3-2009

Rationality and Indeterminacy

Cristina Bicchieri
University of Pennsylvania, cb36@sas.upenn.edu

Rationality and Indeterminacy

Abstract

Much of the history of game theory has been dominated by the problem of indeterminacy. The very search for better versions of rationality, as well as the long list of attempts to refine Nash equilibrium, can be seen as answers to the indeterminacy that has accompanied game theory through its history. More recently, the experimental approach to game theory has attempted a more radical solution: by directly generating a stream of behavioral observations, one hopes that behavioral hypotheses will be sharper, and predictions more accurate. This article looks at several attempts to address indeterminacy, including the shift to evolutionary models. However, because its goal is to establish whether rational choice models are inescapably doomed to produce indeterminate outcomes, it pays much more attention to the experimental turn in game theory, the difficulty it encounters, and the promising results obtained by more realistic models of rationality that include a social component.

Disciplines

Behavioral Economics | Epistemology | Philosophy of Science | Social Psychology

PART II

MICROECONOMICS

CHAPTER 6

RATIONALITY AND INDETERMINACY

CRISTINA BICCHIERI

1. Indeterminacy

Much of the history of game theory has been dominated by the problem of indeterminacy. The very search for better, more encompassing versions of rationality, as well as the long list of attempts to refine Nash equilibrium, can be seen as answers, or attempted solutions, to the indeterminacy that has accompanied game theory through its history. More recently, the experimental approach to game theory has attempted a more radical solution: by directly generating a stream of behavioral observations, and thus controlling some crucial parameters, one hopes that behavioral hypotheses will be sharper, and predictions more accurate. I shall look at several attempts to address indeterminacy, including the shift to evolutionary models. However, because my goal is to establish whether rational choice models are inescapably doomed to produce indeterminate outcomes, I will pay much more attention to the experimental turn in game theory, the difficulty it encounters, and the promising results obtained by more realistic models of rationality that include a social component. The sophisticated reader should bear with some initial review of familiar ideas for the sake of following the historical (and logical) thread, from early attempts to address indeterminacy out to novel ideas and solutions.

There are at least two kinds of indeterminacy we may want to distinguish. One, which I will dub *epistemic indeterminacy,* is something we all have to live with. Our knowledge of the world is limited, and the outcomes of our choices

usually are not deterministic; instead, any choice corresponds to several possible outcomes, each tied to the occurrence of a particular state of the world. Though we cannot predict which outcome will occur, we can assess the probability with which the corresponding state of the world will occur and can choose on the basis of this probabilistic assessment. Rational choice in this context simply means maximizing expected utility, where the subjective utility of an act is the weighted sum of the desirability of its consequences, and the weights are the probabilities we assess for each of the possible consequences. Decision theory formalizes all this: It tells us which strategies are rational, that is, coherent with the subject's preferences, with respect to certain and uncertain outcomes, and with her beliefs about all the variables that are relevant to choice but that she cannot control.

Though we live with epistemic indeterminacy, some form of *predictability* in the context of individual decision making is still possible. By predictability I mean the ability of a third party to predict what sort of action an individual will take. For example, suppose there are reliable statistical data available and we know our subject knows them, we know his preferences, and have every reason to believe he is both practically and epistemically rational.[1] In this case, we can in fact predict what this person will choose. The prediction becomes a little more complicated in case there are no objective probabilities to rely on, but suppose again that we happen to know a person's subjective probabilistic assessments. In this case, provided again that we know that person's preferences and have every reason to believe such individual is both practically and epistemically rational, we can predict his choice.[2]

This simple model of rational choice has been criticized as too abstract and demanding: on one end, the amount of knowledge required for third-party predictability is quite extreme, and on the other end, the decision maker is often unable to even imagine all the possible consequences of her actions, calculate the probabilities, maximize as the rationality recipe recommends, and so on. What I want to stress here, however, is not our obvious condition as cognitive misers. Instead, I want to draw attention to the fact that, even if we were perfect cognitive machines and/or perfect predictors, there are contexts in which rationality may not help us make a choice or predict what another's choice will be. Such contexts are very common; whenever we interact with other individuals and the outcome of such interaction depends upon the joint actions of all the parties, we face *strategic indeterminacy*.

To act rationally in a strategic context is much more difficult because the consequences of an action depend upon what *all* the parties involved do. That is, outcomes are jointly determined by the parties' independent actions. The interactive decision problem of an agent can be represented in general terms as follows: The agent will choose a plan of action considering that the consequences of his/her choice also depend on a combination of unknown and uncontrollable variables, including other agents' plans of action. Rephrasing the problem in terms of decision theory, we may say that the agent is *rational* if she maximizes her expected

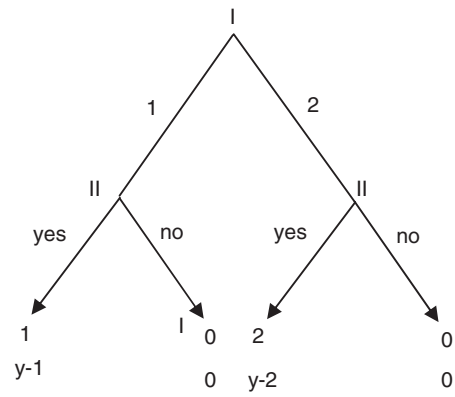

Figure 6.1.

utility, calculated by assigning subjective probabilities to the possible values of all the relevant variables she cannot control, and taking into account her present information. The fundamental difference between decision theory and game theory is that the latter tries to *explain subjective beliefs using strategic reasoning*.

For example, think of a two-player bargaining game in which player I moves first, and can sell to player II a good for a low price ($1) or a high price ($2). Player II can accept (yes) or reject (no) I's proposal; if II rejects, both get 0 (Figure 6.1). I's problem is one of maximizing expected utility (in this case, expressed in terms of money), given what he knows (or believes) about II's knowledge, preferences, and beliefs.

Note that how much information is possessed by the players is crucial in determining the outcome, and our knowledge of such information is equally critical in determining the possibility of predicting what the players will do. Suppose that the rules of the game and players' rationality are *common knowledge*.[3] If II is known to be rational, then I can *infer* that II will refuse if the price is greater than *y*, and accept if the price is lower than *y*. However, without knowing the value of *y*, player I would still be unable to predict II's response to his offer. If we also assume that payoffs are common knowledge, then I would know the value of *y*, *predict* how II will react to each offer, and choose accordingly.[4] In our case, if I knows that $y > 2$, then he expects II to accept both prices. If instead I knows that $1 < y < 2$, then he expects II to accept only the low price.

Notice that I's subjective beliefs about player II's choices are *endogenous*. He infers them from his knowledge of the structure of the game, of player II's payoffs and rationality. In sum, in a strategic context, what is rational for a player to do depends upon what he expects other players will do, which in turn is inferred from the knowledge that a player has about other players and the structure of the game. *Each party to a strategic interaction is at the same time a chooser and a predictor of other parties' choices*. In this case, epistemic and strategic indeterminacy are entwined.

2. Nash Predictions

An easy solution to the problem of strategic indeterminacy would be to collapse it into epistemic indeterminacy: if players' subjective beliefs were treated as *exogenous* to the game, other players' choice profiles would become states of the world that have a given probability of occurring, and one might just try to maximize expected utility with respect to those states, bypassing the fact that such states are determined by what other players believe their counterparts are going to choose. Is this an interesting solution to strategic indeterminacy? Not if we think that the notion of *Nash equilibrium* is important.

Nash equilibrium (Nash 1951) is the standard solution concept for noncooperative games. Informally, a Nash equilibrium specifies players' actions and beliefs such that (a) each player's action is optimal given his beliefs about other players' choices; (b) players' beliefs are correct. Thus, an outcome that is not a Nash equilibrium requires either that a player chooses a suboptimal strategy, or that some players misperceive the situation. More formally, a Nash equilibrium is a vector of strategies $(\sigma_1^*,\ldots,\sigma_n^*)$, one for each of the n players in the game, such that each σ_i^* is optimal given (or is a *best reply* to) σ_{-i}^*.[5]

Nash equilibrium is an appealing solution concept for noncooperative games for several reasons.[6] It captures an important feature of individual rationality, that is, that being rational means maximizing one's expected utility under the constraint represented by what one expects other individuals to choose. It is supported by correct beliefs, in the sense that, if players are in equilibrium, their beliefs about each other's strategy choice are correct. Finally, the concept of Nash equilibrium depicts the idea of a self-enforcing agreement. Were players to agree in preplay negotiation to play a particular strategy combination, they would have an incentive to stick to the agreement only in case the agreed upon combination is a Nash equilibrium. There are many real-life situations in which there is no third party available to monitor and enforce compliance with an agreement: many transactions are conducted with a handshake in the expectation that the parties will fulfill their promises. Indeed, when this happens it means that it is in the parties' interest to fulfill the terms of the agreement, that there is no incentive to unilaterally deviate from it. It means, in other words, that the agreement is a Nash equilibrium.

Game theorists typically assign a predictive value to Nash equilibrium. In a well-known passage of their book, *Theory of Games and Economic Behavior* (1944), von Neumann and Morgenstern said that rational players who know (i) all there is to know about the structure of the game they are playing, (ii) all there is to know about the beliefs and motives of the other players, (iii) that every player is rational, (iv) that every player knows (i)–(iii), (v) that every players knows (i)–(iv), and so on, will be able to infer the optimal strategy for every player. In that case, each player will behave rationally by maximizing his expected utility conditional on what he expects the others to do. This states what could be rightly called the

central dogma of game theory: that rational players will always jointly maximize their expected utilities, or play a *Nash equilibrium*.

Ken Binmore (1987/1988) has argued that there are two possible interpretations of Nash equilibrium. According to the *evolutive* interpretation, a Nash equilibrium is an observed regularity. Players know the equilibrium, and test the rationality of their behavior given this knowledge acquired from experience. The players (and the game theorist) can accordingly predict that a given equilibrium will be played, since they are accustomed to coordinate upon that equilibrium and expect (correctly) others to do the same. According to the more commonly adopted *eductive* interpretation, instead, a game is a unique event. In this case it makes sense to ask whether players can deduce what others will do from the information available to them. The players (and the game theorist) can predict that an equilibrium will be played just in case they have enough information to infer players' choices. The standard assumptions game theorists make about players' rationality and knowledge should in principle be sufficient to guarantee that an equilibrium will obtain. The customary assumptions are:

C(SG + PF). The structure of the game (SG) and players' preferences (PF) are common knowledge among the players;
C(Rat). The players are rational (Rat) and this is common knowledge.

When a game has a unique Nash equilibrium, we can predict that it will be played if we are able to show that players, armed with common knowledge of rationality and of the structure of the game, will infer the Nash solution. If players have dominated strategies, C(Rat) entails that they will eliminate them, and this is common knowledge (we assume that the consequences of C(SG + PF) and C(Rat) are common knowledge, too). Often, after we have eliminated strictly dominated strategies for one player, we may find that there are now strictly dominated strategies for another player, which will be eliminated as well. This process of successive elimination can continue until there are no more strictly dominated strategies left. If a unique strategy remains for each player, we say the game has been solved by *iterated dominance*. It is easy to prove that a strategy profile thus obtained is a Nash equilibrium (Bicchieri 1993).

Common knowledge of rationality, preferences, and strategies may thus facilitate the task of predicting an opponent's strategy but, as I argued elsewhere (Bicchieri 1993), it does not guarantee that the resulting prediction will be correct. This is because the concept of Nash equilibrium embodies a notion of individual rationality, since each player's equilibrium strategy is a best reply to the opponents' strategies, but, unfortunately, it does not specify how players come to form the beliefs about each other's strategies that support equilibrium play. Beliefs, that is, can be internally consistent but fail to achieve the interpersonal consistency that guarantees that an equilibrium will be attained. Bernheim (1984) and Pearce (1984) have argued that assuming players' rationality (and common knowledge thereof) can only guarantee that a strategy will be *rationalizable,* in the sense of

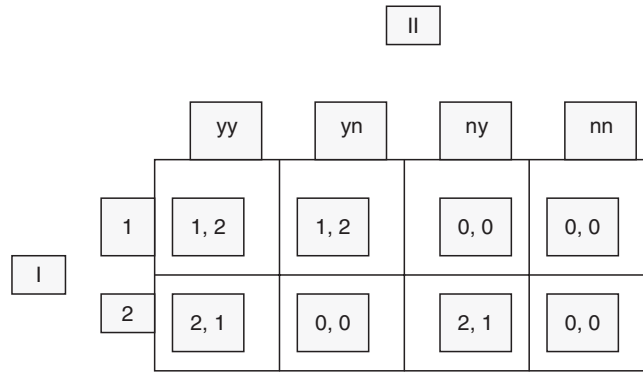

Figure 6.2.

being supported by internally consistent beliefs about other players' choices and beliefs. Yet a combination of rationalizable strategies may not constitute a Nash equilibrium. On the other hand, the fact that a Nash equilibrium is always a combination of rationalizable strategies is of no help in predicting it will be played.

Consider for a moment the normal form representation of the Figure 6.1 game:

I am assuming here that I knows II's payoffs, and that y=3. Player II has four strategies: always accept (*yy*), always reject (*nn*), accept if $1, reject if $2 (*yn*), and reject if $1 and accept if $2 (*ny*). For player II, *nn* is strictly dominated by *yy,* so C (Rat) allows I to exclude it (and II knows it). However, all three other strategies of player II are rationalizable, as well as the two strategies of player I. Moreover, there are three pure strategy Nash equilibria of this game: (2,*yy*), (1,*yn*) and (2, *ny*), and no way to predict, by C(Rat) and C(SG+PF), which of them, if any, will be played.

The *refinements of Nash equilibrium* program are precisely attempts to eliminate some equilibria as being unreasonable. If we consider again the extensive form game of Figure 6.1, we can imagine player II threatening to reject if I asks a high price. This threat is embedded in the equilibrium (1,*yn*), but is it a credible threat? Player I knows that y=3 and that II is rational, hence I knows that, faced with a choice between 0 and 1, II will always choose 1, that is, will always accept the high price. Thus the (1,*yn*) equilibrium should be eliminated as unreasonable. In this case, we have applied the simplest refinement of *subgame perfection*. Briefly, a Nash equilibrium s^* is perfect if it remains an equilibrium in every proper subgame of the original game G. In our simple example, we can calculate the final result by backward induction: We take the subgames starting at II's decision node and look at the optimal choice for player II. If y=3, the optimal choice for II is to accept at both subgames. Knowing that, player I will always choose to set a high price. Another way to look at the same problem is to perturb the game by assuming that every strategy has a very small probability of being chosen (Selten 1975). So in

the equilibrium (1,*yn*) player II is indifferent between *yy* and *yn* simply because II is certain that I will choose 1, the low price. However, if there is a small probability that I, by mistake, chooses the high price, then *yy* gives II a higher utility than *yn*.

The problem with the refinements literature is that it lacks a coherent interpretation of deviations from equilibrium play (Bicchieri 1988, 1993). A deviation may be a mistake, but it may also be a signal. There is no general model of belief revision that would include different refinements as special cases of some general, substantive criterion of belief change. Moreover, in order to predict a specific equilibrium outcome, it must be assumed that players share (and have common knowledge that they share) the same reasonable principles. Multiplicity of equilibria only aggravates a problem already present in cases in which the equilibrium is unique: the equilibrium depends on parameters that are not known to an external observer trying to predict the outcome of the game. Note that, in this case, the players themselves are external observers of their interactive environment who have to guess what their opponents will do, which in turn depends upon the opponents' expectations about other players' choices, and so on.

3. Evolution and Learning

One way to solve strategic indeterminacy is to think in terms of evolution. Evolutionary models describe aggregate dynamics, without explaining in great detail how such dynamics are generated by individual behaviors. In fact, individuals in such models are often represented as strategy bearers whose choices are fixed or, when they have the possibility of shifting strategies, such shifts are not determined by what they expect others to do but rather by how well they have done in the past. In such models, a Nash equilibrium is no longer interpreted as a unique event; it is instead conceived as an observed regularity, about which we want to know how it was reached and what accounts for its stability. When multiple equilibria are possible, we want to know why players converged to one in particular and then stayed there. In this case, the selection process is not the result of complicated, multistage reasoning; it simply results from some form of natural selection.

Evolutionary theories are inspired by population biology (e.g., Maynard Smith & Price 1973). These theories dispense with the notion of the decision maker, as well as with best responses/optimization, and use in their place a natural selection, survival-of-the-fittest process together with mutations to model the frequencies with which various strategies are represented in the population over time. In a typical evolutionary model, players are preprogrammed for certain strategies, and are randomly matched with other players in pairwise repeated encounters. The relative frequency of a strategy in a population is simply the proportion of players in that population who adopt it. The theory focuses on how the strategy profiles

	D	H
D	B/2	0
H	B	B-C/2

Figure 6.3.

of populations of such agents evolve over time, given that the outcomes of current games determine the frequency of different strategies in the future. As an example, consider the symmetric game in Figure 6.3 and suppose that there are only two possible behavioral types: hawks and doves.

A hawk always fights and escalates contests until it wins or is badly hurt. A dove sticks to displays and retreats if the opponent escalates the conflict; if it fights with another dove, they will settle the contest after a long time. Payoffs are expected changes in fitness due to the outcome of the game. Fitness here means just reproductive success (e.g., the expected number of offspring per time unit). Suppose injury has a payoff in terms of loss of fitness equal to C, and victory corresponds to a gain in fitness B. If hawk meets hawk, or dove meets dove, each has a 50% chance of victory. If a dove meets another dove, the winner gets B and the loser gets nothing, so the average increase in fitness for a dove meeting another dove is $B/2$. A dove meeting a hawk retreats, so her fitness is unchanged, whereas the hawk gets a gain in fitness B. If a hawk meets another hawk, they escalate until one wins. The winner has a fitness gain B, the loser a fitness loss C. So the average increase in fitness is $(B\text{-}C)/2$. The latter payoff is negative, since we assume the cost of injury is greater than the gain in fitness obtained by winning the contest. We assume that players will be randomly paired in repeated encounters, and in each encounter they will play the stage game of Figure 6.3.

If the population were to consist predominantly of hawks, selection would favor the few doves, since hawks would meet mostly hawks and end up fighting with an average loss in fitness of $(B\text{-}C)/2$, and $o > (B\text{-}C/2)$. In a population dominated by doves, hawks would spread, since every time they meet a dove (which would be most of the time) they would have a fitness gain of B, whereas doves on average would only get $B/2$.

Maynard Smith interpreted evolutionary games as something that goes on at the phenotypic level. The fitness of a phenotype depends on its frequency in the population. A strategy is a phenotype, and a player is just an instance of such a behavioral phenotype. In our example, we have only two behavioral phenotypes: hawks and doves. Evolutionary game theory wants to know how strategies do on average when games are played repeatedly between individuals who are randomly drawn from a large population. The average payoff to a strategy depends on the composition of the population, so a strategy may do very well (in term of fitness) in one environment and poorly in another. If the frequency of hawks in the popula-

tion is q and that of doves correspondingly $(1\text{-}q)$, the average increase in fitness for the hawks will be $q(B\text{-}C)/2+(1\text{-}q)B$, and $(1\text{-}q)B/2$ for the doves. The average payoff of a strategy in a given environment determines its future frequency in the population. Strategies that, on average, earn high payoffs in the current environment are assumed to increase in frequency, and strategies that, on average, earn lower payoffs are assumed to decrease in frequency. If the average payoffs of the different strategies are the same, then the composition of the population is stable. In our example, the average increase in fitness for the hawks will be equal to that for the doves when the frequency of hawks in the population is $q=B/C$. At that frequency, the proportion of hawks and doves is stable. If the frequency of hawks is less that B/C, then they do better than doves, and will consequently spread; if their frequency is larger than B/C, they will do worse than doves and will shrink.

Note that if $C > B$, then $(B\text{-}C)/2 < 0$, so the game in Figure 6.3 has two pure-strategy Nash equilibria: (H, D) and (D, H). There is also a mixed strategy equilibrium in which Hawk is played with probability $q = B/C$ and Dove is played with probability $(1\text{-}q) = C\text{-}B/C$. If the game of Figure 6.3 were played by rational agents who *choose* which behavior to display, we would be at a loss in predicting their choices. We know that from C(SG + PF) and C(Rat) the players cannot infer that a particular equilibrium will be played; moreover, since there are no dominated strategies, all possible outcomes are rationalizable. In our hawk/dove example, however, players are not rational and do not choose their strategies. So if an equilibrium is attained, it must be the outcome of some process very different from rational deliberation. The process at work is natural selection: High-performing strategies increase in frequency whereas low-performing strategies diminish in frequency and eventually go to zero.

The mechanism is quite simple: the bearer of a successful behavioral trait will have more offspring than the bearer of a less successful trait, and each of the descendants will display the same behavioral trait, hence the frequency increase. This is an extremely simplified and probably wrong story of how behavioral traits are transmitted among humans. We have no evidence for the genetic transmission of behavioral traits such as altruism, selfishness, or the tendency to escalate conflicts. There is instead evidence that such traits are culturally transmitted, and a realistic model of how a specific behavioral pattern becomes dominant in a population should, therefore, include a description of how individuals learn behavioral patterns and imitate those who are successful.

We have seen that in a population composed mostly of doves, hawks will thrive, and the opposite would occur in a population composed mainly of hawks. So, for example, if hawks dominate the population, a mutant displaying dove behavior can invade the population, since individuals bearing the dove trait will do better than hawks. The main solution concept used in evolutionary game theory is the *evolutionarily stable strategy* (ESS) introduced by Maynard Smith and Price (1973). A strategy or behavioral trait is evolutionarily stable if, once it dominates in the population, it does strictly better than any mutant strategy, hence it cannot

be invaded. To formalize this concept, let me first make a brief digression. In a symmetric game like hawk/dove, we have a finite set of pure strategies S and a corresponding set Δ of mixed strategies. A population state is equivalent to a mixed strategy $x \in \Delta$. Note that the evolutionary model gives a natural interpretation to mixed strategies as the proportions of certain strategies (or traits) in a population. A state in which each individual plays a pure strategy and the proportion of different strategies correspond to x is called a polymorphic state. Alternatively, we may interpret the population state x as monomorphic, in the sense that each player plays the mixed strategy x. In a two-player game, being matched against a randomly drawn individual in population state x is equivalent to being matched against an individual who plays the mixed strategy x. Hence the average payoff of playing strategy y in population state x is equal to the expected payoff to y when played against the mixed strategy x, that is, $u(y,x)$. The population average in this case is equal to the expected payoff of the mixed strategy x when matched against itself, that is, $u(x,x)$.

In a symmetric, two-player game, x is an ESS if and only if, for all $y \in \Delta$ such that $y \neq x$,

(1) $u(x,x) > u(y,x)$

or

(2) $u(x,x) = u(y,x)$, and $u(x,y) > u(y,y)$.

Condition (1) tells us that strategy x is a unique best reply against itself. If the bulk of the population consists of type x and a small number of mutants of type y enters the population, if x does better against x than y does against x, y will be less fit and disappear. However, if x is a mixed strategy, we know (1) does not hold. In this case, for x to be an ESS, (2) must hold. If both x and y perform equally well against x, then y will be less fit than x if x does better against y than y does against y.

In the hawk/dove game, neither of the two pure behavioral types is evolutionarily stable, since each can be invaded by the other. We know, however, that a population in which there is a proportion $q = B/C$ of hawks and $(1\text{-}q) = C\text{-}B/C$ of doves is stable. This means that the type of behavior that consists in escalating fights with probability $q = B/C$ cannot be invaded by any other type, hence it is an ESS. To show that the mixed strategy $x = (B/C, C\text{-}B/C)$ is an ESS, we have to show that condition (2) is satisfied. Indeed, $u(x,y) - u(y,y) = 1/2C\,(B\text{-}Cq)^2$ is greater than zero for all $q \neq B/C$.

An ESS is a strategy that, when it dominates the population, is a best reply against itself. Therefore, an evolutionarily stable strategy such as $(B/C, C\text{-}B/C)$ is a Nash equilibrium. Though every ESS is a Nash equilibrium, the reverse does not hold; in our stage game, there are three Nash equilibria, but only the mixed strategy equilibrium $(B/C, C\text{-}B/C)$ is an ESS. However, when a strategy is a *unique* best reply to itself, it is both an ESS and a *strict* Nash equilibrium. In this special

case, the reverse also holds: Every strict Nash equilibrium is an ESS. In a strict equilibrium, there exists no other strategy that is an alternative best reply to the equilibrium strategy, and this guarantees noninvadability.

The prior examples show how evolution can at least partially solve the problem of equilibrium selection without imposing heroic cognitive requirements on players. An ESS is, in fact, not just a Nash equilibrium but also a perfect and proper equilibrium (Van Damme 1987). Furthermore, an evolutionary account of how a Nash equilibrium is achieved provides an explanation of the dynamics of the selection process, something which the refinement program cannot do. In the hawk/dove example, we have assumed that the success of a strategy depends on the outcome of pair-wise random matches. It is often the case that a strategy's success depends not on the strategy played by a particular opponent, but on the population-wide frequencies of strategies. When examining behavior in a *population game,* we adopt the concept of an *evolutionarily stable state* (also ESS) (Hofbauer & Sigmund 1998).

Suppose the game has N pure strategies, with an NxN symmetric expected payoff matrix $A=(a_{ij})$. There is an infinite number of players, and each player initially commits to playing exactly one of the N pure strategies. Let p be the $N \times 1$ vector denoting the population-wide proportion of each of the N strategies (player types) in the population at a given time. Let

$$f_i(p) = \sum_j a_{ij} p_j = A_i p$$

denote the fitness of strategy i and let $\sum f_i(p) = Ap.$ denote the population-wide payoff. The population-wide weighted average fitness value is $p^T Ap$. We say that $\hat{p}$ is an *evolutionarily stable state* if for any $p \neq \hat{p}$ in the neighborhood of $\hat{p}$, we have:

$$\hat{p}^T Ap > p^T Ap$$

This captures the idea that the population-wide payoff under $\hat{p}$ is higher (locally) than for any other vector p.[7]

The definitions of evolutionarily stable strategies or states are static. To describe the dynamic process that leads to a certain distribution of strategies in a population, we have models of the selection dynamics that express the growth rate of a strategy i in population state p as a function of i's average payoff in p relative to the average payoff to other strategies in p. Evolutionarily stable state does not refer to a specific dynamic, but biologists and evolutionary game theorists frequently use deterministic *replicator* dynamics (Taylor & Jonker 1978) of the form:

$$(*)\ p_i(t+1) = \frac{p_i(t) A_i p(t)}{p^T(t) A p(t)},$$

where $p(t)$ denotes the population-wide proportions at time t, the denominator is a measure of average strategy fitness in the population at t, and the numerator measures the fitness of strategy i at time t. Strategies with above-average fitness see their proportions increase, and those with below-average fitness see their proportions decrease.[8]

ESS are asymptotically stable fixed points of this replicator dynamic, though the converse need not be true (see, e.g., Samuelson 1997). A similar relationship holds between the replicator dynamic and Nash equilibria: if $\hat{p}$ is a Nash equilibrium of the symmetric NxN game with expected payoff matrix A, then $\hat{p}$ is a stationary state of the replicator dynamic.

In evolutionary theory replication, variation and heredity are the basic assumptions. Any entity capable of replicating itself with differential success will be subject to an evolutionary process. Differential success, in turn, is related to hereditary variations. In biology, replicators are genes and in genetic evolution, variation is provided by random mutations and recombinations of gene sequences. Behavioral patterns can be replicators, too, in the sense that behavioral trait x is replicated when a gene $\boldsymbol{x}$ that predisposes its carriers to behave according to this pattern replicates itself. This means that bearers of gene $\boldsymbol{x}$ will behave in ways that make them reproductively successful, so that in the next generation there will be more copies of $\boldsymbol{x}$. To the extent that behavior x promotes the replication of its predisposing gene, we are correct in saying that the behavior is replicating itself. Individuals are just bearers of such genetic material, hence they are born with fixed behavioral traits. Variation of competing strategies is provided by random mutations and recombinations of gene sequences.

When we think of strategies, however, we usually refer to behaviors that are not genetically inherited. In economic and political applications of game theory, actors can be firms, political parties, nations. Even when actors are individuals, their strategies have a strong cultural component. Evolutionary models can still be applied to explain how Nash equilibria are attained and whether they are stable, but selection mechanisms in this case work through processes of cultural transmission such as learning and imitation. Learning and imitation are subject to mistakes, and new strategies may enter the population either by random mistake or by purposeful innovation. For example, we tend to imitate successful individuals, where success is measured in terms of some shared values. Since it is usually difficult to point to one particular behavior as responsible for successful performance, what is imitated will often consist of a set of behavioral rules, and this in turn may generate mistakes. Payoffs in this case cannot represent fitness changes, but if we give them a utility interpretation, we must provide for interpersonal comparisons of utilities. Indeed, to imitate a more successful individual, one must be able to compare one's payoffs with the payoffs of others, but traditional von Neumann-Morgenstern utilities do not allow for such comparisons.

Evolutionary games provide us with a way of explaining how agents that may or may not be rational and—if so—subject to severe information and calculation restrictions, achieve and sustain a Nash equilibrium. When there exist evolutionarily stable strategies (or states), we know which equilibrium will obtain, without the need to postulate refinements in the way players interpret off-equilibrium moves. Yet we need to know much more about processes of cultural transmission, and to develop adequate ways to represent payoffs, so that the promise of evolutionary games is actually fulfilled.

An alternative to a traditional evolutionary model is a learning model. By learning, we mean a mechanism by which a player's present choice depends on previous experience, which in interactive environments includes the choices made by other players. In learning models, players may be endowed with small or large memories, and be as sophisticated as we want them to be. Some such models assume that players only look at past actions and outcomes, and choose more frequently those actions that are associated to higher payoffs. In this case, we are far from the traditional model of rational choice, and the problem of strategic indeterminacy does not arise. Other models, however, assume that players are also forward-looking; at every stage of the game, players make probabilistic conjectures on the opponents' strategies, and then they maximize expected utility on the basis of such conjectures. Though players are usually assumed to ignore the effects that their own choices have on their opponents' future choices, they are endowed with belief-revision capabilities and modify their conjectures on the basis of their observations of how the opponents have played. It can be proved that, in the long run, if a learning dynamics converges (and thus players' subjective probabilistic assessments coincide with the observed frequencies), the limit is a Nash equilibrium (Fudenberg & Levine 1998). Clearly, in any learning model, observation of other players' choices is crucial. But what does a player observe? Suppose a simultaneous-move game is played repeatedly. After each stage game, since actions and strategies coincide, what a player observes are the opponents' strategies (single uncontingent actions). It is worthwhile to note that in such models one need not assume complete information or common knowledge of rationality. What is important to assume is that players are epistemically rational in a weak sense: their beliefs must be internally consistent, and belief revision is done according to Bayes rule. Strategic indeterminacy seems to be resolved in that, when a successful learning process will lead the players to a Nash equilibrium, such equilibrium can reproduce indefinitely, and thus players' predictions about each other's actions turn out to be accurate.

The problem with these kinds of models is that most of the interactions we want to represent are *dynamic* ones. In a dynamic game, players only observe the terminal nodes that are reached in that play of the game, not the parts of their opponents' strategies that specify how they would have played at information sets in unreached parts of the tree. Thus players cannot observe their opponents' strategies, since, in this case, a strategy does not coincide with an action. Note that a strategy is a complete, contingent plan of action that tells a player how to behave in all sorts of circumstances (i.e., at every information set). In a dynamic game, it is impossible to tell how an opponent *would have* played in circumstances that did not occur; all that can be observed are the actions performed during the game. The problem here is that a conjecture can be compatible with what is observed, but it may be wrong. Consider again the game in Figure 6.1. Suppose player I believes, for whatever reason, that player II would only buy for a low price. In this case, the optimal choice for player I would be to ask for a low price, and if II accepts, players will be locked in the outcome $(1, y)$. Player I will have no reason to change his

initial conjecture, and will never be able to know how II would have reacted to a higher price. In this case, the outcome (1, *y*) is compatible with the Nash equilibrium (1, *yn*), but *yn* might not be the strategy chosen by II.

What we are facing here is an interesting twist: Players can learn to correctly predict the outcome of the game, and thus reproduce it indefinitely, even if they are wrong about each others' strategies. In fact, players can even generate stable outcomes that are incompatible with Nash equilibrium. Fudenberg and Levine (1998) did show that there are situations in which each player chooses a best reply to her conjecture, and that conjecture is compatible with the pattern of play she observes, but the *self-confirming equilibria* thus obtained are not Nash equilibria.

The conclusions we can draw for strategic indeterminacy are not reassuring. On the one hand, we have seen that, in the context of common knowledge of SG, PF and Rat, strategic indeterminacy can be resolved only by endowing players with an unrealistic load of extra information (and common knowledge thereof). For example, players would have to have common knowledge of how to interpret deviations from the equilibrium path, have a common understanding of how to prioritize within a hierarchy of possible interpretations, have common priors about players' types, and so on. Note that in this case we resolve strategic indeterminacy by completely eliminating epistemic indeterminacy. Only in these circumstances Nash predictions can be made by the theorist *and* the players. If instead we abandon full rationality and information in favor of an evolutionary approach, Nash equilibrium can be justified as the outcome of a process of natural selection. In this case, we have completely bypassed the problem of strategic indeterminacy, since players have no need to *reason to* an equilibrium. Finally, learning models cast doubt on the possibility of predicting Nash equilibria. Limited observability of players' actions may prevent convergence to a stationary state, and even with observability, dynamic games may converge to stable and stationary states that are not Nash equilibria.

3. Experiments

As we shall see next, the experimental approach, by controlling the rules of the game, the monetary payoffs, and the amount of knowledge players have about these parameters, seems at first sight a viable solution to the problem of indeterminacy. Observations about players' behavior are generated in the laboratory, where the experimenter can control the game description, the order of moves, players' information about earlier moves, the outcomes and their relation with players' moves, as well as players' knowledge of all of the above. However, since players' preferences over outcomes cannot be easily controlled, the experimenter will have to make hypotheses about players' preferences, and about players' knowledge about each other's preferences. Only in this case we will be able to make predictions about the

outcome of the game. When experimental economists started testing the prediction that players converge to a Nash equilibrium, the default auxiliary hypotheses were that players only have selfish preferences over monetary outcomes, and that this fact is common knowledge among them. The falsification of many such predictions in a variety of games has led some to claim that Nash equilibrium theory has been falsified, but all that was falsified are the auxiliary hypotheses about players' preferences and common knowledge thereof. The challenge now is to make new, better hypotheses about players' utilities, hypotheses that are general enough to explain the results of a variety of experiments, and are specific enough to allow for meaningful predictions. To illustrate the difficulties and potential pitfalls of the new approach, as well as the consequences for the indeterminacy problem, I shall now turn to a well known experimental game that has engaged both theorists and experimentalists in an attempt to make sense of the unexpected results.

In 1982, Guth, Schmittberger and Schwarze published a study in which they asked subjects to play what is now known as an Ultimatum bargaining game. Their goal was to test the predictions of game theory about equilibrium behavior. Their results instead showed that subjects consistently deviate from what game theory predicts. To understand what game theory predicts, and why, let us consider a typical Ultimatum game. Two people must split a fixed amount of money M according to the following rules: the proposer (P) moves first and offers a division of M to the responder (R), where the offer can range between M and zero. The responder has a binary choice in each case: to accept the offer or to reject it. If the offer is accepted, the proposer receives M-x and the responder receives x, where x is the offer amount. If the offer is rejected, each player receives nothing. If rationality (and self-interest) are common knowledge, the proposer knows that the responder will always accept any amount greater than zero, because Accept dominates Reject for *any* offer greater than zero. Hence proposer should offer the minimum amount guaranteed to be accepted, and responder will accept it. For example, if M = \$10 and the minimum available amount is 1 cent, the proposer should offer it and the offer should be accepted, leaving the proposer with \$9.99. This is the result predicted by *perfect equilibrium* theory.

Experiments find, however, that nobody offers 1 cent or even 1 dollar. Note that such experiments are always one-shot and anonymous. That is, subjects play the game only once with an anonymous partner and are guaranteed that their choice will not be disclosed. The absence of repetition is important to distinguish between generous behavior that is dictated by a rational, selfish calculation and genuine generosity. If an Ultimatum game is repeated with the same partner, or if a player suspects that future partners will know of her past behavior, it may be perfectly rational for players who are only interested in their material payoff to give generously, if they expect to be on the receiving side at a future time. On the other hand, a Receiver who might accept the minimum in a one-shot game might want to reject a low offer at the beginning of a repeated game, in the hope of convincing future proposers to offer more.

In the United States, as well as in a number of other countries, the modal and median offers in one-shot experimental games are 40% to 50% of the total amount, and the mean offers are 30% to 40%. Offers below 20% are rejected about half the time.[9] These results are robust with respect to variations in the amount of money that is being split, and cultural differences (Camerer 2003). For example, we know that raising the stakes from $10 to $100 does not decrease the frequency of rejections of low offers (those between $10 and $20), and that in experiments run in Slovenia, Pittsburgh, Israel, and Tokyo, the modal offers were in the range of 40% to 50% (Hoffman et al. 1998; Roth et al. 1991).

If we go by the default assumption that players only value their monetary outcomes, then we must conclude that the prediction that players will choose the perfect equilibrium has been falsified. However, as I already mentioned, what has been falsified are the auxiliary hypotheses about players' preferences (and their common knowledge of such preferences). Individuals' behavior across games suggests that money is not the sole consideration, and instead there is a concern for fairness, so much so that subjects are prepared to punish at a cost to themselves those that behave in inequitable ways.[10] A concern for fairness is just one example of a more general fact about human behavior: we are often motivated by a host of factors of which monetary incentives are one, and often not the most important. When faced with different possible distributions, we usually care about how we fare with respect to others, how the distribution came about, who implemented it, and why. The variety of reasons we have for behaving one way or another should be incorporated into a utility function, and economists have recently started to develop richer, more complex models of human behavior that try to explain what we have always known: We do care about other people's outcomes. Thus a better way to explain what is observed in experiments is to provide a richer definition of rationality: People still maximize their utilities, but the arguments of their utility functions include other people's utilities.

In what follows, I will look at two possible explanations for the generous distributions we observe in Ultimatum games. There is no room here to provide a detailed account of how to test these explanations against some interesting variations of the game, and the reader is referred to the relevant literature.[11] I want only to note that such testing is not always easy to conduct. The problem is that we still have quite rudimentary theories of how motives affect behavior. And to test a hypothesis about what sort of motives induce us to act one way or another, we have to be very specific in defining such motives, and the ways in which they influence our choices. In the Ultimatum game, the uniformity of responders' behavior suggests that people do not like being treated unfairly. That is, if subjects perceive an offer of 20% or 30 of the money as unfair, they may reject it to "punish" the greedy proposer, even at a cost to themselves.[12] One possible hypothesis we may make is that both proposers and responders are showing a *social preference* for fair outcomes, or an aversion to inequality.[13] If we make this hypothesis, we can still explain the experimental results with a traditional rational choice model, where the agents' preferences take into account the payoffs of others.

In models of inequality aversion, players prefer both more money and that allocations be more equal. Though there are several models of inequality aversion, perhaps the best known and most extensively tested is the model of Fehr and Schmidt (1999). This model intends to capture the idea that people may be uneasy, to a certain extent, about the presence of inequality, even if they benefit from the unequal distribution. Given a group of L persons, the Fehr-Schmidt utility function of person i is

$$U_i(x_1,...,x_L) = x_i - \frac{\alpha_i}{L-1}\sum_j \max(x_j - x_i, 0) - \frac{\beta_i}{L-1}\sum_j \max(x_i - x_j, 0)$$

where x_j denotes the material payoff that person j gets. α_i //I////I// is a parameter that measures how much player i dislikes disadvantageous inequality (an "envy" weight), and β_i//I////I// measures how much i dislikes advantageous inequality (a "guilt" weight).[14] One constraint on the parameters is that $0 < \beta_i$//I////I// $< \alpha_i$//I////I//, which indicates that people dislike advantageous inequality less than disadvantageous inequality. The other constraint is β_i//I////I// < 1, so that agents do not suffer terrible guilt when they are in relatively good positions. For example, a player would prefer getting more without affecting other people's payoff, even though that results in an increase of the inequality.

Applying the model to the Ultimatum game I just described, the utility function is simplified to

$$U_i(x_1, x_2) = x_i - \begin{cases} \alpha_i(x_{3-i} - x_i) & \text{if } x_{3-i} \geq x_i \\ \beta_i(x_i - x_{3-i}) & \text{if } x_{3-i} < x_i \end{cases} \qquad i = 1,2$$

Obviously if the responder rejects the offer, both utility functions are equal to zero, that is, $U_{1reject} = U_{2reject} = 0$. If the responder accepts an offer of x, the utility functions are as follows:

$$U_{1accept}(x) = \begin{cases} (1+\alpha_1)M - (1+2\alpha_1)x & \text{if } x \geq M/2 \\ (1-\beta_1)M - (1-2\beta_1)x & \text{if } x < M/2 \end{cases}$$

$$U_{2accept}(x) = \begin{cases} (1+2\alpha_2)x - \alpha_2 M & \text{if } x < M/2 \\ (1-2\beta_1)x + \beta_2 M & \text{if } x \geq M/2 \end{cases}$$

The responder should accept the offer if and only if $U_{2accept}(x) > U_{2reject} = 0$. Solving for x, we get the *threshold for acceptance:* $x > \alpha_2 M/(1+2\alpha_2$//I////I//). Evidently if α_2//I////I// is close to zero, which indicates that player 2 (R) does not care much about being treated unfairly, the responder will accept very mean offers. On the other hand, if α_2//I////I// is sufficiently big, the offer has to be close to half to be accepted. In any event, the threshold is not higher than $M/2$, which means that hyper-fair offers (more than half) are not necessary for the sake of acceptance.

Note that for the proposer, the utility function is monotonically decreasing in x when $x \geq M/2$. Hence a rational proposer will not offer more than half of

the money. Suppose $x \leq M/2$; two cases are possible depending on the value of $\beta_1//I////I//$. If $\beta_1//I////I//>1/2$, that is, if the proposer feels sufficiently guilty about treating others unfairly, the utility is monotonically increasing in x, and his best choice is to offer $M/2$. On the other hand, if $\beta_1//I////I//<1/2$, the utility is monotonically decreasing in x, and hence the best offer for the proposer is the minimum one that would be accepted, i.e. (a little bit more than) $\alpha_2 M//I////I//(1+2\alpha_2//I////I//)$. Lastly, if $\beta_1//I////I//=1/2$, it does not matter how much the proposer offers, as long as it is between $\alpha_2 M//I////I//(1+2\alpha_2//I////I//)$ and $M/2$. Note that the other two parameters, α_1 $//I////I//$and $\beta_2//I////I//$, are not identifiable in Ultimatum games.

As noted by Fehr and Schmidt, the model allows for the fact that individuals are heterogeneous. Different $\alpha//I////I//$'s and $\beta//I////I//$'s correspond to different types of people. Although the utility functions are common knowledge, the exact values of the parameters are not. The proposers, in most cases, is not sure what type of responders they are facing. Along the Bayesian line, her belief about the type of the responder can be formally represented by a probability distribution P on $\alpha_2//I////I//$ and $\beta_2//I////I//$. When $\beta_1//I////I//>1/2$, the proposer's rational choice does not depend on what P is. When $\beta_1//I////I//<1/2$, however, the proposer will seek to maximize the expected utility:

$$EU(x)=P(\alpha_2 M/(1+2\alpha_2)<x)\times((1-\beta_1)M-(1-2\beta_1)x)$$

Therefore, the behavior of a rational proposer in the Ultimatum game is determined by her own type ($\beta_1//I////I//$) and her belief about the type of the responder. The experimental data suggest that for many proposers, either $\beta//I////I//$ is big ($\beta//I////I//>1/2$), or they estimate the responder's $\alpha//I////I//$ to be large. The choice of the responder is only determined by his type ($\alpha_2//I////I//$) and the offer. Small offers are rejected by responders with a positive $\alpha//I////I//$.

The positive features of the Fehr-Schmidt utility function are that it can rationalize both positive and negative outcomes, and that it can explain the observed variability in outcomes with heterogeneous types. One of the major weaknesses of their model, however, is that it has a consequentialist bias. Players only care about final distributions of outcomes, not about how such distributions come about. However, recent experiments have established that how a situation is framed matters to an evaluation of outcomes, and that the same distribution can be accepted or rejected depending on "irrelevant" information about the players or the circumstances of play (Bicchieri 2006; Camerer 2003). Another difficulty with this approach is that, if we assume the distribution of types to be constant in a given population, then we should observe, overall, the same proportion of fair outcomes in Ultimatum games. Not only this does not happen, but we also observe individual inconsistencies in behavior across different situations in which the monetary outcomes are the same. If we assume that individual preferences are stable, then we would expect similar behaviors across Ultimatum games. If instead we conclude that preferences are context-dependent, then we should provide a mapping from contexts to preferences that indicates in a fairly predictable way how and why a given context or

situation changes one's preferences. Of course, different situations may change a player's expectation about another player's envy or guilt parameters, and we could thus explain why a player may change her behavior, depending on how the situation is framed. In the case of Fehr and Schmidt's utility function, however, experimental evidence implies that a player's *own* β (or α) changes value in different situations (Bicchieri 2006, Chapter 3). Yet nothing in their theory explains why one would feel consistently more or less guilty (or envious) depending on the decision context.

4. Norms and Expectations

To make clear what I mean, let us consider the results of a questionnaire distributed to 100 Carnegie Mellon undergraduate students that depicted three situations in which the payoffs were the same, but the descriptions of the situation significantly differed.[15]

1. Imagine you must choose how to allocate $10 between yourself and person Y, whom you don't know. You must allocate the money in one of two ways:

A. You and person Y both get $0
B. You get $2 and person Y gets $8

82% of the students choose B, (2, 8).

2. In this scenario, you can offer whatever you want, but Y lets you know that she wants to be offered more than $5. Y announces that you must offer $8 (and keep $2) or she will reject the offer. To prove this to you, Y takes a "commitment pill" that will biologically compel her to reject any offer of less than $8.

A. You keep more than $2 of the $10 for yourself and Y rejects—both you and Y get $0
B. You offer to keep only $2 of the $10 and Y accepts—you get $2 and Y gets $8

Only 49% of the students choose B (2, 8).

3. In this scenario, Y wants to be offered more than $5. Y lives on an island with a different culture, but where the people are of similar wealth to you. In Y's culture, the "last mover" is perceived as the person who controls this game, and is expected to get more of the money. An anthropologist, who will phone your offer to Y, informs you that in Y's culture, any offer less than $8 is viewed as insulting. If you do not offer the split $2, $8, Y will reject your offer. The anthropologist tells you that if the roles were reversed, Y would offer you $8. There is no anthropologist telling Y what you think is fair.

Here 63 percent of the students choose B (2, 8).

Note that all these choices are consequentially equivalent: either both parties get 0, or we have a (2,8) distribution. As I argued before, most models of social preference are consequentialist, but the results I am reporting show that people assign a value to the process through which the outcome is obtained.[16] The (2,8) outcome in the "commitment pill" choice is clearly less attractive; here the responder is rejecting potentially fair offers and, by taking the pill, has given herself an unfair advantage.[17]

In the anthropologist scenario, the responder obeys a different rule, and doesn't know you don't know it. Moreover, it is clear that the rule is symmetrical: Were the responder in the proposer's role, he would just keep $2. If we consider the three cases, and the students' responses, it seems that what makes the difference is the presence (or absence) of social norms that can be violated, and violation consistently elicits a negative reaction in a majority of participants.

The first case presents a simple choice. Only a person with a strong aversion to inequality would choose the (0, 0) outcome, and lose $2. Such people exist, but their number is quite small. Because there is no clear rule about how to behave, the (2, 8) outcome seems the obvious choice. The second case instead is one in which Y is patently unfair, and the majority of students choose to punish him, at a cost to themselves. The last case is one in which a different norm is at work, and thus the receiver who expects $8 is not seen as greedy or manipulative.

If a person has a strong aversion to inequitable outcomes, the number of rejections should stay the same, irrespective of the description of the situation. But the above examples and many experimental results show that this is not the case; most people are extremely sensitive to the way a situation is framed, and when fairness is at stake, many will choose to punish transgressions at a cost to themselves.[18] Preferences, that is, are conditional on the decision context. But what exactly *maps* a context into a specific interpretation that involves, among other things, expectations, beliefs, and causal attributions about other people's motives and future behaviors? I have argued elsewhere (Bicchieri 2006, Chapter 2) that we interpret any situation we are in, and especially new ones, according to scripts that represent stored, generic knowledge about classes of situations. We have scripts that describe what happens at parties, lectures, family reunions, party meetings, and so on. Such scripts contain roles, sequence of actions rules, beliefs and expectations regarding individuals' roles, as well as prescriptions for unexpected occurrences. Scripts are typically shared within a given culture, and, indeed, what is apparent from a variety of experiments is that individuals share a common understanding and interpretation of the experimental situation and the kind of behavior that is most appropriate in those circumstances.[19] Social norms, I have argued, are embedded into scripts. In the typical Ultimatum game, once a fair division script is activated, players will have definite beliefs about what the proposer should offer, especially if they do not have any specific information about him. If a fairness norm is prompted, not only will one expect to get a fair share, but one will be ready to attribute an unfair share to the greediness of the proposer, feel outraged, and retaliate.

Since script activation involves activation of the appropriate expectations and beliefs, the *expectations* subjects have about what others do in the same situation, as well as about what they believe is expected of them, play an important role in guiding their choices. The majority of individuals do not show a consistent disposition to behave in a cooperative, trusting, or fair way. People do not punish transgressors in all circumstances, nor do they positively reciprocate in all cases in which reciprocation is a possible choice. Rather, individuals change their behavior according to the way the situation is framed, which in turn generates very different expectations about what other individuals similarly situated would do, as well as beliefs about what one is expected to do in such situations. Experimental data show that such expectations, when elicited, are interpersonally consistent, and I want to argue that the social norms that generate them are key to understanding experimental behavior and can offer a solution to the indeterminacy problem.

My definition of social norm (see Appendix) is different from the traditional sociological ones, in that I understand a social norm to be a behavioral rule that is supported by (and consists of) the empirical and normative expectations of those who abide by it. People, I have argued, have a *conditional preference* for following a norm, provided their expectations are met (Bicchieri 2006, Chapter 1). In an Ultimatum game, for example, the proposer will have an incentive to be fair if she believes the responder expects a fair share, and in the absence of any other information that is precisely what most proposers expect. Note that I am not assuming the proposer *wants or prefers* to be fair, unconditionally. Surely there are such individuals, but we need not count on them to have fair distributions. It is enough to assume that most individuals conditionally prefer to be fair given that they believe that (a) others typically behave in a fair way, and (b) they are expected to choose a fair division. Whether their motive is fear of retaliation or just the recognition of others' legitimate expectations is not relevant to the present discussion.

The norm-based utility function I introduced in (2006) can now be applied to the Ultimatum game. Let $\pi_i(I)$ be the payoff function for player i. The norm-based utility function of player i depends on the strategy profile s, and is given by

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

where $k_i \geq 0$ is a constant representing i's sensitivity to the relevant norm. Such sensitivity may vary with different norms; for example, a person may be very sensitive to equality and much less so to equity considerations. The first maximum operator takes care of the possibility that the norm instantiation (and violation) might be ambiguous in the sense that a strategy profile instantiates a norm for several players simultaneously (as would be the case, for example, in a social dilemma with three players). The second maximum operator ranges over all the players other than the norm violator. In plain words, the discounting term (multiplied by k_i) is the maximum payoff deduction resulting from all norm violations.

In the traditional Ultimatum game, the norm usually prescribes a fair amount the proposer ought to offer. The norm functions that represent this norm are the

following: N_1 is a constant N function, and N_2 is nowhere defined.[20] If the responder rejects, the utilities of both players are zero.

$$U_{1\text{reject}}(x)=U_{2\text{reject}}(x)=0$$

Given that the proposer offers x and the responder accepts, the utilities are the following:

$$U_{1accept}(x) = M - x - k_1 \max(N_1 - x, 0)$$
$$U_{2accept}(x) = x - k_2 \max(N_2 - x, 0)$$

where N_i denotes the amount player i thinks he should get/offer according to some social norm applicable to the situation, and k_i is non-negative. Note that k_1 measures how much player 1 dislikes to deviate from what he takes to be the norm. To obey a norm, sensitivity to the norm need not be high. Fear of retaliation may make a proposer with a low k behave according to what fairness dictates but, absent such risk, his disregard for the norm will lead him to be unfair. I assume here it is common knowledge that $N_1 = N_2 = \text{N}$, which is reasonable in the traditional Ultimatum game. Again, the responder should accept the offer if and only if $U_{2\text{accept}}(x) > U_{2\text{reject}} = 0$, which implies the following *threshold for acceptance:* $x > k_2N/(1+k_2)$. Notice that an offer larger than the norm dictates is not necessary for the sake of acceptance.

For the proposer, the utility function is decreasing in x when $x \geq N$, hence a rational proposer will not offer more than N. Suppose $x \leq N$. If $k_1 > 1$, the utility function is increasing in x, which means that the best choice for the proposer is to offer N. If $k_1 < 1$, the utility function is decreasing in x, which implies that the best strategy for the proposer is to offer the least amount that would result in acceptance, that is, a little bit more than the threshold $k_2N/(1+k_2)$. If $k_1 = 1$, it does not matter how much the proposer offers, provided the offer is between $k_2N/(1+k_2)$ and N.

It should be noted that k_1 plays a very similar role as that of β_1 in the Fehr-Schmidt model. In fact, if we take N to be $M/2$ and k_1 to be $2\beta_1$, the two models agree on what the proposer's utility is. It is equally apparent that k_2 in this model is analogous to α_2 in the Fehr-Schmidt model. There is, however, an important difference between these parameters. The α's and β's in the Fehr-Schmidt model measure people's degree of aversion toward inequality, which is a very different disposition than the one measured by the k's, that is, people's sensitivity to different norms. The latter will usually be a stable disposition, and behavioral changes may thus be caused by changes in focus or in expectations. A theory of norms can explain such changes, whereas a theory of inequity aversion does not.

It is also the case that the proposer's belief about the responder's type figures in her decision when $k_1 < 1$. The belief can be represented by a joint probability over k_2 and N_2, if the value of N_2 is not common knowledge. The proposer should choose an offer that maximizes the expected utility

$$EU(x) = P(k_2 N_2 / (1 + k_2) < x) \times (M - x - k_1 (N_1 - x)).$$

If we now apply the above utility function to the questionnaire we just discussed, it is reasonable to assume that in the Ultimatum game the norm prescribes (5, 5) offers, and the proposer should thus never expect to keep less than 5. Her choice in the pill scenario is whether to induce rejection or accept $2:

$$U_{1reject}\ (x) = 0$$

$$U_{1accept}\ (x) = 2 - k_1\ \ (5 - 2)$$

If $k_1 > 2/3$, *reject* is the utility-maximizing choice. In the anthropologist scenario instead, the conditions for obeying the (5, 5) norm fail: the responder neither knows the rule nor expects the proposer to follow it. Actually, the relevant norm is (2, 8). The discounting term drops out and it is again preferable to offer 8.

Consider once more the game in Figure 6.1. Let us suppose that it represents a situation in which, for whatever reason, low price is the norm. In this case, even if player I does not know the value of *y*, she knows that player II will reject an offer of 2 for any $y > 2$ *and* $k_{II} > 1$.[21] Player I is thus playing a Bayesian game in which player II can be one of different types.[22] However, his prior probabilities are influenced by the norm's existence. The norm points to the equilibrium (1, *yn*) and, in case the norm is de facto followed in the population, player I will have good reason to assess a high probability to $k_{II} > 1$. Note that assessing the *k* of player II is crucial even if I were to know the value of *y*. For example, if $y=3$ and in the absence of a norm, player I would most certainly choose the high price. The presence of a norm, however, drastically changes the situation, since now player I has to assess the probability that II cares about the norm. In this case, it is better for player I to choose the low price, even if he personally does not care that much about following the norm. Expectations, in other words, are crucial to our decision to obey norms.

I said before that norms are a way to solve the indeterminacy problem, and I have argued elsewhere (Bicchieri 2006) that established norms are equilibria. However, since social norms often go against our self-interest, especially when we narrowly interpret self-interest as a desire for material incentives, a social norm need not be an equilibrium of an ordinary game in which payoffs represent self-interested preferences. Thus, for example, a cooperative norm cannot be a Nash equilibrium of a Prisoner's Dilemma (PD) game. If such a norm exists and is followed, however, the original PD game would be transformed (at least for the norm-followers) into the subsequent, very different game:

In the traditional PD, each player's preference ranking is $DC > CC > DD > CD$. B in Figure 6.4 stands for best, S for second best, and so on. In the symmetric coordination game instead, each norm follower's preference ranking is $CC > DD > DC > CD$.[23] That is, the players who follow a cooperative norm will do it because their empirical and normative expectations have been met, hence they *prefer* to obey the norm. The new coordination game has two *strict* Nash equilibria, one of which is Pareto superior to the other.[24, 25] When a norm of cooperation exists and is

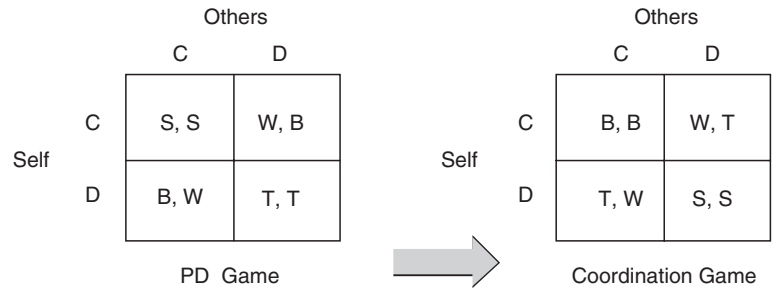

Figure 6.4.

obeyed, a game like the PD above is *transformed into a coordination game:* Players' payoffs in the new game will differ from the payoffs of the original game, since their preferences and beliefs will be as in conditions 2, 2(a) and 2(b) or 2(b′) in the Appendix. Indeed, if a player knows that a cooperative norm exists and expects a sizeable part of the population to follow it, then, provided she also believes she is expected (and maybe also prefers) to follow such norm, she will have a preference to conform to the norm in a situation in which she has the choice to cooperate or to defect. Note that what I am saying implies that a social norm, unlike a convention, is never a solution of an original coordination game, though it is an equilibrium of the new, transformed game it *creates*.

More formally, and to further illustrate the norm-based utility function introduced above, consider the PD we are discussing. The norm-based function for either player is defined at C and undefined at D. The utility function for player 1 is then the following:

$$U_1(C,C)=\pi_1(C,C)-k_1(\pi_1(C,C)-\pi_1(C,C))=\pi_1(C,C)$$
$$U_1(D,D)=\pi_1(D,D)-k_1(\pi_1(D,D)-\pi_1(D,D))=\pi_1(D,D)$$
$$U_1(C,D)=\pi_1(C,D)-k_1(\pi_1(C,C)-\pi_1(C,D))$$
$$U_1(D,C)=\pi_1(D,C)-k_1(\pi_2(C,C)-\pi_2(C,D))$$

Player 2's utility function is similar. The game turns out to be a coordination game with two equilibria when $U_1(D,C) < U_1(C,C)$ and $U_2(D,C) < U_2(C,C)$, that is, when[26]

$$k_1 > \frac{\pi_1(D,C)-\pi_1(C,C)}{\pi_2(C,C)-\pi_2(C,D)}$$
$$k_2 > \frac{\pi_2(D,C)-\pi_2(C,C)}{\pi_1(C,C)-\pi_1(C,D)}$$

Otherwise it remains a PD.

It is important to note that my definition of social norm does not entail that *everybody* conforms. In fact, the definition (see Appendix) says that a social norm may exist and not be followed. For some, the PD in our example is never trans-

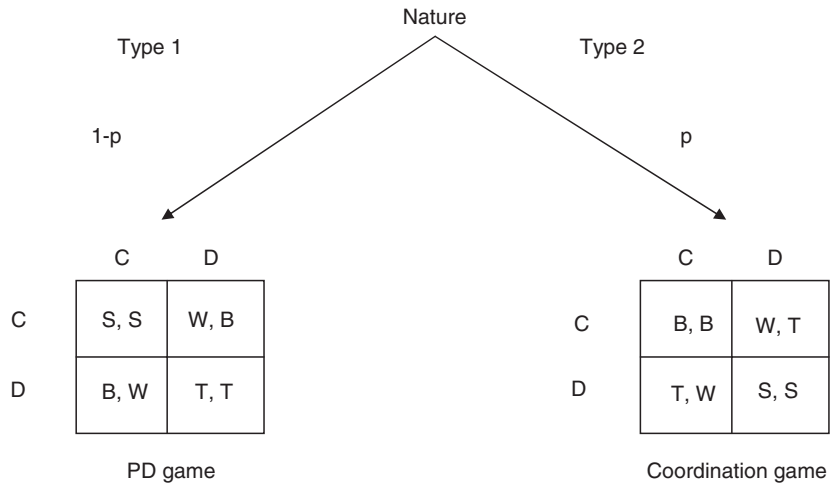

Figure 6.5.

formed into any other game. And even a person who starts playing a coordination game like the one described above may revert to playing the regular PD if she realizes that *empirical expectations* have been violated. Let me clarify this point with a simple example. Suppose an actor is faced with a finitely repeated PD, and suppose the situation is such that a cooperative norm is primed. The player knows there exists a cooperative norm that applies to this kind of situation. The player also knows that there are several types of players, some of whom would not see the game as he does. To make matters easy, suppose there are two types of players, those who simply see the game as a PD, and those that follow a cooperative norm.[27] In this case we may model the choice situation as a Bayesian game (Figure 6.5) in which Nature picks a player type with a given probability, so that with prior probability p the opponent one faces is playing a coordination game, and with probability (1-p) he is playing a PD. If a norm-follower assesses a sufficiently high probability to being matched with a similar type, he will cooperate.[28]

Note that, when faced with a defection, a player will reassess his probabilities, and possibly revert to playing the equilibrium strategy (defect) for the traditional PD. Thus, one might say that the *existence* of a norm always presents a conditional follower with a Bayesian game: If the normative and empirical expectations conditions are fulfilled, she will attach a higher probability to being matched with a similar player type (a norm follower), and act accordingly. But she must also be prepared to revise her probabilistic assessment in case experience contravenes her previous expectations.[29] Note that the existence of a social norm facilitates equilibrium selection in the Bayesian game faced by the conditional norm's followers. If the probability of being matched with a similar type is high enough, *C,C* is the selected equilibrium. Otherwise *D,D* will be selected.

5. Conclusions

The search for better, empirically grounded theories of how agents make decisions has led experimental economists to make many new assumptions about what motives guide us, and to incorporate these motives in utility functions. We are a long way from designing utility functions that are general enough to capture the richness of experimental data available. A promising way to proceed is to include a truly social component into utility functions. An example is the introduction of social norms, since their existence shapes our choices through the expectations they generate. We can maintain a traditional rational choice model, but make more interesting and realistic auxiliary hypotheses about preferences and motives. Epistemic indeterminacy is mitigated by the existence of social norms since, as I have argued, norms come in packages that include beliefs, expectations, causal attributions, and so on. Strategic indeterminacy is mitigated, too, since the existence of a norm points to a specific equilibrium, allowing players to coordinate on it. I use the term *mitigate* because players still have to assess their opponents' sensitivity to the relevant norm, and their choice will depend on this assessment. Experimental evidence, however, points to the fact that players, when a norm applies to the situation they are in, tend, *ceteris paribus,* to have quite uniform expectations of their opponents' caring about it (they are expected to care) and act accordingly in a uniform way.[30] Bringing more empirical data of this kind into our models is the only way I can see to solve the indeterminacy problem that has plagued otherwise excellent choice models for too long.

Appendix: Conditions for a Social Norm to Exist

Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motive game. R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subseteq P$ such that, for each individual $i \in P_{cf}$:

Contingency: i knows that a rule R exists and applies to situations of type S;
Conditional preference: i prefers to conform to R in S on the condition that:

(a) *Empirical expectations: i* believes that a sufficiently large subset of P conforms to R in S;

and either

(b) *Normative expectations: i* believes that a sufficiently large subset of P expects i to conform to R in S;

or

(b′) *Normative expectations with sanctions: i* believes that a sufficiently large subset of *P* expects *i* to conform to *R* in *S*, prefers *i* to conform and may sanction behavior.

A social norm *R* is *followed* by population *P* if there exists a sufficiently large subset $P_f \subseteq P_{cf}$ such that, for each individual $i \in P_f$, conditions 2(a) and either 2(b) or 2(b′) are met for *i* and, as a result, *i* prefers to conform to *R* in *S*.

NOTES

1. By *practical rationality* I mean that an agent will choose that action that best fulfills her goals, given her beliefs about the situation. By *epistemic rationality* I refer to the rationality of an agent's beliefs. This may simply mean that probabilistic beliefs obey the axioms of probability calculus, but it may also mean that an agent will use all the statistical data that are available to her (see Bicchieri 1993, Chapter 1)

2. In fact, we can even *infer* a person's utility by looking at a sequence of choices she made, provided we assume she is consistent. F. P Ramsey (1931) was the first to show how, by observing a series of bets an individual is prepared to make, it is possible to infer both her preferences and probabilistic beliefs.

3. An event *p* is *common knowledge* among the players if all the players know that *p*, all know that all know that *p*, and so on. (Lewis 1969; Aumann 1976).

4. When players have common knowledge of the rules of the game and of their mutual preferences, the game is one of *complete information*. In our example, if I does not know *y*, or if I knows *y* but he does not know that II knows that, then the game is one of incomplete information (Harsanyi 1967–1968).

5. Note that optimality is only conditional on a fixed σ_{-i}, not on all possible σ_{-i}. A strategy that is a best reply to a given combination of the opponents' strategies may fare poorly vis a vis another strategy combination.

6. One important virtue of Nash equilibrium is that for games with a finite number of pure strategies and finitely many players, a Nash equilibrium always exists, at least in mixed strategies (Nash 1951).

7. By contrast, $\hat{p}$ is a symmetric *Nash equilibrium* if $\hat{p}^T A\hat{p} \geq p^T A\hat{p}$ for all feasible *p*.

8. Note that (*) is a deterministic system, which allows some strategies to become extinct, in the sense that $pi(t) = 0$ for some *i*,*t*. To prevent extinction, mutations are added, but a discussion of how to modify (*) to include mutations and how to interpret the latter would take us too far from the present topic. For an analysis of stochastic models, see Foster and Young (1990).

9. Guth et al. (1982) were the first to observe that the most common offer by proposers was to give half of the sum to the responder. The mean offer was 37 percent of the original allocation. In a replication of their experiments, they allowed subjects to think about their decision for one week. The mean offer was 32 percent of the sum, which is still very high.

10. We know that responders reject low offers even when the stakes are as high as three months' earnings (Cameron 1995). Furthermore, experiments in which third parties have a chance to punish an unfair proposer at a monetary cost to themselves show that (moderately) costly punishment is frequent (Fehr & Fishbacher 2000).

11. See Camerer (2003, Chapter 3) and Bicchieri (2006, Chapter 3).

12. Note, again, that the experiments I am referring to were all one-shot, which means that the participants were fairly sure of not meeting again; therefore, punishing behavior cannot be motivated as an attempt to convince the other party to be more generous the next time around. Similarly, proposers could not be generous because they were expecting reciprocating behavior in future interactions.

13. By *social preference* I refer to how people rank different allocations of material payoffs to self and others.

14. The term max (x_j—x_i, 0) denotes the maximum of x_j—x_i and 0; it measures the extent to which there is disadvantageous inequality between i and j.

15. The questions were devised by Jason Dana and Daylian Cain, who were taking my course on social norms.

16. I have extensively discussed this point in Bicchieri (2006, Chapter 3)

17. T. Schelling (1960) presents several cases of 'commitment strategies' that help one of the parties to get the upper hand in negotiating an agreement.

18. See, for example, Fehr et al. (2003), Dana, Weber *et al* (2003), Frey and Bohnet (1995), Hoffman, McCabe et al. (1994), Bicchieri and Chavez (2007).

19. Bicchieri and Chavez (2007).

20. Intuitively, N_2 should proscribe rejection of fair (or hyperfair) offers. The incorporation of this consideration, however, will not make a difference in the formal analysis.

21. Note that for $1<y<2$, I knows II will only accept the low price.

22. When players are uncertain as to the type of player they are facing, they will assess some probability that the other player is of a certain type. Typically, the list of all possible types and their prior probability of occurring in the population are taken to be common knowledge among the players (Harsanyi 1967, 1968).

23. For a justification of this ranking, see Bicchieri 2006, 16–19.

24. In a strict Nash equilibrium each player's strategy is a unique best reply to the other players' strategies. This means that a strict Nash equilibrium cannot include weakly dominated strategies.

25. A coordination game is a game in which there are at least two Nash equilibria in pure strategies, and players have a mutual interest in reaching one of these equilibria (*CC* or *DD* in our game), even if different players may prefer different equilibria (which is not the case in the above example).

26. Note that $U_1(D,C)$ stands for the utility of player 1 when 1 plays *D* and 2 plays *C*. Analogously, $U_2(D,C)$ stands for the utility of player 2 when 1 plays *C* and 2 plays *D*.

27. In a finitely repeated game, even a selfish' player may want to cooperate for a while, if it is not common knowledge that all players are rational and selfish (Kreps et al. 1982). This consideration, however, has no bearing on my argument, since until a defection is observed a player cannot distinguish between a forward-thinking selfish type and a true cooperator.

28. If players use an availability heuristic to come to this probability assessment, the probability of playing a coordination game might initially be much higher. That is, if a player is the type who follows a cooperative norm, that player tends to believe there is a high probability that others are like him or her.

29. Recent experiment I conducted showed that, even if subjects do not particularly care about a (fairness) norm, their expectations about their partner's sensitivity to it drive their choices (Bicchieri & Chavez 2007). Moreover, we also discovered that subjects are very sensitive to what other people in their situation have done, and when there is conflict between normative and empirical expectations, the latter always win (Bicchieri & Xiao 2007).

30. Bicchieri and Chavez (2007), Bicchieri and Lev-on (2007).

REFERENCES

Aumann, R. (1976). "Agreeing to Disagree." *Annals of Statistics* 4: 1236–1239.

Bernheim, D. (1984). "Rationalizable Strategic Behavior." *Econometrica* 52: 1007–1028.

Bicchieri, C, (1993). *Rationality and Coordination*. Cambridge: Cambridge University Press.

Bicchieri, C. (1988). "Strategic Behavior and Counterfactuals." *Synthese* 76: 135–169.

Bicchieri, C. (2006). *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge, England: Cambridge University Press.

Bicchieri, C. & Chavez, A. (2007). "The Fragility of Fairness: How Beliefs Affect Behavior in Ultimatum Games." Discussion paper, Goldstone Research Unit, University of Pennsylvania, May 2007.

Bicchieri, C. & Lev-on, A. (2007). "Computer-Mediated Communication and Cooperation in Social Dilemmas: An Experimental Analysis." *Politics, Philosophy and Economics* 6 (2): 139–168.

Bicchieri, C. & Xiao, E. (2007). "Do the Right Thing: But Only If Others Do So." http://d.repec.org/n?u=RePEc:pra:mprapa:4609&r=cbe

Binmore, K. (1987, 1988). "Modeling Rational Players I and II." *Economics and Philosophy* 3: 9–55, and 4: 179–214.

Camerer, C. (2003). *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton: Princeton University Press.

Cameron, L. (1995). "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia." *Working paper, Princeton Department of Economics—Industrial Relations Sections* 345.

Dana, J., Weber, R. & Kuang, J. (2003). "Exploiting Moral Wriggle Room: Behavior Inconsistent with a Preference for Fair Outcomes." *Carnegie Mellon Behavioral Decision Research Working Paper* 349.

Fehr, E. & Fischbacher, U. (2003). "The Nature of Human Altruism." *Nature* 425: 785–791.

Fehr, E. & Schmidt, K. (1999). "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics* 114 (3): 817–868.

Foster, D., & Young, H.P. (1990). Stochastic Evolutionary Game Dynamics. *Theoretical Population Biology* 38: 219–232.

Frey, B. & Bohnet, I. (1995). "Institutions Affect Fairness: Experimental Investigations." *Journal of Institutional and Theoretical Economics* 151 (2): 286–303.

Fudenberg, D. & Levine, D.K. (1998). *The Theory of Learning in Games*. Cambridge, MA: MIT Press.

Guth, W., Schmittberger, R. & Schwarze, B. (1982). "An Experimental Analysis of Ultimatum Bargaining." *Journal of Economic Behavior and Organization* 3: 367–388.
Harsanyi, J. (1967–68). "Games with Incomplete Information Played by 'Bayesian' Players." Parts 1, 2, and 3. *Management Science* 14: 159–182, 320–332, 468–502.
Hofbauer, J. & Sigmund, K, (1998). *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.
Hoffman, E., McCabe, K.A,. Shachat, K., & Smith, V. (1994). "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior* 7: 346–380.
Hoffman, E., McCabe, K.A. & Smith, V. (1998). "Behavioral Foundations of Reciprocity: Experimental Economics and Evolutionary Psychology." *Economic Inquiry* 36: 335–352.
Kreps, D., Milgrom, P., Roberts, J., and Wilson, R. (1982). Rational Cooperation in the Finitely Repeated Prisoner's Dilemma. *Journal of Economic Theory* 27: 245–252.
Lewis, D. (1969). *Convention: A Philosophical Study*. Cambridge, MA: Cambridge University Press.
Maynard Smith, J. & Price, G. (1973). "The Logic of Animal Conflict." *Nature* 246: 15–18.
Nash, J. (1951). "Non-cooperative Games." *Annals of Mathematics* 54: 286–295.
Pearce, D. (1984). "Rationalizable Strategic Behavior and the Problem of Perfection." *Econometrica* 52: 1029–1050.
Ramsey, F.P. (1931). "Truth and Probability." In R. B. Braithwaite, Ed., *The Foundations of Mathematics and Other Logical Essays*. London: Routledge and Kegan Paul.
Roth, A.E., Prasnikar, V. Okuno-Fujiwara, M. & Zamir, S. (1991). "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review* 81 (5): 1068–1095.
Schelling, T. (1960). *The Strategy of Conflict*. Cambridge: Harvard University Press.
Selten, R. (1975). "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4: 22–55.
Taylor, P.D. & Jonker, L.B. (1978). "Evolutionary Stable Strategies and Game Dynamics." *Mathematical Bioscience,* 40: 145–156.
Van Damme, E. (1987). *Stability and Perfection of Nash Equilibrium*. Berlin: Springer.
Von Neumann, J. & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.