberg	(0, 0)	(2, 4)

Figure 4.11 The Battle of the Sexes Game

A story that goes along with the game could be as follows. Unbeknownst to each other, one morning a couple bought tickets to different musical events on that same night. When they go for lunch that day, they agree to go to the same event come what may, but she had to run off to another meeting before they could decide which one. In an extraordinary stroke of bad luck, both of their iPhones die on them just after lunch, and both have out-of-office appointments all afternoon. Given they cannot communicate, how do they coordinate their evening? Player 1, Claire, prefers Bananarama to Schoenberg, and going to either concert with her partner to going alone. Player 2, Jeff, prefers Schoenberg to Bananarama, and also going to either concert with Claire to going alone.

There are two obvious Nash equilibria (in pure strategies), and that would be the end of the story if it wasn't for refinements of the concept. If that was the end of the story, game theory would really not be very helpful in determining what rational agents ought to do in strategic situations. Who would have thought that it's rational for Jeff to go to the Bananarama concert if Claire does and to the Schoenberg if she goes there (and vice versa for Claire)?

Luckily, there are further refinements, one of which we've in fact already encountered: the subgame-perfect Nash equilibrium. A strategy profile is a subgame-perfect equilibrium if it represents a Nash equilibrium of every subgame of the original game. The concept applies only to sequential games, so let us amend the story a little. Let us suppose that Claire works until 6 p.m. that day and Jeff until 7.30 p.m., and both know that. So Claire will make her decision first (which is known by Jeff). The resulting *sequential* game is depicted in Figure 4.12.

Informally speaking, a subgame is game that begins at any node of the original game and contains that node and all its successors (unless the subgame's initial node is in an information set which is not a singleton; in that case all nodes in that information set belong to the subgame). This

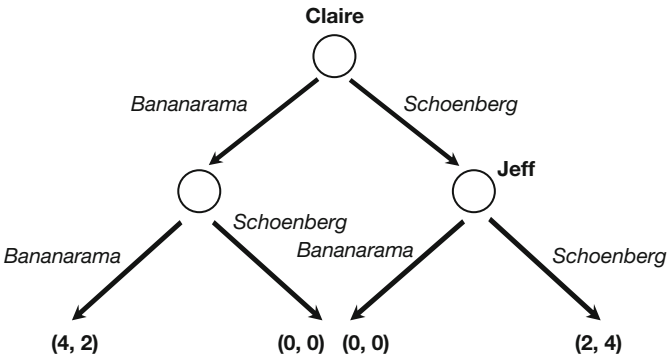

Figure 4.12 The Battle of the Sexes in Extensive Form

game has three subgames: the original three-node game (every game is a subgame of itself) and two games that begin at Jeff's decision points. The Eurobonds Game of Figure 4.2 has only two subgames: the original game and the subgame that begins with South's moves (because both nodes of that stage of the game are in the same information set).

As we have seen, one solves a game in extensive form by backward induction. Backward induction is a method to find the subgame-perfect equilibrium of a game. In the left node, Jeff would choose "Bananarama," and "Schoenberg" in the right node. Predicting that, Claire chooses "Bananarama," and thus the only subgame-perfect equilibrium in this game is ("Bananarama," "Bananarama"). The purpose of backward induction is to eliminate incredible threats. Jeff can threaten all he wants to go to Schoenberg, because once Claire has made up her mind, he could only lose by exercising his threat. By invoking an asymmetry—allowing Claire to move first—we can reduce the number of pure-strategy Nash equilibria from two to one.

The elimination of incredible threats creates a paradoxical situation, however, which can be illustrated by the four-stage Centipede Game of Figure 4.13. Its subgame-perfect solution is for player 1 to move "down" at the first stage (we are moving left to right instead of top to bottom). This is because at the last stage, player 2 would move "down," which is predicted by player 1, who would therefore move "down" at the third stage, and so on. At any intermediate stage, either player may ask: "How did we get here?" Consider stage 2. Given player 1's rational move at stage 1 was "down," the only explanation that we reached this stage is that player 1 is irrational or made a mistake. But if so, player 2 may expect that player 1 continues to play that way, and intend to stop at a later stage when the payoffs are higher. Rationality for player 2 in fact may *require* that he does not play "down" at stage 2 if he believes that player 1 is not fully rational. Suppose player 2 believes that player 1 moves "right" no matter what. If so, player 2's best move is to wait until the last stage of the game to move "down." Thus, if the players do not assume other players to be rational, their own rationality allows or asks of them to continue in the game. Indeed, in empirical tests of the game, subjects play a few rounds before ending the game (see for instance McKelvey and Palfrey 1992).

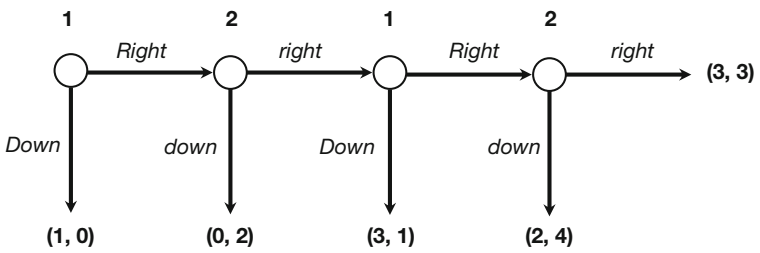

Figure 4.13 The Centipede Game

To *end up* in the subgame-perfect equilibrium outcome (1, 0), we therefore have to assume both players be perfectly rational and have common knowledge of rationality. But if we make these assumptions, we can never *reason* our way towards that outcome.

One standard way around this paradox in the literature is to invoke so-called "trembling hands" (Selten 1975). A player's hand trembles when she makes mistakes despite being fully rational. Player 1 might for instance be resolved to play "down" at the first node but then accidentally push the wrong button and go "right." As long as this happens with some positive probability, the game can be solved by backward induction.

The idea of a trembling-hand (perfect) equilibrium further refines the subgame-perfect equilibrium, as can be seen in the Tremble Game of Figure 4.14 (Rasmusen 2006: 111). In this game there are three Nash equilibria: ("top," "left"), ("top," "right") and ("bottom," "left"), two of which are subgame perfect: ("top," "left") and ("bottom," "left"). However, the possibility of trembling rules out ("bottom," "left") as an equilibrium. If player 2 has a chance of trembling, player 1 will prefer to play "top" in order to secure his payoff of 1. Player 2 chooses "left" because if player 1 trembles and plays "bottom" by mistake, she prefers "left" to "right." ("top," "left") is therefore the unique trembling-hand equilibrium.

The "standard way around" the paradox is not convincing. Depending on how large one player estimates the probability with which the other player's hand trembles and his own payoffs, it might well be rational to continue in the Centipede Game. It does not matter whether the opponent is assumed to be irrational or make mistakes, the consequences for each player's deliberations of what best to do next are the same. To respond that trembles occur randomly and with small probability does not help. In order to reach the final stage of the game quite a few trembles have to have occurred. Trembles influence results systematically.

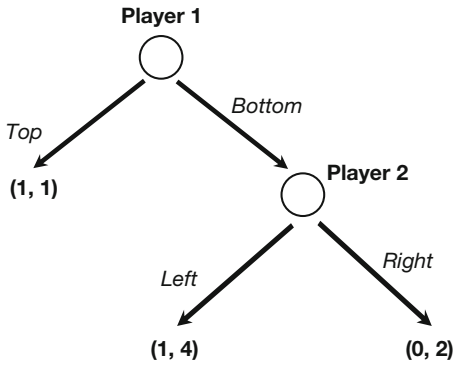

Figure 4.14 The Tremble Game

Let us examine a second paradox. In the Chain Store Game, a Dutch monopolist called Alban Heym has branches in 20 towns. The chain faces 20 potential competitors, one in each town, who can choose "enter" or "stay out." They do so sequentially and one at a time. If a potential competitor chooses "stay out," he receives a payoff of 1, while the monopolist receives a payoff of 5. If he chooses "enter," he receives a payoff of either 2 or 0, depending on the monopolist's response to his action. The monopolist must choose between two strategies, "cooperative" or "aggressive." If the monopolist chooses the former, he and the competitor receive a payoff of 2, and if he chooses the latter, each player receives a payoff of 0. The last round of the game is depicted in Figure 4.15.

The subgame-perfect equilibrium is easy to see. Entering into a price war, Alban Heym can only lose. The competitor knows this, and since he has a higher payoff when the market is shared than when he stays out of it, he will enter. ("enter," "cooperative") is the equilibrium reached by backward induction in this game.

But the managers of Alban Heym think they can outwit the competitor. They reason that if Alban Heym demonstrates being tough by playing "aggressive" in early rounds, potential entrants will be deterred and the chain store can reap higher profits (cf. Selten 1978). A game theorist might respond: "Look, your threat is not credible. You will most certainly not enter a price war in the final round—there is nothing to gain from it. Thus, the competitor will enter for sure. In the penultimate round, there is again no reason to fight. It will be costly, and it has no effect on the final round. Continue to reason thusly, and you will see that deterrence is not a valid strategy."

Indeed, the managers of Alban Heym made a mistake. If "aggressive" is a preferable strategy in early rounds of the game, then this should be reflected in the payoffs. The managers analyzed a different game, not the Chain Store Game. However, one can show that if there is a small amount

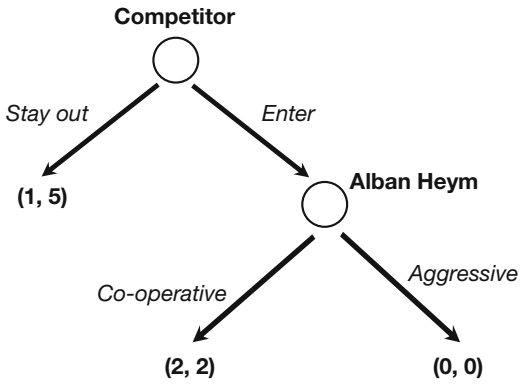

Figure 4.15 The Chain Store Game

of uncertainty about the payoffs, it may be rational for a monopolist to build up a reputation by fighting entry initially. This leads to another refinement of the subgame-perfect equilibrium called *sequential equilibrium* (Kreps and Wilson 1982).

The idea is that at least one player's profile is determined exogenously and unobservable to the other player. In the Chain Store Game, for instance, the monopolist may be "weak" or "strong" depending on whether his preferences are as in Figure 4.15 or the converse, as on the right-hand side of Figure 4.16.

The game is solved by finding for a player *i*, at each information set after which *i* is to play, the best reply to the strategies of the other players. In this game, Alban Heym can build a reputation of being tough and thereby deter potential entrants from trying to enter the market.

In each case considered above, the "refinement" of the Nash equilibrium consisted in adding a structural feature to the game which changed its nature. In no case was the refinement justified on the basis of considerations of what would be reasonable in a situation to do. One consequence is that the refinements often do not reduce indeterminacy but rather add to it. Both the Centipede and the Chain Store Game show that in order to settle on specific equilibria one has to weaken common knowledge assumptions, which means that one is moving away from attempting to defeat indeterminacy.

Game theory, understood as a theory of rational decision-making, is thus highly problematic. The Nash equilibrium is ill-justified. Even if it were justified, it would solve few problems because most games have multiple Nash equilibria. Thus far, the refinement program has produced few results that can be defended from the point of view of rationality and that have helped to reduce indeterminacy.

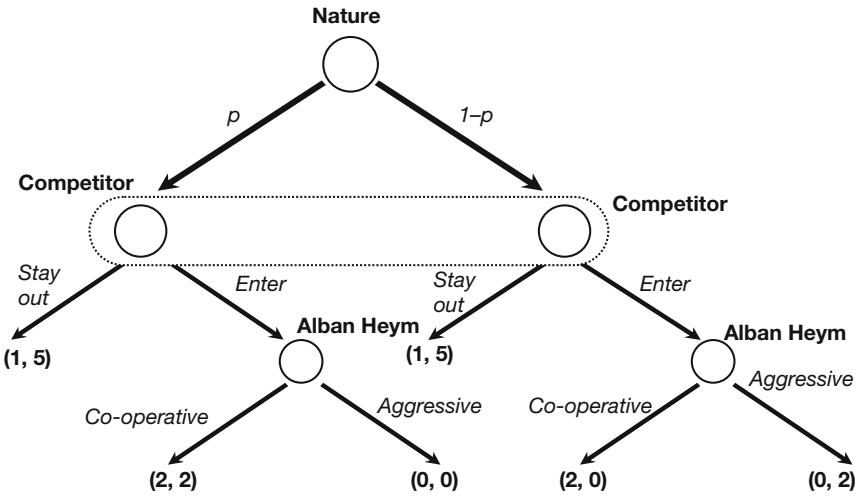

Figure 4.16 The Chain Store Game with Uncertainty

Game Theory as Explanatory Theory

It has sometimes been argued that game theory, as theory of rational choice, would be a very good candidate for an empirical theory of social phenomena (see for instance Grüne-Yanoff and Lehtinen 2012). A theory which predicts that people act rationally is self-fulfilling. People who accept the theory and therefore predict that other people behave rationally have an incentive to act rationally themselves. A theory which predicts that other people behave *ir*rationally does not have this benefit. Agents acting on such a theory have all the more reason to deviate from the theory because doing so will improve their performance. Rational-choice theories therefore have a stabilizing influence on social phenomena.

But of course this argument has any bite only to the extent that game theory is successful as a theory of rationality. As we saw in the last section, it is not. Not all is lost yet, though. Perhaps game theory isn't so good a theory of rationality but it might nevertheless be a useful predictive and explanatory theory. Perhaps the justification of playing a Nash equilibrium strategy in this or that refinement is wanting, but if people play the strategies the theory predicts anyway, who cares? As long as the theory has empirical content, or, more carefully, as long as it is useful for modeling empirical phenomena, it may have its virtues. In this section we will see that even as explanatory or predictive theory, game theory is very problematic.

In order to make predictions (only a theory that predicts empirical phenomena can also explain them), any theory must in one way or another be brought to bear on empirical phenomena. Most genuine theories transcend the realm of the observable, i.e., they contain terms that refer to unobservable states of affairs. If the arguments that were given here concerning the nature of preferences are correct, one of the core elements of game theory refers to something unobservable. Carl Hempel called the principles that connect a theory's theoretical (or unobservable) vocabulary with an observational vocabulary "bridge principles" (Hempel 1966). We need bridge principles for a theory to have empirical content.

The architecture of games is given by their payoff structure. As we saw in the previous section, the payoffs are utility indices indicating preference ranking. This is unfortunate. If they indicated material outcomes, the game theorist could straightforwardly determine which game is being played in a given situation because material outcomes are observable (perhaps not literally, but for all intents and purposes of the game theorist). The usual story that goes along with what I have called the Eurobonds Game is one about a proposal made to two prisoners, which is why the game is normally known as Prisoner's Dilemma:

> In the Prisoner's Dilemma, two prisoners … are being interrogated separately. If both confess, each is sentenced to eight years in prison; if

> both deny their involvement, each is sentenced to one year. If just one confesses, he is released but the other prisoner is sentenced to ten years.
> (Rasmusen 2006: 20; footnote suppressed)

The corresponding matrix could be written as shown in Figure 4.17.

		Prisoner 2	
		Confess	Deny
Prisoner 1	Confess	(8 years, 8 years)	(0, 10 years)
	Deny	(10 years, 0)	(1 year, 1 year)

Figure 4.17 The Prisoner's Dilemma in Game Form

The advantage of presenting games in this way—that application to empirical situations is easier—is frustrated by the fact that game theory could not make any predictions without knowledge of players' preferences. In some cases, what people prefer may be relatively straightforward, such as here. It is very reasonable to assume that almost everybody has the following preference ranking: acquittal > 1 year > 8 years > 10 years. In other words, it is reasonable to assume that utility is strictly decreasing by number of years in prison and therefore that the Eurobonds Game of Figure 4.2 is an adequate transformation of the game form of Figure 4.17 (for the notion of a game form, see Weibull 2004).

What is reasonable to assume in one case should not be blindly accepted as a rule more generally. In other words, one should allow the utility function:

$$U = U(M),$$

where M designates the material outcomes of a game, to vary between people and across situations. Although in early experimental applications of game theory, subjects were assumed to care only about their own material gains and losses, this is clearly *not* a substantive hypothesis of game theory as such. As game theorist and experimenter Ken Binmore remarks:

> Actually, it isn't axiomatic in economics that people are relentlessly selfish. … Everybody agrees that money isn't everything. Even Milton Friedman used to be kind to animals and give money to charity.
> (Binmore 2007: 48)

Thus, institutions, social and cultural norms and other immaterial facts may affect people's valuations of the material outcomes of a game. It is by no means obvious for instance how people rank the different outcomes of the Ultimatum Game (Figure 4.4, above). If a strong fairness norm is at work, players might rank (material) outcomes as follows (5, 5) > (0, 0) > (6, 4) > (4, 6) > (7, 3), etc., which is not at all strictly increasing in material outcomes.

There is now a growing literature on what functional form U might take. For instance, in Fehr and Schmidt's theory of fairness, the utility function for two-person games has the following form (Fehr and Schmidt 1999: 822):

$$U_i(x) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\},\ i \neq j,$$

where the x's are player i's and j's monetary payoffs and α and β are parameters measuring how disadvantageous and advantageous inequality affects a player's utility. In Cristina Bicchieri's formulation, social norms are explicit arguments in the function (Bicchieri 2006: 52):

$$U_i(s)=\pi_i(s) - k_i \max_{s_{-j} \neq L_{-j}} \max_{m \neq j}\{\pi_m(s_{-j}, N_j(s_{-j})) - \pi_m(s), 0\},$$

where $s = (s_1, s_2, ..., s_n)$ is a strategy profile, $\pi_i(s)$ is the material payoff function for player i and $k_i \geq 0$ is a constant representing a player's sensitivity to the relevant norm. A norm for player i is represented by the function N_i: $L_{-i} \rightarrow S_i$, where $L_{-i} \subseteq S_{-i}$, S_i is player i's strategy set and S_{-I} the set of strategy profiles for the other players.

Generally speaking, though, economists are loath to make any substantial assumptions about people's utility functions. They instead believe that one can learn people's preferences in one situation and use that knowledge to make predictions about what the same people will prefer in another situation. Learning about what people prefer in choice situations is called "preference elicitation." Often this is done by having experimental subjects play a subgame of a game of interest. Suppose we are interested in eliciting the players' preferences in the Eurobonds Game. First we have to write down the game in game form with material outcomes. For simplicity, let us assume that the main material outcome is growth rates. In extensive form, the game could then look as shown in Figure 4.18.

To elicit the South's preferences, we have it choose over strategies in the subgames (Figure 4.19).

To elicit the North's preferences we simply switch roles (we can do so because the material payoffs are symmetrical). If both North and South prefer "profligate" to "frugal" in both subgames we know that Figure 4.2 is a correct rendering of the game, and we can use game-theoretic tools to solve it and make a prediction.

The problem with this elicitation procedure is that it assumes that preferences are quite strongly context-independent. What that means, and why this is not always a reasonable and safe assumption to make, can be

illustrated with the following game, called the "Trust Game" (Figure 4.20). Here an investor can choose between keeping his investment or transferring it to a trustee. If he does the latter the money is quintupled. The trustee then decides whether to keep the money or return half of it to the investor.

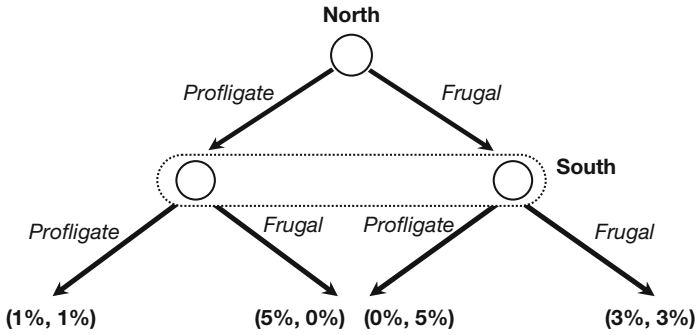

Figure 4.18 The Eurobonds Game in Game Form

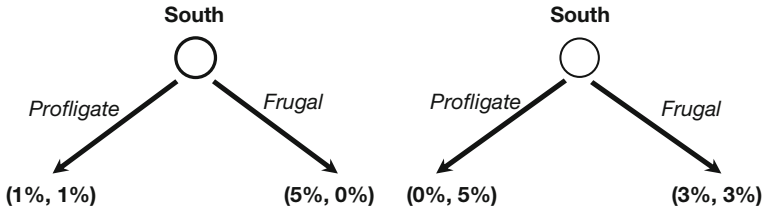

Figure 4.19 Two Subgames of the Eurobonds Game in Game Form

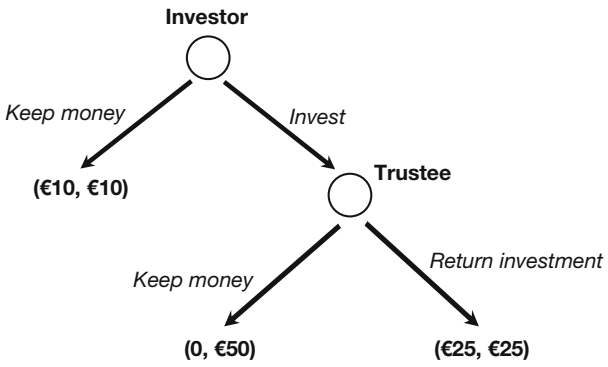

Figure 4.20 The Trust Game in Game Form

Consider the subgame beginning at the second decision node. On its own this subgame is in fact a version of the so-called Dictator Game in which a player decides whether or not to split a pie between herself and a second player. Unlike in the Ultimatum Game (recall Figure 4.4), in the Dictator Game the second player does not have the option to "reject," and the first player therefore does not have to fear punishment. It is thus in the subgame of the Trust Game. We can then expect people to choose similarly in both games, and indeed, the assumption that people choose similarly is implicitly made in using this elicitation procedure.

Suppose an individual chooses "Keep money" in the subgame when played on its own. Is it reasonable to expect her to do the same in the full game? When the full game is played, other norms may affect players' decisions than when only the subgame is played. If I am trusted I may wish to reward the trusting partner by returning his investment, even at a material cost to me, and even when the investor has no way to punish me for a selfish decision. Or I may act on equity considerations. Unlike in the Dictator Game, in the Trust Game the investor helps to produce the pie, and I might therefore think that she deserves a return or that I am obliged to pay her back because it is her money.

Be the normative considerations as they may, subjects do in fact choose differently in the Dictator Game and in the Investment Game (see for instance Cox 2004). This means that their preferences over outcomes are not context-independent. An aspect of the relevant context in this case is whether another player has moved first and thereby helped to create the opportunity to "reward" or "pay back." Without context independence, this particular elicitation procedure is invalid.

There are other problems with this procedure. Francesco Guala points out that it cannot be used to test games in which reciprocity matters (Guala 2006). That this is correct is not hard to see: if my being "nice" (in the sense of moving in such a way as to give a co-player a higher payoff than I could have done had I chosen differently) depends on my co-player's being "nice"; i.e., if my preferences depend on my co-player's preferences, one cannot learn about these preferences in situations where other players do not exist.

It is important to note, however, that game theory is not wedded to this particular elicitation procedure, or any other procedure for that matter. Weibull 2004, for instance, proposes for a slightly more complicated version of a Dictator Game in which four outcomes have to be ranked that "the experimentalist asks subject *A* to state her ordinal ranking of the four plays for each of the 24 possible (strict) ordinal rankings that *B* may have, and likewise for subject *B*." The experimenter could then find a matching pair, inform each player about the other's preference and make a prediction on that basis.

There are numerous problems with this proposal. First, it might be too far away from the revealed-preference approach for it to be appealing to economists. Even if one refuses to *identify* preferences with choices, one could still

hold that choices are *the best guide* to people's preferences and thus insist on people's preferences being elicited in choice tasks. Second and relatedly, as we saw in the previous chapter, people sometimes reverse their preferences between a valuation-task and a choice-task (Lichtenstein and Slovic 1971, 1973). It is therefore not clear that asking people to rank outcomes is a reliable guide to what they will do later in the choice-task. Third, there may be no matching pair of rankings. Fourth, even if the procedure worked for laboratory experiments, it is quite obviously completely unusable for predicting and explaining field phenomena.

There is a deeper issue which has nothing to do with any of the specific problems of specific elicitation procedures. Any theory needs bridge principles in order to be applicable to empirical phenomena for theory testing, prediction and explanation. In game theory, bridge principles come in two types. Type one are assumptions about the form of people's utility functions. As mentioned above, if people's utilities could be assumed to be strictly increasing (or decreasing) in their own material gains and losses (and independent of everything else), one could use observations of material outcomes together with the tools of the theory to derive predictions. That simple assumption is implausible, but nothing prevents economists from providing more complex functions in which other people's material gains and losses and social and cultural norms (to give a few examples) play a role. The other type of bridge principle are elicitation procedures. Here preferences are estimated without necessarily assuming that utility functions must have some specific form. The two types are sometimes used jointly. The utility functions in the literature on fairness and social norms all come with free parameters which have to be estimated from people's choice behavior. Here, then, a type-one assumption about the general form of a utility function is combined with a type-two elicitation procedure that fills in the details.

So far, so good. The problem is that none of these bridge principles is part of the core set of claims of the theory. Rather, there is quite a large menu of potential principles to choose from, and economists make free use of the menu in specific applications. But the choice is not systematic and well reasoned. *If* the assumption that people always prefer more money for themselves to less were part of game theory, then the theory could be used for predictions and explanations. But it's not (see the Binmore quotation above), and for a good reason: the theory would have been refuted in countless instances. *If* preference elicitation through choices in subgames were part of the core of game theory, preference interdependence could not be represented within the theory. But it's not, and for an equally good reason: preference interdependence is an empirically important phenomenon.

Economists do not like to make substantial assumptions of this kind. Their theory of rationality is a "formal," not a "substantial" theory, we are told. But a formal theory does not by itself allow the prediction and explanation of empirical phenomena. Thus bridge principles, which provide the theory with substance, are added on an ad hoc basis. The ad hocness of the

conjoining of core theory with principles is problematic for both predictions and explanations. Genuine predictions cannot be made because it is known only after the fact which of a variety of bridge principles should be used in conjunction with the theory. After the fact we can always find some game-theoretical model that in conjunction with some bridge principle captures the phenomenon of interest. But have we thereby explained it? There are three main accounts, models or senses of explanation: rational-choice, causal and unification (see Chapters 2 and 7). Game theory is not a theory of rationality (or if it is, it's a very problematic one), and thus game-theoretic "explanations" are no rational-choice explanations. Game theorists are adamant that they do not intend to model the underlying decision processes of agents that are responsible for people's choices, and therefore they are no causal explanations, either. Nor are they unifying, as I will argue in Chapter 7. For now let us conclude that as long as there are no more systematic insights into what bridge principles should be regarded as part of the core of the theory (in other words, as long as there is no more systematic work on what utility functions people have—work of the kind we find in Fehr and his colleagues and in Bicchieri—and on elicitation procedures of the kind we find in Weibull), game theory remains deeply problematic, both as a theory of rational choice and as explanatory theory.

Study Questions

1 The Eurobonds Game caricatures the actual situation in which Europe found itself in 2012 quite badly. What are the most important differences? Do you think there is a core of truth within it?
2 Rewrite the games in the section on "Is It Always Rational to Play a Nash Equilibrium in One-Shot Games?" as sequential games. Is the Nash equilibrium now a more convincing solution concept?
3 Compare the criticism of game theory as explanatory theory in the section on "Game Theory as Explanatory Theory" with the criticism of rational-choice theory in the previous chapter. Are there commonalities between the two?
4 Model an empirical strategic situation (such as the Cuba crisis or firms competing for business) as a two-person one-shot game and solve it. What obstacles do you encounter?
5 In your view, what parameters should a utility function that is useful as a bridge principle to apply games to empirical situations have? Defend your answer.

Suggested Readings

Game theory began with von Neumann and Morgenstern 1944. A good critical introduction to the issues discussed here is the chapter on game theory in Hargreaves Heap *et al*. 1992. Grüne-Yanoff 2008, Grüne-Yanoff and Lehtinen 2012 and Ross 2010a are useful philosophical discussions. I learned about the importance of bridge principles for the application of game theory from Guala 2008.