CHAPTER

11 Behavioral Game Theory

11.1 Introduction

Analytical game theory is in many ways a huge success story: it is increasingly becoming the foundation of other subdisciplines of economics (including microeconomics) and it has migrated to philosophy, biology, political science, government, public policy, and elsewhere. But, as we will see in this chapter, its descriptive adequacy and normative correctness are controversial. **Behavioral game theory** aims to study the degree to which analytical game theory succeeds in capturing the behavior of real people engaging in strategic interaction, and proposes extensions of analytical game theory in the interest of capturing that behavior. Yet some of the proposed extensions to analytical game theory do not in fact constitute deviations from neoclassical orthodoxy. Thus, there is nothing distinctively behavioral about some of the models discussed under the heading "behavioral game theory." Other models, however, constitute real deviations from neoclassical orthodoxy.

11.2 Social preferences: Altruism, envy, fairness, and justice

Much of the literature on social preferences is driven by data from two games: the **ultimatum game** and the **dictator game**. Both are played by two agents: a **proposer** (Player I) and a **responder** (Player II). Here, these games are outlined as they are presented to participants in laboratory experiments, where outcomes are described in terms of dollars and cents rather than in terms of the utilities that players derive from them. In order to analyze the interaction, we need to transform the dollars and cents into utilities. Strictly speaking, you do not even know what game the players are playing until you have identified payoffs in utility terms. But doing so is far from obvious, as we will see.

The ultimatum game has two stages. At the outset, the proposer is given a fixed amount of money; for purposes of this discussion, let us suppose it is $10. In the first stage, Player I proposes a division of the dollar amount; that is, the proposer offers some share of the $10 to the other player. The proposer might propose to give it all away (leaving nothing for himself), to give none of it away (leaving all of it for himself), or to offer some fraction of the $10 to the other player (leaving the balance for himself). For example, the proposer might offer $4, leaving $6 for himself. In the second stage, the responder accepts or rejects the proposed division. If she accepts, both ame

success story: it is increas-
plines of economics (includ-
hilosophy, biology, political
re. But, as we will see in this
correctness are controversial.
ree to which analytical game
l people engaging in strategic
al game theory in the interest
posed extensions to analytical
s from neoclassical orthodoxy.
about some of the models dis-
eory." Other models, however,
hodoxy.

lism, envy, fairness,

driven by data from two games:
Both are played by two agents:
r II). Here, these games are out-
laboratory experiments, where
nd cents rather than in terms of
order to analyze the interaction,
to utilities. Strictly speaking, you
playing until you have identified
rom obvious, as we will see.
t the outset, the proposer is given
this discussion, let us suppose it
a division of the dollar amount;
f the $10 to the other player. The
y (leaving nothing for himself), to
himself), or to offer some fraction
e balance for himself). For exam-
$6 for himself. In the second stage,
osed division. If she accepts, both

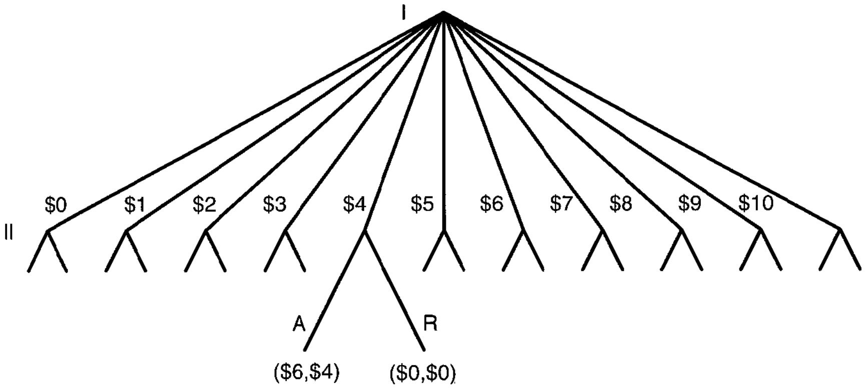

Figure 11.1 The ultimatum game (in dollar terms)

players receive their proposed share; if she rejects, neither player receives anything. The ultimatum game can be represented as in Figure 11.1. In this figure, I have omitted all branches corresponding to fractional amounts, and I have left out all but one set of branches representing Player II's decision in the second stage.

Example 11.1 Dividing the cake When two children have to divide a cake, they sometimes follow a procedure in which the first child splits the cake in two and the second chooses first. Given that Kid II will choose the largest piece, Kid I will want to divide the cake as evenly as possible, thereby guaranteeing a 50–50 split. We can easily imagine a variation of this procedure, in which Kid I proposes a division of the cake, and Kid II gets to approve (in which case each child gets the proposed piece) or disapprove (in which case the parents give the entire cake to the dog). The new procedure would constitute an example of the ultimatum game.

The ultimatum game has been extensively studied by experimental economists. According to Camerer's survey of the results:

> The results ... are very regular. Modal and median ultimatum offers are usually 40–50 percent and means are 30–40 percent. There are hardly any offers in the outlying categories 0, 1–10, and the hyper-fair category 51–100. Offers of 40–50 are rarely rejected. Offers below 20 percent or so are rejected about half the time.

Based on these results, we should expect responders to reject offers below $2 when playing the (one-shot, anonymous) game in Figure 11.1. But such low offers would be rare. By and large, we should expect offers in the $3–$5 range. Many people have drawn the conclusion that these results are inconsistent with analytical game theory.

The observed outcomes are quite consistent with Nash equilibrium predictions, however, even when players care about nothing but their own dollar payoffs. Let us assume that each individual is simply trying to maximize his or her dollar payoffs, and that $u(x) = x$. Player I must choose an amount to offer the other player; suppose that he offers $4. Player II's strategy is a little more convoluted. Because a strategy must specify what a player will do under all possible circumstances (see Section 10.2), Player II's strategy must specify what she will do at each of the nodes where she might find herself. Suppose that Player II rejects all proposed divisions in which Player I offers less than $4 and accepts all others. If so, the two players are in equilibrium. If Player I decreased his offer, it would be rejected, and both would receive nothing; if he increased his offer, it would be accepted, but he would receive less. Given that Player I offers $4, Player II can do no better than accepting. In brief, data on the ultimatum game do not represent a puzzle from the point of view of the theory we learned in Section 10.2, since observed outcomes are consistent with Nash equilibrium predictions. (Given the many equilibria of this game, though, this is not saying much.)

Nevertheless, many people think the Nash equilibrium prediction is problematic, since it requires players to reject positive offers. One way to articulate the problem is to say that the Nash equilibrium in the game requires players to reject what is in effect a dominant strategy (namely, accepting) in the subgame that starts at the node where Player II moves. Another way to articulate the problem is to say that the equilibrium is not subgame perfect and that Player II's threat to reject a low offer is not credible (see Section 10.4). We might, therefore, restrict our analysis to subgame-perfect equilibria. Yet there is only one subgame-perfect equilibrium in this game. In this equilibrium, Player I offers nothing and Player II accepts all offers. This might be counterintuitive. But it is a Nash equilibrium because (a) given Player I's offer, Player II would be no better off if she rejected it, and (b) given that Player II accepts all offers, Player I can do no better than to keep all the money for himself. It is a subgame-perfect equilibrium, because Player II's strategy is also a Nash equilibrium strategy in all subgames: no matter what she has been offered, she cannot improve her payoff by rejecting the offer. A prediction based on the idea of subgame-perfect equilibrium, given our assumption about the two players' utility function, is in fact inconsistent with the experimental results.

The dictator game resembles the ultimatum game, except for the fact that the second stage has been eliminated. In dollar terms and assuming that the proposer starts out with $10, the dictator game can be represented as in Figure 11.2; again, I have left out all the branches representing fractional amounts. On the assumption that the players' utility function remains $u(x) = x$, there is only one Nash equilibrium and therefore only one subgame-perfect equilibrium: the case in which Player I offers nothing to the responder and keeps all the money for himself.

Example 11.2 Charitable donations One example of a dictator game played in the real world involves a person's decision about whether to give money to charity. Whenever you walk by a beggar, for example, you are in effect playing a dictator game in which you must propose some allocation of the money in your pocket with the panhandler. If you walk on, you have in effect picked the maximally selfish, Nash-equilibrium allocation.

st choose an amount to
er II's strategy is a little
at a player will do under
I's strategy must specify
ht find herself. Suppose
Player I offers less than
n equilibrium. If Player I
would receive nothing; if
would receive less. Given
n accepting. In brief, data
rom the point of view of
d outcomes are consistent
y equilibria of this game,

equilibrium prediction
positive offers. One way
equilibrium in the game
minant strategy (namely,
e where Player II moves.
hat the equilibrium is not
t a low offer is not credible
our analysis to subgame-
perfect equilibrium in this
g and Player II accepts all
Nash equilibrium because
better off if she rejected it,
er I can do no better than to
erfect equilibrium, because
trategy in all subgames: no
nprove her payoff by reject-
bgame-perfect equilibrium,
ity function, is in fact incon-

game, except for the fact that
ir terms and assuming that
game can be represented as
nches representing fractional
lity function remains $u(x) = x$,
re only one subgame-perfect
nothing to the responder and

ple of a dictator game played
bout whether to give money to
ample, you are in effect playing
me allocation of the money in
n, you have in effect picked the
l.

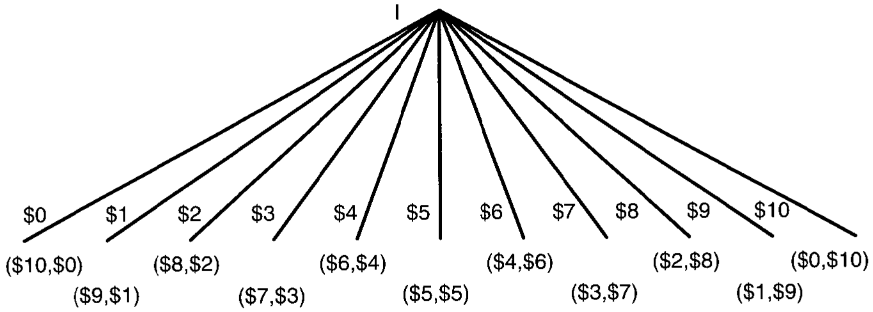

Figure 11.2 The dictator game (in dollar terms)

Experimental evidence suggests that proposers in the (one-shot, anonymous) dictator game typically offer less than proposers in the (one-shot, anonymous) ultimatum game. That said, many proposers are nevertheless willing to share a substantial amount (10–30 percent) of their initial allocation. In the version of the game in Figure 11.2, this means that the proposer would be willing to share $1–$3 with the responder, even though the latter has no way to penalize proposers who offer nothing.

The literature on **social preference** grapples with these phenomena. This literature is based on the assumption that people sometimes care not only about their own attainment but about other people's attainment too. We can model this by assuming that a person *P*'s utility function $u_p(\cdot)$ has two or more arguments. Thus, *P*'s utility function may be given by $u_p(x, y)$, where x is *P*'s attainment and y is the other person *Q*'s attainment.

It is quite possible for *P* to derive positive utility from *Q*'s attainment, so that $u_p(x, y)$ is an increasing function of y. For example, *P*'s utility function may be $u_p(x,y) = \frac{3}{5}\sqrt{x} + \frac{2}{5}\sqrt{y}$. If so, *P* is said to be **altruistic** and to have **altruistic preferences**. Some parents, relatives, friends, and admirers are willing to make real sacrifices in order to improve another person's situation. This is easily explained if we assume that their utility is (in part) a function of the other person's attainment. Altruism in this sense might be what Adam Smith had in mind when he said: "[There] are evidently some principles in [man's] nature, which interest him in the fortune of others, and render their happiness necessary to him" (quoted in Section 1.2).

There is no requirement that *P* derive *positive* utility from *Q*'s attainment. In fact, $u_p(x, y)$ may be a decreasing function of y. For example, *P*'s utility function may be $u_p(x,y) = \sqrt{x} - \sqrt{y}$. This specification entails that *P*'s utility goes up when *Q*'s attainment goes down and *vice versa*. If so, *P* is said to be **envious**. Some Prius hybrid-car owners derive deep satisfaction from rising gasoline prices. This cannot be explained by reference to the financial effects of gasoline prices on the Prius owner: though fuel-efficient, a Prius remains a gasoline-powered car, so rising gasoline prices will hurt Prius owners too. But it can be explained if we assume that the disutility Prius owners derive from getting less gasoline for their dollar is outweighed by the utility they derive from knowing that SUV owners suffer even more.

There is no reason to restrict our analysis to these functional forms. According to a common interpretation of John Rawls's theory of justice (see Section 6.2), societies should be ordered with respect to justice based on the welfare of the least fortunate in each society. Thus, a person with **Rawlsian preferences** might try to maximize the minimum utility associated with the allocation. If each individual derives $\sqrt{x}$ utiles from his or her private consumption x, the Rawlsian P might maximize $u_p(x,y) = \min(\sqrt{x}, \sqrt{y})$. Rawls uses the term "justice as fairness" to describe his theory, so Rawlsian preferences could also be described as preferences for **fairness**.

Another agent might care about the degree of inequality among the relevant agents, so as to rank allocations based on the absolute difference between the best and worst off. Such an agent is said to be **inequality averse** and to have inequality-averse preferences. If each individual derives $\sqrt{x}$ utiles from his or her private consumption x, the inequality-averse P might wish to minimize the absolute difference between each person's utility. This amounts to maximizing $u_p(x,y) = -|\sqrt{x} - \sqrt{y}|$. Such agents care about equality for its own sake, unlike the Rawlsians who (given the definition above) care about equality only insofar as it benefits the least well off. Because the inequality-averse agent ends up assessing the outcomes of ultimatum and dictator games so similarly to the Rawlsian agent, I will not discuss this case further.

Utilitarians like Bentham, whom we came across in Section 1.2, believe that we should pursue the greatest good for the greatest number. Thus, a utilitarian agent might try to maximize the total amount of utility derived from private consumption. If each individual derives $\sqrt{x}$ utiles from his or her private consumption x, the utilitarian P might maximize $u_p(x,y) = \sqrt{x} + \sqrt{y}$. So understood, **utilitarian preferences** constitute a special case of altruistic preferences. Obviously, this list is far from exhaustive: any agent who derives utility from another agent's private consumption counts as having social preferences.

To see how the shape of the proposer's utility function affects his assessment of the various outcomes in the ultimatum and dictator games, see Table 11.1.

Table 11.1 Dictator game utility payoffs (maxima in boldface)

Payoffs	Player P's utility function $u_P(x, y)$				
(x, y)	$\sqrt{x}$	$\sqrt{x} + \sqrt{y}$	$\sqrt{x} - \sqrt{y}$	$\min(\sqrt{x}, \sqrt{y})$	$\frac{3}{5}\sqrt{x} + \frac{2}{5}\sqrt{y}$
(\$10, \$0)	**3.16**	3.16	**3.16**	0.00	1.90
(\$9, \$1)	3.00	4.00	2.00	1.00	2.20
(\$8, \$2)	2.83	4.24	1.41	1.41	2.26
(\$7, \$3)	2.65	4.38	0.91	1.73	**2.28**
(\$6, \$4)	2.45	4.45	0.45	2.00	2.27
(\$5, \$5)	2.24	**4.47**	0.00	**2.24**	2.24
(\$4, \$6)	2.00	4.45	−0.45	2.00	2.18
(\$3, \$7)	1.73	4.38	−0.91	1.73	2.10
(\$2, \$8)	1.41	4.24	−1.41	1.41	1.98
(\$1, \$9)	1.00	4.00	−2.00	1.00	1.80
(\$0, \$10)	0.00	3.16	−3.16	0.00	1.26

hese functional forms.
's theory of justice (see
to justice based on the
person with **Rawlsian**
ility associated with the
his or her private con-
= min $(\sqrt{x}, \sqrt{y})$. Rawls
eory, so Rawlsian prefer-
ess.
equality among the rele-
solute difference between
inequality averse and to
al derives $\sqrt{x}$ utiles from
rse P might wish to mini-
's utility. This amounts to
are about equality for its
finition above) care about
ff. Because the inequality-
matum and dictator games
s this case further.
s in Section 1.2, believe that
t number. Thus, a utilitarian
utility derived from private
from his or her private con-
$(x,y) = \sqrt{x} + \sqrt{y}$. So under-
case of altruistic preferences.
ent who derives utility from
aving social preferences.
unction affects his assessment
ictator games, see Table 11.1.

na in boldface)

ction $u_P(x, y)$	
in $(\sqrt{x}, \sqrt{y})$	$\frac{3}{5}\sqrt{x} + \frac{2}{5}\sqrt{y}$
0.00	1.90
1.00	2.20
1.41	2.26
1.73	**2.28**
2.00	2.27
2.24	2.24
2.00	2.18
1.73	2.10
1.41	1.98
1.00	1.80
0.00	1.26

When the payoff is ($0, $0) all kinds of agents receive zero utility. Egoists and enviers prefer the outcome where they get all the money. Utilitarians and Rawlsians prefer outcomes where the dollar amount is split evenly. Finally, an altruist who gives a little more weight to her own private utility than does a utilitarian might prefer the outcome ($7, $3) to all others.

It goes without saying that certain kinds of social preference would go a long way toward explaining proposers' behavior in the ultimatum and dictator games. Altruists, Rawlsians, and utilitarians actually *prefer* more equal outcomes. For such agents, there is nothing mysterious about the fact that they voluntarily offer non-zero amounts to responders.

Exercise 11.3 Altruism and the ultimatum game Imagine the ultimatum game from Figure 11.1 played by two utilitarians with $u(x,y) = \sqrt{x} + \sqrt{y}$. Find the unique subgame-perfect equilibrium in this game.

One important insight underscored by the literature on social preferences is that the game people are playing depends on their utility functions. The following exercise illustrates how agents with different utility functions end up playing very different games, even when their interactions superficially might look identical.

Exercise 11.4 Social preferences and the prisoners' dilemma Find the Nash equilibria in pure strategies in Table 11.2, when played by:
(a) Two egoists, for whom $u(x,y) = \sqrt{x}$.
(b) Two utilitarians, for whom $u(x,y) = \sqrt{x} + \sqrt{y}$.
(c) Two enviers, for whom $u(x,y) = \sqrt{x} - \sqrt{y}$.
(d) Two Rawlsians, for which $u(x,y) =$ min $(\sqrt{x}, \sqrt{y})$.

Notice that this game (in dollar terms) has the payoff structure of the prisoners' dilemma (Table 10.5 on page 225).

Table 11.2 Prisoners' dilemma (in dollar terms)

	C	D
C	$16,$16	$0,$25
D	$25,$0	$9,$9

Social preferences are fascinating and important. Economists who do not allow for the possibility of social preferences run the risk of committing terrible mistakes, whether they are trying to explain or predict behavior or to design optimal incentives. If we are mistaken about the players' utility function, we will not even know what game they are playing. Consequently, our analysis of their interaction is likely to fail. Consider the game in Table 11.2. Superficially, when expressed in dollar terms, this looks like a prisoners' dilemma. But as Exercise 11.4 showed, the players may in fact be playing a very different game.

Notice, however, that the entire analysis in this section can be completed without departing from neoclassical orthodoxy. As we know from Sections 1.1 and 2.6, the standard approach makes no assumptions about the nature of people's preferences. Consequently, it makes no assumptions about what can enter as an argument in people's utility function. It is sometimes argued that the results from the dictator and ultimatum games refute the "selfishness axiom" of neoclassical economics. But this charge is misguided: not only is there no such axiom in the calculus, but selfishness is not even entailed by the theory. Hence, there is nothing specifically behavioral about models of social preferences; if anything, the analysis shows the strength and power of the neoclassical framework.

11.3 Intentions, reciprocity, and trust

There is, however, something awkward about the account offered in the previous section. In order to accommodate the behavior of the proposer in the dictator game, we postulate that proposers are largely altruistic. But in order to accommodate the behavior of responders in the ultimatum game, this approach is inadequate. As you can tell from the third column of Table 11.1, an altruist with utility function $u(x,y) = \sqrt{x} + \sqrt{y}$ would prefer any outcome to ($0, $0). In a subgame-perfect equilibrium, therefore, an altruistic responder would accept all offers (see Exercise 11.3). But this is inconsistent with the observation that low offers frequently are rejected. Of all the agents described in the table, only the enviers prefer ($0, $0) to a sharply unfavorable division like ($8, $2). And it would be inconsistent to postulate that people are simultaneously altruistic (to explain their behavior in the dictator game) and envious (to accommodate their behavior in the ultimatum game).

There are other awkward results. In a variation of the ultimatum game, responders were found to reject the uneven division ($8, $2) if the proposer had the choice between ($8, $2) and the even division ($5, $5), but accept it if the proposer had the choice between ($8, $2) and the maximally uneven division ($10, $0). This makes no sense from the point of view of a responder who evaluates final outcomes in accordance with either one of the social preference functions in the previous section. According to each of those models, either ($8, $2) is better than ($0, $0) or it is not; the choices available to the proposer do not matter at all.

To some analysts, these results suggest that responders do not base their decisions on the final outcome of the game alone, but are (at least in part) responsive to what they see as the proposer's **intentions**. In this view, people are willing to reward people who are perceived as having good intentions and to punish people who are perceived as having bad intentions. A proposer who offers $2 rather than $0 is interpreted as having good intentions, even if the resulting allocation is uneven, whereas a proposer who offers $2 rather than $5 is interpreted as having bad intentions. Sometimes, these results are discussed in terms of **reciprocity** or **reciprocal altruism**. Respondents are said to exhibit **positive reciprocity** when they reward players with good intentions and **negative reciprocity** when they punish proposers with bad intentions. Thus, a responder in the ultimatum game who rejects a small positive offer from the proposer is said to exhibit negative reciprocity.

Reciprocity is often invoked in discussions of the **trust game**. This game is played by two players: a **sender** (Player I) and a **receiver** (Player II). At the

ility function. It is some-
ultimatum games refute
. But this charge is mis-
lculus, but selfishness is
thing specifically behav-
g, the analysis shows the
k.

trust

account offered in the pre-
vior of the proposer in the
rgely altruistic. But in order
. the ultimatum game, this
third column of Table 11.1,
would prefer any outcome
efore, an altruistic responder
this is inconsistent with the
d. Of all the agents described
sharply unfavorable division
ulate that people are simulta-
e dictator game) and envious
m game).
ation of the ultimatum game,
ion ($8, $2) if the proposer had
($5, $5), but accept it if the pro-
aximally uneven division ($10,
of a responder who evaluates
he social preference functions in
models, either ($8, $2) is better
he proposer do not matter at all.
at responders do not base their
alone, but are (at least in part)
intentions. In this view, people
ived as having good intentions
ving bad intentions. A proposer
having good intentions, even if
a proposer who offers $2 rather
ns. Sometimes, these results are
l altruism. Respondents are said
ard players with good intentions
proposers with bad intentions.
who rejects a small positive offer
reciprocity.
ons of the **trust game**. This game
) and a **receiver** (Player II). At the

outset, both are awarded some initial amount, let us suppose $10. In the first stage, the sender sends some share $x of his $10 to the receiver. The amount is sometimes called an "investment." Before the investment is received by the receiver, it is multiplied by some factor, let us say three. Thus, the receiver receives $3x. In the second stage, the receiver returns to the sender some share $y of her total allocation $10 + $3x. The final outcome, then, is ($10 – $x + $y, $10 + $3x – $y). The game is called the trust game for the obvious reason that the sender might **trust** the receiver to return some of her gains to him. When both agents maximize $u(x) = x$, there is only one subgame-perfect equilibrium in this game. Since the receiver maximizes by keeping all her money, y will equal zero. Notice that the receiver is in effect playing a dictator game with the sender as a beneficiary. Given that none of the share $x will be returned to him, the sender will keep all his money and x will equal zero. Notice that the resulting allocation ($10, $10) is Pareto inferior to many other attainable allocations. If $x = 10$ and $y = 20$, for example, the final outcome is ($20, $20).

Example 11.5 Investment decisions Suppose that you have the opportunity to invest in a promising business venture, but that you have no way to recover your expenses in case your business partner turns out to be unreliable. In order to captures the gains from trade, you simply must trust the partner to do the job. If so, you and your business partner are playing a trust game against each other. If you play the subgame-perfect equilibrium strategy, you will never invest. But if you never invest, you will never capture any of the available surplus.

Experimental economists have found that senders in the (one-shot, anonymous) trust game on the mean send about half of their initial allocation, and that receivers return a little less than what was invested. Given the figures from the previous paragraph, we should expect senders to send about $5 and receivers to return somewhere between $4 and $5. Thus, Player II succeeds in capturing some, but not all, of the available surplus. (There is a great deal of variability across studies, however.)

Why would a responder care to return some of the sender's investment when the latter has no way to penalize a receiver who returns nothing? According to one frequent answer, the receiver feels like she must reciprocate the sender's investment. Thus, a receiver who returns some of the sender's investment is said to exhibit positive reciprocity. The receiver's behavior is also consistent with altruism and inequality aversion. Meanwhile, the sender's behavior is thought to reflect the expectation that his investment will be repaid in combination with some degree of altruism.

A similar analysis obtains in the case of prisoners' dilemma and **public-goods games**. We know the prisoners' dilemma from Section 10.2. In a typical public-goods game, there are n players. For purposes of this discussion, let us assume that there are three players. Each is given an initial allocation, say $10. The game has only one stage, in which all players move simultaneously. Each player has the option of transferring a share of their initial allocation to a public account. The money in the public account is multiplied by some factor between one and three, say two, and split evenly between the players.

Given the nature of the game, the Pareto optimal outcome results when all players transfer all their money to the public account. In this case, the payoff is ($20, $20, $20). But the Pareto optimal outcome is not a Nash equilibrium, for each player can improve his outcome by transferring less. Indeed, there is only one Nash equilibrium in this game, and it is when nobody transfers any money to the public account and the outcome is ($10, $10, $10). Public-goods games therefore have certain structural similarities with the prisoners' dilemma.

The nature of this interaction should be familiar. Perhaps a set of roommates all prefer the state in which all assist in washing the dishes to the state in which nobody washes the dishes, but no matter what the others do, each roommate prefers not to wash any dishes. Thus, nobody washes the dishes and all suffer the Pareto-inferior outcome. Or, all members in a neighborhood association prefer the state in which all members spend one day a year cleaning up the neighborhood to the state in which nobody spends any time cleaning up the neighborhood, but no matter what the others do, each member prefers not to clean the neighborhood. Thus, nobody cleans the neighborhood and all suffer the Pareto-inferior outcome.

Yet, in experimental studies, cooperation remains a remarkably stubborn phenomenon. In the prisoners' dilemma, the fraction of people playing the cooperative strategy is not 100 percent. But neither is it zero, even when people are playing the one-shot game anonymously. And in anonymous, one-shot public-goods games, Robyn M. Dawes and Richard H. Thaler report:

> While not everyone contributes, there is a substantial number of contributors, and the public good is typically provided at 40–60 percent of the optimal quantity. That is, on average, the subjects contribute 40–60 percent of their stake to the public good. In [one study], these results held in many conditions: for subjects playing the game for the first time, or after a previous experience; for subjects who believed they were playing in groups of 4 or 80; and for subjects playing for a range of monetary stakes.

Why do people cooperate in one-shot prisoners'-dilemma and public-goods games? The experimental results are consistent with a high level of trust in the other players and with a desire to reciprocate what they expect to be generous contributions from the others. Contrary to predictions in our discussion about cheap talk (see Section 10.2), pre-play communication actually increases cooperation in prisoners' dilemmas and contributions in public goods games. In this sense, talk – even when cheap – might serve to promote reciprocity. Other explanations are consistent with the experimental data. Players might, for example, be altruistic as well. It should be noted that when the game is repeated, the level of contributions tends to decline. Thus, repetition appears to bring the players in closer accord with subgame-perfect-equilibrium predictions.

Predictions based on egoistic utility functions in combination with game theoretic equilibrium concepts suggest that people will be unable to coordinate their actions even when it is in their interest to do so. Yet, this is a needlessly pessimistic vision of human nature. The economist Elinor Ostrom won the

utcome results when all
. In this case, the payoff
not a Nash equilibrium,
rring less. Indeed, there
when nobody transfers
s ($10, $10, $10). Public-
rities with the prisoners'

r. Perhaps a set of room-
ing the dishes to the state
what the others do, each
obody washes the dishes
embers in a neighborhood
end one day a year clean-
dy spends any time clean-
others do, each member
ly cleans the neighborhood

ins a remarkably stubborn
tion of people playing the
r is it zero, even when peo-
nd in anonymous, one-shot
d H. Thaler report:

ial number of contributors, and
nt of the optimal quantity. That
ent of their stake to the public
conditions: for subjects playing
rience; for subjects who believed
jects playing for a range of mon-

s'-dilemma and public-goods
with a high level of trust in the
what they expect to be gener-
predictions in our discussion
nmunication actually increases
utions in public goods games.
t serve to promote reciprocity.
erimental data. Players might,
e noted that when the game is
ecline. Thus, repetition appears
subgame-perfect-equilibrium

ions in combination with game
ople will be unable to coordinate
to do so. Yet, this is a needlessly
nomist Elinor Ostrom won the

2009 Nobel Prize for exploring ways in which people develop sophisticated mechanisms that allow them to reach beneficial outcomes in trust and public-goods style games. There is plenty of evidence from the field and from the lab suggesting that people do succeed in coordinating their behavior under a wide range of conditions. Some roommates do succeed in developing mutually acceptable arrangements to make sure the dishes get washed, and some neighborhood associations do succeed in getting their members to participate in neighborhood-cleanup operations. Bad social and political philosophy, and bad social and political institutions, might result from the false assumption that cooperation cannot emerge spontaneously.

Like social preferences, it is quite possible that a story about intentions, trust, and reciprocity can be incorporated into the traditional neoclassical model. There appears to be no principled reason why it cannot be done, and some game theorists have tried. Because the specifications are a little more complicated than those in the previous section, they have been left out of this discussion. That said, it might well be possible to fit the analysis of intentions, reciprocity, and trust into the traditional neoclassical framework. Or it might not.

11.4 Limited strategic thinking

John Maynard Keynes, one of the most influential economists of the twentieth century, drew an analogy between investing in the stock market and participating in a certain kind of newspaper contest. In *The General Theory of Employment, Interest and Money*, Keynes wrote:

> [Professional] investment may be likened to those newspaper competitions in which the competitors have to pick out the six prettiest faces from a hundred photographs, the prize being awarded to the competitor whose choice most nearly corresponds to the average preferences of the competitors as a whole; so that each competitor has to pick, not those faces which he himself finds prettiest, but those which he thinks likeliest to catch the fancy of the other competitors, all of whom are looking at the problem from the same point of view. It is not a case of choosing those which, to the best of one's judgment, are really the prettiest, nor even those which average opinion genuinely thinks the prettiest.

The basic structure of this strategic interaction is captured by the **beauty-contest game**. Here, n players simultaneously pick a number between zero and 100 inclusive. The person whose number is closer to seven-tenths of the average number wins a fixed prize. (The fraction does not have to be seven-tenths, but it must be common knowledge among the players.)

There is only one Nash equilibrium in this game. In this equilibrium, every player picks the number zero and everyone ties for first place. Suppose everyone picked 100, the highest number available. If so, the winning number would be 70. Thus, no rational person would ever pick a number greater than 70. But if no one picks a number greater than 70, the winning number cannot be greater than 49. Yet, if nobody picks a number greater than 49, the winning number cannot be greater than about 34. And so on, all the way down to zero.

Real people, however, do not play the Nash equilibrium strategy in the one-shot game. When played for the first time, answers might fall in the 20–40 range. Interesting things happen when the game is repeated with feedback about the average number in the previous round. During subsequent rounds, the average number will decrease and eventually approach zero. This suggests that, over time, real people converge to the Nash equilibrium prediction.

The favored explanation of the result in the one-shot game is based on the idea that people have different degrees of sophistication. "Level-0" players just pick a number between zero and 100 randomly. "Level-1" players believe that all the other players are level-0 players. Level-1 players, therefore, predict that the mean number will be 50, and therefore pick $0.7 * 50 = 35$. Level-2 players believe that all other players are level-1 players and that the average will be 35, and therefore pick $0.7 * 35 \approx 25$, and so on. Using statistical techniques, behavioral game theorists can estimate what proportion of each sample is level 0, level 1, and so on. The results suggest that most humans are level-1 or level-2 players.

One fascinating feature of this game is that, even if you know the unique Nash equilibrium strategy, you may not want to play it. As long as you expect other players to play out-of-equilibrium strategies and choose a positive number, you will want to do so too. And as long as other players expect you to expect them to pick a positive number, they will want to pick a positive number. And so on. Thus, everyone plays a positive number and an out-of-equilibrium strategy. But, although you want to pick a number greater than zero, the number must not be too high: the aim is to stay one step ahead of the other players.

In keeping with Keynes's analogy, it has been suggested that this kind of game captures the dynamics of real markets and can explain bubbles in stock and real-estate markets. Even if all investors know that the market will ultimately crash, and that the unique Nash equilibrium strategy is to exit the market, they might assume that others will continue to buy for just a little while longer. As long as individual investors think that they can stay one step ahead of the competition and exit the market just before everybody else does, they will want to continue to buy. In the process, of course, they will drive prices even higher.

Example 11.6 Rock-paper-scissors, cont. The game rock-paper-scissors has a unique Nash equilibrium, in which both players randomize with probability 1/3, 1/3, and 1/3, and in which both players have an equal probability of winning (see Exercise 10.13 on page 231). So it might surprise you to hear that there is a World Rock Paper Scissors Society and a World Championship competition. According to the Society website, rock-paper-scissors is a game of skill, not chance: "Humans, try as they might, are terrible at trying to be random [and] in trying to approximate randomness become quite predictable."

Here is one piece of advice. The pros will tell you that "rock is for rookies," because inexperienced males tend to open with rock. Thus, if your opponent is one, open with paper. If your opponent is slightly more sophisticated and may be thinking you will open with rock and therefore opens with paper, open with scissors. If you are playing against an even more sophisticated agent, who will for the reason just identified open with scissors, open with rock.

Here is another piece of advice. Inexperienced players will not expect you to call your throw ahead of time. Thus, if you announce that you are throwing rock next, an inexperienced opponent will assume that you will not and instead will choose something other than paper. So you will want to throw rock. If you play against a slightly more sophisticated agent, though, they will expect you to throw rock after announcing that you will throw rock; so what you need to do is to throw scissors.

As in the beauty-contest game, the goal is to stay exactly one step ahead of your opponent. A similar analysis might apply to people's performance in the centipede game (see Exercise 10.26 on page 237). The typical finding is that people Pass until a few stages from the end, when they Take. This outcome would be expected if neither player thinks the other will play the unique subgame-perfect equilibrium, and that both attempt to Take one stage before the other one does.

Unlike models of social preferences, in this case there is little hope of capturing observed behavior in the one-shot game within the traditional neoclassical model. Things are quite different in the repeated version of the game. As the same group plays the game again and again, they approximate equilibrium predictions.

11.5 **Discussion**

According to the great Austrian economist Friedrich A. Hayek, the existence of spontaneous coordination constitutes the central problem of economic science. Hayek wrote:

> From the time of Hume and Adam Smith, the effect of every attempt to understand economic phenomena – that is to say, of every theoretical analysis – has been to show that, in large part, the co-ordination of individual efforts in society is not the product of deliberate planning, but has been brought about, and in many cases could only have been brought about, by means which nobody wanted or understood.

In this view, economists have never seriously doubted *that* coordination takes place; the question is *how* it emerges and is sustained. Much of the behavioral game theory literature on social preferences, trust, and reciprocity was developed in large part to answer this question. Obviously, this chapter does not contain a complete account of strategic interaction that fails to fit the picture painted by analytical game theory, or of the models behavioral game theorists have offered to capture the way in which people really interact with each other.

To what extent is the work presented under the heading "behavioral game theory" compatible with the traditional neoclassical framework? As we have seen, much of what goes under the heading is in fact consistent with analytical game theory. Models of social preferences are clearly consistent, since they proceed by allowing a person P's utility function to reflect another person Q's attainment. The degree to which ideas of intentions, reciprocity, and trust can be incorporated into neoclassical theory remains unclear, though game theorists have tried. Models that try to capture people's limited ability to think strategically, by contrast, are more clearly inconsistent with any model that relies on Nash or subgame-perfect equilibrium concepts.

In defense of analytical game theory, it has been argued that neoclassical theory is only intended to apply under sharply circumscribed conditions. Thus, the prominent game theorist Ken Binmore writes:

> [Neoclassical] economic theory should only be expected to predict in the laboratory if the following three criteria are satisfied:
>
> - The problem the subjects face is not only "reasonably" simple in itself, but is framed so it seems simple to the subjects;
> - The incentives provided are "adequate";
> - The time allowed for trial-and-error adjustment is "sufficient".

Binmore recognizes that he is also denying the predictive power of neoclassical economics in the field, and adds: "But have we not got ourselves into enough trouble already by claiming vastly more than we can deliver?" Binmore's view is nicely illustrated by the beauty-contest game (see Section 11.4). Here, real people's behavior differs dramatically from Nash equilibrium predictions during the first round, but converges to it as the game is repeated and players learn the outcome of previous rounds. The same thing might be true for public-goods and other games. Binmore's defense of analytical game theory does not constitute an argument against behavioral game theory, however, provided behavioral game theory is intended to apply when the three conditions are not satisfied.

ADDITIONAL EXERCISES

Exercise 11.7 Gandhi A leader of India's independence movement and advocate of non-violent social change, Mahatma Gandhi is supposed to have said: "The true measure of any society can be found in how it treats its most vulnerable members." If we interpret this line as an expression of social preferences, what kind of preferences would that be?

Problem 11.8 Equilibrium concepts *We know from Section 10.4 that some game theorists think the Nash equilibrium concept is problematic, and that the concept of subgame-perfect equilibrium better captures the behavior of rational agents. But as the single subgame-perfect equilibrium in the ultimatum game suggests (see Section 11.2), there is something funny about subgame-perfect equilibrium too. For one thing, the subgame-perfect equilibrium requires the responder to accept an offer of $0. And in most Nash equilibria, the responder does better than this. In your view, which equilibrium concept offers the best prediction of the behavior of rational agents: Nash or subgame-perfect equilibrium?*

reflect another person
ıs, reciprocity, and trust
unclear, though game
ꞩ limited ability to think
ıt with any model that
ꞁts.

argued that neoclassi-
:cumscribed conditions.
ɜs:

to predict in the laboratory

ɩly" simple in itself, but is

ɹfficient".

ive power of neoclassical
ɔt ourselves into enough
ɫeliver?" Binmore's view
Section 11.4). Here, real
equilibrium predictions
ne is repeated and play-
ɜ thing might be true for
ıf analytical game theory
ɪ game theory, however,
ɩly when the three condi-

ndence movement and
Gandhi is supposed to
e found in how it treats
is line as an expression
ould that be?

*from Section 10.4 that
ept is problematic, and
ter captures the behav-
rfect equilibrium in the
something funny about
ubgame-perfect equilib-
And in most Nash equi-
view, which equilibrium
ational agents: Nash or*

FURTHER READING

The best-known and most thorough treatment of behavioral game theory is Camerer (2003), which includes the quotation summarizing results from the ultimatum game (p. 49). Kagel and Roth (1995) contains thorough treatments of experimental methods and results; Durlauf and Blume (2010) offers more concise and up-to-date treatments. A good review of cooperation in laboratory experiments is Dawes and Thaler (1988), which among other things summarizes the results of public-goods experiments (p. 189). The World RPS Society (2011) will tell you how to beat anyone at rock-paper-scissors. The two historical quotations are from Keynes (1936, p. 156) and Hayek (1933, p. 129); for more on Hayek's take on information and cooperation, see Angner (2007). The Binmore quotation is from Binmore (1999, p. F17).