

## 12 Reasoning about strategic interaction

### Solution concepts in game theory

*Andrew M. Colman*

Game theory is concerned with rational choice in decisions involving two or more interdependent decision makers. Its range of applicability is broad, including all decisions in which an outcome depends on the actions of two or more decision makers, called *players*, each having two or more ways of acting, called *strategies*, and sufficiently well-defined preferences among the possible outcomes to enable numerical *payoffs* reflecting these preferences to be assigned.

Decision theory has a certain logical primacy in psychology, because decision making drives all deliberate behaviour, and game theory is the portion of decision theory dealing with decisions involving strategic interdependence. This chapter focuses on reasoning in games, and in particular on theoretical problems of specifying and understanding the nature of rationality in strategic interaction. These problems are far from trivial, because even simple games present deep and mysterious dilemmas that are imperfectly understood and have not been solved convincingly.

The notion of rationality underlying game theory is *instrumental rationality*, according to which rational agents choose the best means to achieve their most preferred outcomes. This means-end characterization of rational choice is conspicuously neutral regarding an agent's preferences or desires, a point that was stressed by the Scottish philosopher David Hume in a frequently quoted passage of his *Treatise of Human Nature*: 'Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them. . . . A passion can never, in any sense, be call'd unreasonable, but when founded on a false supposition, or when it chuses means insufficient for the design'd end.' (1739-40, 2.III.iii). Hume conceded only that preferences based on 'false supposition' are unreasonable or irrational. Contemporary philosophers and game theorists take an even more permissive view, requiring only that preferences should be consistent. Although everyday language contains both internal reason statements (*P has a reason for doing x*) and external reason statements (*There is a reason for P to do x*), the philosopher Bernard Williams (1979) has shown that 'external reason statements, when definitely isolated as such, are false, or incoherent, or really something else misleadingly expressed' (p. 26). A person's reasons for acting in a particular way are invariably internal, hence an action is instrumentally rational, *relative to the agent's knowledge and beliefs at the time of acting*, if it is the best means to achieve the most preferred outcome, provided only that the knowledge and beliefs are not inconsistent or incoherent. Thus, if I am thirsty, and I come upon a jar of powder that I believe to be cocoa but is actually rat poison, I act rationally, relative to my knowledge and beliefs, if I dissolve the powder in hot milk and drink the infusion, even though my preference for doing so is based on a 'false supposition'.

Instrumental rationality is formalized in *expected utility theory*, introduced as an axiomatic system by

von Neumann and Morgenstern (1947), in an appendix to the second edition of their book, *Theory of Games and Economic Behavior*. It is based on the idea that a rational agent has complete and consistent preferences among the available outcomes and also among gambles involving those outcomes. The theory assigns numerical utilities to the outcomes in such a way that players who always choose utility-maximizing options (strategies or gambles) can be shown to be acting in their own best interests and therefore to be instrumentally rational. In game theory, utilities are represented by payoffs, and the theory, as presented by von Neumann and Morgenstern, is primarily *normative*, inasmuch as its basic aim is to determine what strategies rational players should choose to maximize their payoffs. It is not primarily a *positive* or *descriptive* theory that predicts what strategies human players are likely to choose in practice. It can none the less be argued (Colman, in press) that it becomes a positive theory by the addition of a bridging hypothesis of weak rationality, according to which people try to do the best for themselves in any given circumstances. Granted that deviations from perfect rationality are inevitable, because human decision makers have bounded rationality, the bridging hypothesis provides game theory with a secondary objective, that of making testable predictions, thus justifying the otherwise inexplicable enterprise of experimental gaming (reviewed by Camerer, 2003; Colman, 1995, chaps 5, 7, 9; Kagel and Roth, 1995, chaps 1-4; Pruitt and Kimmel, 1977).

The fundamental problem that we encounter when we attempt to determine rational play in games is that individual players have incomplete control over the outcomes of their actions. In individual decision making, expected utility theory provides a clear and unambiguous interpretation of rationality. A rational decision maker chooses the option with the highest expected utility or, if there is a tie for top place, one of the options with the highest expected utility. But a game does not generally have a strategy that is best in this straightforward sense, because a player's preferences range over outcomes, not strategies, and outcomes are determined partly by the choices of other players.

The remainder of this chapter will be devoted to the most prominent suggestions that have been put forward for solving this problem. I shall focus principally on *non-cooperative* games, except for a brief discussion of cooperative games near the end. The distinction between these two classes of games was introduced by Nash (1951). In non-cooperative games, the players act independently, whereas in cooperative games they are free to negotiate coalitions based binding and enforceable agreements. The following sections are devoted to an examination of the ideas behind the major solution concepts – general principles designed to yield rational solutions to particular classes of games. These fundamental issues are seldom discussed in the game-theoretic literature.

## **NASH EQUILIBRIUM**

The leading solution concept for non-cooperative games is undoubtedly *Nash equilibrium*. A Nash equilibrium (or *equilibrium point*, or *strategic equilibrium*, or simply *equilibrium*) is a profile of strategy choices, one for each of the  $n$  players in a game, such that each player's strategy is a *best reply* to the  $n - 1$  others. A best reply is a strategy that maximizes a player's payoff, given the strategies chosen by the others. It follows from the definition that any non-equilibrium profile of strategies is necessarily self-destabilizing, inasmuch as at least one player has an incentive to deviate

from it. It is often claimed, conversely, that an equilibrium point is self-supporting and self-enforcing, but we shall see that this is not always the case. An important psychological property of an equilibrium point is that it gives the players no cause to regret their strategy choices when those of their co-players are revealed.

		<b>II</b>	
		<b>C</b>	<b>D</b>
<b>I</b>	<b>C</b>	<b>4, 4</b>	<b>1, 2</b>
	<b>D</b>	<b>2, 1</b>	<b>3, 3</b>

*Figure 12.1 Assurance Game.*

The equilibrium concept can be illustrated by the Assurance Game, the payoff matrix of which is displayed in Figure 12.1. This is a simple two-person game, first introduced by Sen (1969), in which Player I chooses between the rows arbitrarily labelled *C* and *D*, Player II independently chooses between the columns labelled *C* and *D*, and by convention the pair of numbers in each cell are the payoffs to Player I and Player II in that order. What defines this game as the Assurance Game is the rank order of the payoffs rather than their absolute values – it is still an Assurance Game if the payoffs are 10, 0, -5, and -10, for example, provided that the highest payoff goes to each player in the (*C*, *C*) outcome, the second-highest payoff to each in the (*D*, *D*) outcome, and so on. The identities of other named games also depend on their ordinal structures.

Sen (1969) gave the following interpretation to illustrate how the Assurance Game might arise in an everyday strategic interaction, at least in the dreaming spires of academia. Two people face the choice of going to a lecture or staying at home. ‘Both regard being at the lecture *together* the best alternative; both, staying at home the next best; and the worst is for him or her to be at the . . . lecture without the other’ (p. 4, footnote 5). Given these preferences, the strategic structure corresponds to Figure 12.1. It is clear that the (*C*, *C*) outcome in the top-left cell of the payoff matrix is an equilibrium point because, for Player I, *C* is the best reply to Player II’s *C*, and for Player II, *C* is the best reply to Player I’s *C*. But there is another equilibrium point at (*D*, *D*), where strategies are also best replies to each other. It yields lower payoffs for both players, and it seems intuitively obvious that rational players would choose their *C* strategies, because (*C*, *C*) is not only an equilibrium point but is the best equilibrium point for both players (technically, it *payoff dominant*, and that is something we need to examine more closely later). In fact, (*C*, *C*) is uniquely *Pareto-efficient* in the sense that no other outcome gives either player a higher payoff without giving the other player a lower payoff.

The equilibrium concept was first formalized by Nash (1950a, 1951), who gave two separate proofs that every finite game – that is, every game with a finite number of players, each having a finite number of strategies – has at least one equilibrium point, provided that *mixed strategies* are brought into consideration. A mixed strategy is a probability distribution over a player’s (pure) strategies. For

example, if a player has two pure strategies, such as  $C$  and  $D$  in the Assurance Game, then one feasible mixed strategy involves choosing randomly between them with equal probabilities assigned to each, by tossing a coin, for example; another mixed strategy involves 60%–40% randomization, and so on. In fact, with the payoffs shown in Figure 12.1, it is easy to verify that if both players choose  $C$  and  $D$  with equal probability, these mixed strategies form a third Nash equilibrium, with expected payoffs of  $2\frac{1}{2}$  to each player. In the increasingly popular Bayesian interpretation of game theory, a mixed strategy is viewed construed as uncertainty in the mind of a co-player about which pure strategy will be chosen (Harsanyi, 1973).

What makes Nash equilibrium so important is a theoretical discovery that if a game has a uniquely rational solution, then it must be an equilibrium point. This proposition was deduced by von Neumann and Morgenstern (1944, pp. 146-8), using a celebrated *Indirect Argument*, and prominently expounded by Luce and Raiffa (1957, pp. 63-5, 173). In its current interpretation, the Indirect Argument rests on the standard *common knowledge and rationality* assumptions of game theory. The first of these is that the specification of the game, embodied in a payoff matrix in the case of a two-person game, and everything that follows logically from it, are common knowledge among the players. The second is that the players are instrumentally rational, invariably choosing strategies that maximize their utilities or payoffs, relative to their knowledge and beliefs, and that this too is common knowledge in the game. The concept of *common knowledge* was introduced by Lewis (1969, pages 52-68) and formalized by Aumann (1976). Roughly speaking, a proposition is common knowledge among a group of players if every player knows it to be true, knows that every other player knows it to be true, knows that every other player knows that every other player knows it to be true, and so on.

According to the standard assumptions, the specification of the game and the players' rationality are common knowledge in the game. From these assumptions, it can be proved that any uniquely rational solution must be an equilibrium point. First, an immediate implication of the common knowledge and rationality assumptions is that any conclusion that a player validly deduces about a game will be deduced by the co-player(s) and will be common knowledge in the game. This logical implication is called the *transparency of reason* (Bacharach, 1987). It implies that, if it is uniquely rational for Player 1 to choose Strategy  $s_1$ , Player 2 to choose Strategy  $s_2$ , ..., and Player  $n$  to choose Strategy  $s_n$ ,

then  $s_1, s_2, \dots, s_n$  must be best replies to one another, because, by the transparency of reason, each

player anticipates the others' strategies and, to maximize utility, chooses a best reply to them.

Because  $s_1, s_2, \dots, s_n$  are best replies to one another, they are in Nash equilibrium by definition. This

establishes that if a game has a uniquely rational solution, then that solution must necessarily be an equilibrium point. A deep and subtle problem that is often overlooked is that the converse does not necessarily hold, because the Indirect Argument rests on an unproved assumption that a game has a uniquely rational solution (Sugden, 1991). A Nash equilibrium, even if unique, is not necessarily a rational solution, because a game may have no uniquely rational solution.

## **Unstable equilibrium**

If a particular outcome is a Nash equilibrium, that is not a sufficient reason for a rational player to choose the corresponding equilibrium strategy. This can be seen, first, in certain games with only mixed-strategy equilibrium points, such as the game shown in Figure 12.2. This game has no pure-strategy equilibrium point. Its unique equilibrium point is the mixed-strategy solution in which Player I randomizes between Strategies *C* and *D* with probabilities 2/3 and 1/3 respectively, and Player II randomizes between Strategies *C* and *D* with probabilities 1/3 and 2/3 respectively. If both players use these equilibrium strategies, then Player I's expected payoff is 3• and Player II's is 2•, and neither player can benefit by deviating. The transparency of reason may seem to imply that each player will therefore expect the other to choose the prescribed equilibrium strategy. But if Player I expects Player II to choose (1/3*C*, 2/3*D*), then this nullifies Player I's reason for choosing (2/3*C*, 1/3*D*), because *any* pure or mixed strategy yields an identical expected payoff of 3• against Player II's mixed strategy, and the same argument applies, *mutatis mutandis*, to Player II. This is a valid deduction, and the transparency of reason ensures that it is common knowledge. It implies that neither player has any reason to expect the other to choose a mixed equilibrium strategy. In the mixed-strategy case, not only does the fact that a particular outcome is a Nash equilibrium fail to provide a player with a sufficient reason for choosing the corresponding equilibrium strategy but, on the contrary, it appears to vitiate any reason that a player might have for choosing it.

		<b>II</b>	
		<b>C</b>	<b>D</b>
<b>I</b>	<b>C</b>	<b>3, 3</b>	<b>4, 2</b>
	<b>D</b>	<b>5, 1</b>	<b>3, 3</b>

*Figure 12.2* Game with a unique mixed-strategy equilibrium point.

This suggests that the [(2/3*C*, 1/3*D*), (1/3*C*, 2/3*D*)] mixed-strategy equilibrium solution in Figure 12.2 is unstable. A player can deviate from it unilaterally without suffering any penalty, although there is no positive incentive to do so. Harsanyi (1973) argued, however, that this instability is apparent rather than real, provided that an element of uncertainty is introduced into the modelling of the game. He suggested that a player should always be assumed to have a small amount of uncertainty about a co-player's payoffs. If games with solutions in mixed strategies are modelled by *disturbed games* with randomly fluctuating payoffs, deviating slightly from the values in the payoff matrix, then mixed-strategy equilibrium points disappear and are replaced by pure-strategy equilibrium points, and the fluctuating payoffs interact in such a way that rational players choose strategies with the probabilities prescribed by the original mixed-strategy solution. If the game shown in Figure 12.2 is disturbed, then it will no longer have a mixed-strategy solution. Player I will receive a higher payoff from either the *C* or the *D* strategy – *C* in 2/3 of disturbed games and *D* in 1/3 – and for Player II these proportions will be reversed. Thus, although rational players will simply choose their best pure strategies without

making any attempt to randomize, they will choose them with the probabilities of the classical mixed-strategy solution.

**Subgame-perfect equilibrium**

There is worse to come for Nash equilibrium. Some equilibrium points require players to choose strategies that are arguably irrational. This anomaly was discovered by Selten (1965, 1975), who developed a refinement of Nash equilibrium, called the *subgame-perfect equilibrium*, specifically to eliminate it. A simple example of an imperfect equilibrium is shown in Figure 12.3.

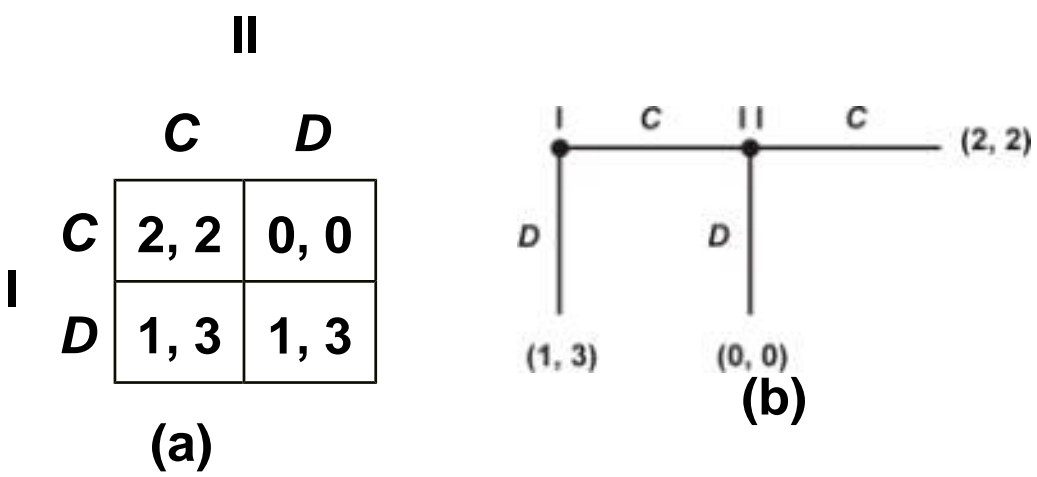


Figure 12.3 Game with an imperfect equilibrium point. (a) Normal form. (b) Extensive form.

In the payoff matrix shown in Figure 12.3(a), both (C, C) and (D, D) are equilibrium points, but (C, C) is subgame-perfect, in Selten’s terminology, and (D, D) is imperfect, requiring an irrational choice from one of the players. This emerges from an examination of the *extensive form* of the game shown in Figure 12.3(b), a graph depicting the players’ moves as if they moved sequentially, starting with Player I on the left. If the game were played sequentially, and if the second decision node were reached, then a utility-maximizing Player II would choose C, to secure a payoff of 2, not D, yielding 0. Working backwards, Player I would anticipate Player II’s reply and would therefore choose C rather than D, to secure 2 rather than 1. Thus we can conclude that (C, C) is the only rational equilibrium point of the extensive-form game, and it is therefore subgame-perfect. Because the (D, D) equilibrium point could not be reached by rational behaviour in the extensive form, it is imperfect in the normal form. A subgame-perfect equilibrium is one that induces payoff-maximizing choices in every branch or subgame of its extensive form.

Selten (1975) introduced the concept of *trembling-hand equilibrium* to identify and eliminate imperfect equilibria. At every decision node in the extensive form of a game there is assumed to be a small probability ε (epsilon) that the player’s rationality will break down for some unspecified reason, resulting in a mistake or unintended move. The introduction of these small error probabilities produces a *perturbed game* – slightly different from Harsanyi’s (1973) disturbed games, in which it is the payoffs rather than the players’ actions that go astray. In Selten’s theory, whenever a player’s hand ‘trembles’, the erroneous move is assumed to be determined by a random process, and every move that could possibly be made at every decision node therefore has some positive probability of

being played. Assuming that the players' trembling hands are common knowledge in a game, Selten proved that only the subgame-perfect equilibria of the original game remain equilibrium points in the perturbed game, and they continue to be equilibrium points as the probability  $\epsilon$  tends to zero. According to this widely accepted refinement of the equilibrium concept, the standard game-theoretic assumption of rationality is reinterpreted as a limiting case of incomplete rationality.

**PAYOFF DOMINANCE**

Undoubtedly the most serious deficiency of Nash equilibrium as a solution concept is its systematic indeterminacy, arising from multiplicity of equilibrium points. We have already encountered this problem in the Assurance Game in Figure 12.1. Equilibrium points are convincing solutions to strictly competitive (finite, two-person, zero-sum) games in which one player's gain is invariably equal to the co-player's loss, because in such games, if there are multiple equilibrium points, then they are invariably *equivalent* and *interchangeable*. Two equilibrium points  $(E, F)$  and  $(E', F')$  are equivalent if the payoffs are the same in each, and they are interchangeable if  $(E, F')$  and  $(E', F)$  are also equilibrium points. Then it *makes no difference* which equilibrium strategies the players choose, because (it is easy to prove) the outcome is invariably an equilibrium point with the same payoffs. Figure 12.4, for example, shows a typical strictly competitive game. Following the convention for strictly competitive games, only Player I's payoffs are shown – Player II's are simply the negatives of these, Player I's gains being Player II's losses. The four outcomes  $(C, D)$ ,  $(C, E)$ ,  $(D, D)$ ,  $(D, E)$ , are all equilibrium points, and as long each player chooses an equilibrium strategy –  $C$  or  $D$  for Player I and  $D$  or  $E$  for Player II – the strategies are in equilibrium and payoffs are the same: 2 units to Player I and  $-2$  to Player II.

		<b>II</b>		
		<b>C</b>	<b>D</b>	<b>E</b>
<b>C</b>	<b>4</b>	<b>2</b>	<b>2</b>	
<b>D</b>	<b>7</b>	<b>2</b>	<b>2</b>	
<b>E</b>	<b>3</b>	<b>0</b>	<b>1</b>	

*Figure 12.4* Strictly competitive game with multiple equilibrium points.

The classic solution of strictly competitive games is widely accepted, although occasional sceptical voices have been raised against it from the beginning (see especially Ellsberg, 1956). But games that are not strictly competitive often have multiple equilibrium points that are non-equivalent and non-interchangeable, and as a consequence lack determinate equilibrium solutions. The Assurance Game

shown in Figure 12.1 is a case in point: there are pure-strategy equilibrium points at  $(C, C)$  and  $(D, D)$ , but they are non-equivalent because the payoffs are different in each, and non-interchangeable because if Player I chooses  $C$  and Player II  $D$ , for example, then the resulting  $(C, D)$  outcome is not an equilibrium point. In the Assurance Game, both players obviously prefer  $(C, C)$  to  $(D, D)$ , but the Nash equilibrium criterion, on its own, is indeterminate, and games typically have several equilibrium points.

In their influential book, *A General Theory of Equilibrium Selection in Games*, Harsanyi and Selten (1988), suggested a principle that they called the *payoff-dominance principle* to help solve the problem of Nash indeterminacy. (They also suggested a secondary *risk-dominance principle* that is not directly relevant to this discussion.) Given two equilibrium points in a game, one payoff-dominates (or Pareto-dominates) the other if it gives every player a higher payoff than the other. In the Assurance Game of Figure 12.1,  $(C, C)$  payoff-dominates  $(D, D)$ , because it gives both players higher payoffs. The payoff-dominance principle is the proposition that if one equilibrium point payoff-dominates all others in a game, then rational players will choose the strategies corresponding to it. Harsanyi and Selten proposed that the payoff-dominance principle should be regarded as part of every player's 'concept of rationality' and should be common knowledge among the players.

Payoff dominance is the leading principle of equilibrium selection, and its intuitive force is generally acknowledged (Colman, 1997; Colman and Bacharach, 1997; Crawford and Haller, 1990; Lewis, 1969; Sugden, 1995). But why is it intuitively compelling, and why should rational players use it? To expose the phenomenon in its starkest form, let us consider the Hi-Lo Matching Game shown in Figure 12.5(a).

		<b>II</b>				<b>II</b>	
		<b>C</b>	<b>D</b>			<b>C</b>	<b>D</b>
<b>I</b>	<b>C</b>	4, 4	0, 0	<b>I</b>	<b>C</b>	4, 4	0, 5
	<b>D</b>	0, 0	3, 3		<b>D</b>	0, 0	3, 3
<b>(a)</b>				<b>(b)</b>			

Figure 12.5 (a) Hi-Lo Matching Game. (b) Modified Hi-Lo Matching Game.

The Hi-Lo Matching Game is really just a simplified version of Sen's Assurance Game with out-of-equilibrium payoffs stripped out. There are two pure-strategy equilibrium points at  $(C, C)$  and  $(D, D)$ , and  $(C, C)$  obviously payoff-dominates  $(D, D)$ . In spite of this, the standard common knowledge and rationality assumptions of game theory provide no rational justification for preferring  $C$  to  $D$ . That is why Harsanyi and Selten (1988) had to introduce the payoff-dominance principle as an axiom. A rational player would choose  $C$  rather than  $D$  if there were a reason to expect the co-player to choose  $C$ , but there is no such reason, because the co-player faces exactly the same quandary, and this leads

to an infinite regress. It is impossible to derive a mandate for choosing  $C$  from the common knowledge and rationality assumptions. This point is widely misunderstood, presumably because choosing  $C$  seems intuitively rational, and it is therefore worth pausing to discuss two common fallacies.

First, it is tempting to argue, as a referee of a journal article that I submitted once did, that  $C$  yields a payoff of 4 or zero, whereas  $D$  yields 3 or zero, therefore a rational player should choose  $C$  to maximize expected utility under uncertainty. This argument is easily refuted by considering the modified version in Figure 12.5(b). For Player I,  $C$  yields 4 or zero, whereas  $D$  yields 3 or zero, just as in the original version in Figure 12.5(a), but it is obvious that no rational Player I would choose  $C$ . In the modified game, Player II receives a higher payoff by choosing  $D$  than  $C$  against *both* of Player I's strategies and will therefore certainly choose  $D$ , if rational. By the transparency of reason, Player I knows this and, if rational, will therefore choose  $D$  in order to secure a payoff of 3 rather than zero. The only rational outcome in Figure 12.5(b), and of course the only Nash equilibrium, is  $(D, D)$ . The second fallacy is to assume that the Bayesian principle of insufficient reason can be used to assign equal probabilities to the co-player's strategies. According to this principle, we are entitled to consider two events as equally probable if we have no reason to consider one more probable than the other. If this were valid, then in the original Hi-Lo Matching Game of Figure 12.5(a), Player I might assume that Player II's strategies are equally probable, in which case it would certainly be rational for Player I to choose  $C$ , because it would yield a (subjective) expected utility of  $(\frac{1}{2} \times 4) + (\frac{1}{2} \times 0) = 2$ , whereas the expected utility of a  $D$  choice would be  $(\frac{1}{2} \times 0) + (\frac{1}{2} \times 3) = 1\frac{1}{2}$ . But then, then by the transparency of reason, Player II would anticipate Player I's  $C$  strategy and, to maximize utility, would also choose  $C$  – with certainty. Player I would anticipate *this*, and we have a contradiction. Starting from the assumption that Player II's  $C$  and  $D$  strategies are equally probable, we have proved that their probabilities are 1 and 0 respectively. From the assumption that Player II is equally likely to choose  $C$  or  $D$ , we have proved that Player II is certain to choose  $C$  – *reductio ad absurdum*. No method of assigning subjective probabilities to co-players' strategies yields up a valid reason for choosing  $C$  in the Hi-Lo Matching Game.

There is simply no way of justifying the payoff-dominance principle in the Hi-Lo Matching Game, or in the Assurance Game of Figure 12.1, or in any other game, on the basis of the standard knowledge and rationality assumptions. Surprisingly, payoff dominance is not rationally justifiable, in spite of its intuitive appeal. But there is experimental evidence to show that human decision makers coordinate on payoff-dominant solutions with considerable ease, even in matching games with far more strategies than the Hi-Lo Matching Game (Mehta, Starmer, and Sugden, 1994) and that players are strongly influenced by payoff dominance in more complex games as well (Cooper, *et al.*, 1990). Various modifications of the assumptions have been suggested to account for this phenomenon, the most prominent being team reasoning and Stackelberg reasoning.

### **Team reasoning and Stackelberg reasoning**

Team reasoning (Sugden, 1993; Bacharach, 1999) is based on the idea that, in certain circumstances, players act to maximize their collective payoff, relative to their knowledge and beliefs, rather than their individual payoffs. A team-reasoning player first identifies a profile of strategy choices that maximizes the collective payoff of the players, and if this profile is unique, plays the corresponding

individual strategy that is a component of it. This involves a radical revision of the standard assumptions, according to which decision makers maximize individual payoffs. But examples of joint enterprises abound in which people appear to be motivated by collective rather than individual interests. On sports fields and battlefields, in commercial companies, university departments, and families, anecdotal evidence suggests that people sometimes choose actions according to what is good for 'us', though their individual preferences may not coincide with the collective interest. In some circumstances de-individuation may even occur, with people tending to lose their sense of personal identity and accountability (Colman, 1991; Dipboye, 1977; Zimbardo, 1969). Team reasoning leads naturally to the selection of payoff-dominant equilibrium points such as  $(C, C)$  in Figures 1 and 5(a). Experimental research has confirmed the intuition that there are circumstances in which decision makers prefer outcomes that maximize collective payoffs. Park and Colman (2001) reported an experiment in which 50 participants were presented with vignettes designed to elicit various social value orientations. In two vignettes, describing scenarios in which payoffs go into a common pool and the participants benefit jointly from cooperative outcomes, preferences were strongly and significantly biased towards joint rather than individual payoff maximization, and qualitative analysis of verbally expressed reasons for choices indicated that team-reasoning explanations, alluding directly or indirectly to collective payoff maximization, were invariably given in these two vignettes.

A second suggestion for explaining the payoff-dominance phenomenon is Stackelberg reasoning, suggested by Colman and Bacharach (1997). The assumption here is that players choose strategies that maximize their individual payoffs on the assumption that any choice will invariably be met by the co-player's best reply, as if players could read each others' minds. In the Hi-Lo Matching Game shown in Figure 12.5(a), for example, if the players assume that any strategy will always be correctly anticipated by the co-player, then Player I might reason that a  $C$  choice will be met with a  $C$  counter-strategy (because Player II prefers 4 to zero), and  $D$  will be met with by  $D$  (because Player II prefers 3 to zero). Player I would receive a payoff of 4 in the first case and 3 in the second, hence if the choice could be anticipated by Player II, then a rational Player I would choose  $C$ . If both players reason like this, then they choose the payoff-dominant  $(C, C)$  equilibrium point in the Hi-Lo Matching Game in Figure 12.5(a), or in the Assurance Game in Figure 12.1. Colman and Bacharach proved that Stackelberg reasoning results in coordination on a payoff-dominant equilibrium point in any game that has one. In some games, Stackelberg reasoning yields strategies that are not in equilibrium, and such games are not Stackelberg soluble. Stackelberg reasoning functions as a strategy generator and Nash equilibrium as a strategy filter.

Is there any evidence that people do, in fact reason in this way? Colman and Stirk (1998) reported an experiment in which 100 randomly paired players made one-off strategy choices in 12 different  $2 \times 2$  games. Nine of the games were Stackelberg soluble and three were not. The players were motivated by substantial monetary payoffs. A significant bias towards Stackelberg strategies emerged in all Stackelberg-soluble games, with large effect sizes. In non-Stackelberg-soluble games, very small and non-significant effects were found. A protocol analysis of players' stated reasons for choices revealed joint payoff maximization to be a reason significantly more frequently in the Stackelberg-soluble games. These results provide strong evidence that Stackelberg reasoning influences players, at least in  $2 \times 2$  games. Both Stackelberg reasoning and team reasoning probably contribute to the payoff-dominance phenomenon, and both require revision of the underlying assumptions of game theory.

# STRATEGIC DOMINANCE

The concept of strategic dominance is illustrated in the familiar Prisoner's Dilemma Game (Figure 12.6). Each player chooses between cooperating (C) and defecting (D). Each receives a higher payoff from defecting than cooperating, irrespective of whether the other player cooperates or defects, but each receives a higher payoff if both cooperate than if both defect. The game's name derives from an interpretation devised by Albert W. Tucker for a seminar at Stanford University Psychology Department in 1950, a few months after the game was discovered at the RAND Corporation in Santa Barbara, California. Two people, charged with joint involvement in a serious crime, are arrested, prevented from communicating with each other, and interrogated separately. The police have insufficient information for a successful prosecution unless at least one of the prisoners discloses incriminating evidence. Each prisoner has to choose between cooperating with the other prisoner by concealing the incriminating evidence (C) and defecting by disclosing it (D). If both cooperate, both are acquitted (the second-best payoff for each); if both defect, both are convicted (the third-best payoff for each); and if only one defects while the other cooperates, then according to a plea bargain offered to them, the one who defects is acquitted with a reward for helping the police (the best possible payoff), and the one who conceals the evidence receives an especially heavy sentence (the worst possible payoff).

		II	
		C	D
I	C	3, 3	0, 5
	D	5, 0	2, 2

Figure 12.6 Prisoner's Dilemma Game.

The Prisoner's Dilemma is ubiquitous in everyday strategic interaction. It is a standard model of bilateral arms races (Brams, 1976, pp. 81-91) and of many similar interactions involving cooperation and competition, trust and suspicion. Rapoport (1962) found a poignant example in Puccini's opera Tosca, after Tosca's lover has been condemned to death, when the police chief, Scarpia, offers to save his life by ordering the firing squad to use blank cartridges if Tosca agrees to have sex with him. Tosca and Scarpia each face a choice between keeping their side of the bargain and double-crossing the other player, and the strategic structure corresponds to Figure 12.6. In the opera, Tosca and Scarpia both defect: Tosca stabs Scarpia as he is about to grab her, and Scarpia turns out not to have ordered the firing squad to use blank cartridges. The diabolically frustrating Prisoner's Dilemma game models cooperation versus competition, trust versus suspicion, and individualism versus collectivism. Multi-person social dilemmas with the same underlying strategic properties have also

been extensively studied (see reviews by Colman, 1995, chaps 6, 7, 9; Foddy *et al.*, 1999; Van Lange *et al.*, 1992).

There is a certain logical inevitability about the unfolding tragedy in *Tosca*, and the unravelling of certain peace processes in political trouble spots. The reason is that *D* is a *dominant strategy* for both players, in the sense that each player receives a higher payoff from defecting than from cooperating, irrespective of the co-player's choice. If Player II chooses *C*, then Player I receives a higher payoff by playing *D* than *C* (Player I gets 5 rather than 3), and similarly if Player II chooses *D* (Player I gets 2 rather than 0). Defecting is thus the unconditionally best strategy for Player I, and by symmetry, the same applies to Player II. It is best for each player to defect whatever the other player does, but this entails a paradox, because each does better if both cooperate than if both defect. The (*D*, *D*) outcome, corresponding to dominant strategies, is the only Nash equilibrium, but if both players choose their dominated *C* strategies, then the outcome (*C*, *C*) is better for each. Rationality is thus self-defeating in the Prisoner's Dilemma Game.

In spite of strategic dominance, experimental evidence (reviewed by Colman, 1995, chap 7) has shown that players frequently cooperate, to their mutual advantage. For example, in the largest experiment, in which the Prisoner's Dilemma Game was played repeatedly, approximately half of all strategy choices were cooperative (Rapoport and Chamah, 1965), and even in experiments using one-shot games, a substantial minority of choices tend to be cooperative (e.g., Deutsch, 1960; Shafir and Tversky, 1992). Real players earn higher payoffs than they would have done had they followed the rational prescriptions of game theory. This is paradoxical, because rationality is defined as expected utility maximization.

As a solution concept, strategic dominance is warmly accepted by decision theorists and game theorists and, like motherhood and apple pie, it is seldom questioned. Its persuasive force seems overwhelming when dominance is strong – when a strategy yields a strictly better payoff than any alternative against all possible counter-strategies, as in the Prisoner's Dilemma Game. In those circumstances, it seems obvious that it is the uniquely rational way of acting. Attempts to justify cooperation in the one-shot Prisoner's Dilemma Game are laughed to scorn by game theorists (see Binmore, 1994, chap. 3). Even if a strategy only weakly dominates all other strategies – if it is at least as good against all counter-strategies and strictly better against at least one – that seems a knock-down argument for choosing it. But there are games that pose bigger challenges to the strategic dominance principle than the Prisoner's Dilemma Game.

The most notorious is Newcomb's problem, discovered by William A. Newcomb and published by Robert Nozick (1969), with the footnote: 'It is a beautiful problem. I wish it were mine'. For a detailed analysis of the problem from various angles, see Campbell and Sowden (1985). Here is a simple version of it. On the table is a transparent box containing £1000 and an opaque box containing either £1m or nothing. A player is offered the choice of taking both boxes or only the opaque box. The player is told, and believes, that a behavioural predictor, such as a sophisticated computer programmed with psychological information about the player, has already put £1m in the opaque box if and only if it has predicted that the player will take only that box, and not the transparent box as well. The player knows that the predictor is always correct or (if it is more credible) correct in 95 per cent of cases, say, although the exact figure is not critical. The problem is summarized in Figure 12.7.

		Predictor	
		Add £1m	No £1m
Player	One box	£1m	£0
	Both boxes	£1m + £1000	£1000

Figure 12.7 Newcomb's problem.

The predictor's payoffs are not shown in Figure 12.7, because they are assumed to play no part in the dilemma. The strategy of taking both boxes is strongly dominant, because it yields more than taking only one box against both of the predictor's counter-strategies – if the predictor has added £1m to the opaque box, then it yields £1m + £1000 rather than just £1000, and if the predictor has not added £1m to the opaque box, then it yields £1000 rather than nothing. That might seem to settle the matter, but the problem is that expected utility theory appears to require a rational player to take only one box. If the player takes both boxes, then the predictor will probably have left the opaque box empty, therefore the player will probably get only £1000, whereas if the player takes only one box, then the predictor will probably have left £1m in it. The player will therefore probably receive a much higher payoff by taking only one box. Thus seemingly irrefutable arguments appear to justify both the one-box and the two-box strategies – expected utility theory appears to justify taking only one box, and strategic dominance taking both. Most people, after pondering the problem, consider the rational strategy to be perfectly obvious, but they are divided as to which strategy that is (Nozick, 1969). Rational players, by definition, maximize expected utility. Newcomb's problem represents a clash between two different ways of reasoning about expected utility, called *evidential* and *causal* expected utility respectively (Nozick, 1993, pp. 41-63). A player who maximizes evidential expected utility uses standard conditional probabilities (such as *the probability that the opaque box contains £1m given that it is chosen*) and infers that players who take only one box usually earn a fortune, whereas people who take both boxes usually do not. If you are a one-box type of person, then the conditional probability that the predictor has put £1m in it is high, and it follows that taking only one box is likely to net you a fortune, whereas if you are a two-box type of person, then there is probably nothing in the opaque box. According to evidential reasoning, the one-box strategy maximizes conditional expected utility and is therefore rational. A player who maximizes causal expected utility uses causally conditional probabilities, reasoning that taking only one box cannot cause £1m to appear in it, if it is not there already, therefore causal expected utility is maximized by taking both boxes. This is often illustrated with the *smoking gene* example. The statistician Ronald A. Fisher (1959) argued that cigarette smoking is a form of behaviour caused by a gene that also causes lung cancer. If this is true, then rational smokers should consider their smoking behaviour as unwelcome evidence that they probably have the gene and are likely to get lung cancer, but it would be futile for them to give up smoking on that account, because doing so would not cause the gene to disappear. On this view, it is

equally futile to take only the opaque box in Newcomb’s problem, because that cannot make money appear in it – the two-box strategy maximizes causal expected utility.

Although causal rather than evidential reasoning obviously applies in the smoking gene case, both evidential and causal reasoning can be defended in the right circumstances (Nozick, 1993). After Newcomb’s problem was aired in *Scientific American* magazine in 1973, no fewer than 148 people wrote to the magazine, and 60 per cent of them favoured the one-box strategy (Nozick, 1974). Experimental evidence (Anand, 1990) has confirmed that many intelligent and well educated people favour the one-box strategy. Human decision makers evidently do not consider strong strategic dominance to be a knock-down argument in Newcomb’s problem, and evidential reasoning is has also been found in other problems (Quattrone and Tversky, 1984).

In some games, strategic dominance is more obviously irrational. These are games in which players’ strategies are not independent each other. I shall illustrate this with a game described in *Luke 10: 30-37* that I shall call the Good Samaritan Game. An onlooker comes across a victim of a mugging. The onlooker has two available strategies, namely to help the victim, like the Good Samaritan, or pass by on the other side, like the Levite. The mugger is still lurking in the vicinity and may violently assault anyone who intervenes. The onlooker’s utilities, taking into account the pain, suffering, and humiliation associated with being mugged (valued at –10 units of utility) and the warm glow that would arise from acting compassionately (worth 5 units of utility), are as shown in the Figure 12.8.

		<b>Mugger</b>	
		<b>Assault</b>	<b>Leave</b>
<b>On-looker</b>	<b>Help</b>	<b>–5</b>	<b>5</b>
	<b>Pass</b>	<b>–10</b>	<b>0</b>

*Figure 12.8* Good Samaritan Game.

The onlooker’s payoff from helping the victim is higher than the payoff from passing by on the other side *whether or not the mugger chooses the assaulting strategy*. This suggests that helping the victim is a strongly dominant strategy and must therefore be unconditionally best for a rational onlooker. The onlooker may reason as follows.

The mugger may assault me whether or not I help the victim. If I’m to be assaulted, then I’m better off helping the victim than passing by on the other side, because then at least my bruises will not be in vain. On the other hand, if the mugger leaves me alone, I’m also better off helping the victim than passing by, because then I’ll have done something good. I receive a higher payoff from helping in either case, therefore it must be rational for me to help the victim.

This argument is seductive but (alas) subtly flawed, because helping the victim may cause the onlooker to be assaulted, and passing by may result in the onlooker being left alone. The problem here is that the condition of *act independence* does not hold. As mentioned near the beginning of the

section on Nash equilibrium, an explicit assumption of non-cooperative game theory is that the players choose their strategies independently. In the Good Samaritan Game, the onlooker's actions are not independent of the mugger's and may have the capacity to influence the mugger's. In the extreme, if the onlooker knew that helping the victim would certainly elicit an assault from the mugger and passing by would certainly not, then the outcomes (*Help, Assault*) and (*Pass, Leave*) on the main diagonal of the payoff matrix would be the only relevant ones and, given the onlooker's utility function, passing by would seem prudent, because (*Pass, Leave*) yields a better payoff to the onlooker than (*Help, Assault*) does.

Formally, according to evidential expected utility reasoning, if the conditional probabilities of the onlooker being assaulted are 1 if the onlooker helps the victim and 0 if the onlooker passes by, then  $Prob(Assault | Help) = 1$ ,  $Prob(Leave | Help) = 0$ ,  $Prob(Assault | Pass) = 0$ ,  $Prob(Leave | Pass) = 1$ , and the conditional expected utility (CEU) of helping and of passing by can be calculated from the payoff matrix shown in Figure 12.8 using standard rules of probability:

$$CEU(Help) = [-5 \times Prob(Assault | Help)] + [5 \times Prob(Leave | Help)] = -5 + 0 = -5$$

$$CEU(Pass) = [-10 \times Prob(Assault | Pass)] + [0 \times Prob(Leave | Pass)] = 0 + 0 = 0$$

It is clear that passing by on the other side yields a higher conditional expected utility than helping. This shows why the argument from strong strategic dominance is fallacious when act independence does not hold. In my opinion, it also exposes the crux of Newcomb's problem. If the actions of the player and the predictor in Figure 12.7 are truly independent, then the dominance argument is valid and a rational player should take both boxes, whereas if act independence is violated by the specification of the problem, then evidential reasoning based on conditional expected utilities apply, and a rational player should take only the opaque box. Differences of opinion about Newcomb's problem seem to me to arise from disagreements about whether or not the specification of the problem implies act independence – if the predictor has paranormal powers, for example, then it might not.

## STABLE SETS AND THE CORE

This chapter has been concerned mainly with non-cooperative games, but a few comments about cooperative games will help to place the earlier sections in perspective. In cooperative games, players are not constrained to choose strategies independently but are able to negotiate coalitions based on binding and enforceable agreements with one another.

The modern history of game theory is often traced to the publication of *Games and Economic Behavior* by von Neumann and Morgenstern (1944). These pioneering theorists failed to derive a generalized equilibrium concept and devoted most of their attention to cooperative games, which they modelled in terms of different ways of dividing a payoff among the players, but it is fair to say that cooperative game theory is still poorly understood. Divisions of the payoff that satisfy conditions of individual and collective rationality are called *imputations*. These are divisions in which the individual players receive at least as much as they could guarantee for themselves by acting independently, and the grand coalition of all players receives the whole payoff, so that nothing is

wasted. Von Neumann and Morgenstern struggled to find a solution concept that would prescribe a uniquely rational imputation for every cooperative game, but they succeeded only in showing that certain *stable sets* of imputations were rational in a specially defined sense. They interpreted these stable sets as ‘standards of behaviour’ governed by social and moral conventions, providing no rational criteria for choosing particular imputations as solutions, and it subsequently transpired that there are games with no stable sets.

Nash (1950b) developed a more radical technique of modelling cooperative games as non-cooperative games and then applying his equilibrium solution concept. This proposed unification of game theory became known as the Nash programme, and it attracted considerable support, but its edge was blunted by Nash indeterminacy, as it emerged that reformulated non-cooperative games typically have multiple equilibrium points.

The most influential solution concept for cooperative games is the *core*, a natural extension of the imputation concept discovered by a postgraduate student (Gillies, 1953). The core of a cooperative game is an imputation in which every possible coalition of players receives at least as much as it could guarantee for itself by acting collectively. This would provide a convincing solution concept were it not for the unfortunate fact that many cooperative games gave empty cores, in the sense that no imputation satisfies all three requirements of individual, coalition, and collective rationality. The simplest example of this is the game of dividing a fixed sum of money among three players by majority vote. For every possible imputation, there is a coalition with the motive and the power to overturn it. For example, if Players I and II agree to take half the payoff each, then Player III can form a coalition with Player I in which Player I gets 60 per cent and Player III 40 per cent, and this coalition has the power to impose its will. I discussed a literary example of an empty core, taken from Harold Pinter’s play, *The Caretaker*, in Colman (1995, pp. 169-75).

## CONCLUSIONS

A noted game theorist once warned that ‘the foundations of game theory are a morass into which it is not wise to wander if you have some place you want to get to in a hurry’ (Binmore, 1994, p. 142). This is a salutary warning, but in this volume we are not in a hurry, and the foundations need to be secured to understand reasoning about interactive decision making. What can be seen clearly through the muddy waters of the morass is that the foundations are in need of maintenance work. The foundations should support whatever theoretical superstructure is required, but in their current state they cannot even support the payoff-dominance principle, leaving unexplained the intuitively obvious solutions to games such as the Hi-Lo Matching Game shown in Figure 12.5(a).

What is surprising and impressive is that we can make any progress at all in understanding reasoning in games. Instrumental rationality, which has a clear and simple interpretation in individual decision making, and can be defined rigorously in terms of expected utility maximization, is difficult to apply in interactive decision making, where the outcomes of a player’s decisions depend partly on the decisions of other players. In spite of this, some progress has been made. I have outlined the major solution concepts, and I have discussed some of the problems that they raise.

The leading solution concept for non-cooperative games is undoubtedly Nash equilibrium. It follows logically from the standard knowledge and rationality assumptions of game theory that any uniquely

rational solution to a game must be a Nash equilibrium. In the special case of strictly competitive games, this yields determinate and persuasive solutions, but in other classes of games, it narrows down the search for a rational solution without generally yielding determinate solutions. Application of the Nash equilibrium solution concept therefore leaves us with a residual problem of equilibrium selection.

The most compelling solution concept of all is strategic dominance. Nothing seems more obvious than the rationality of choosing a strategy that yields a higher payoff than any other against every possible counter-strategy or combination of counter-strategies. If one course of action is unconditionally best in all circumstances that might arise, then it seems obvious that a rational player will invariably choose it. Although this may seem controversial in certain special cases, such as Newcomb's problem, it is unassailable in games that clearly satisfy the condition of act independence. Provided that the players' actions are truly independent, this is therefore a good place to start when seeking a solution to a game. If a player has strategies that are even weakly dominated by others, then delete the dominated strategies. In the resulting reduced game, it may turn out that another player has dominated strategies that can be deleted, and if this process of iterated deletion of dominated strategies is continued, it sometimes converges on a unique solution. If act independence does not hold, then before beginning this process, the game should first be reformulated as a sequential game, with each player having perfect knowledge of any preceding move(s). It is useful to know that when simultaneous-choice games are reformulated as sequential games in this way, they often become soluble by iterated deletion of weakly dominated strategies.

The strategic dominance solution concept is not always helpful, however. In the Assurance Game of Figure 12.1 and the Hi-Lo Matching Game of Figure 12.5(a), for example, it gets us nowhere. In such cases, ad-hoc methods of equilibrium selection such as the payoff-dominance principle may have to be applied. If a game with multiple equilibrium points has one yielding a higher payoff to every player than any other, then it seems obvious that rational players will play their parts in it. Team reasoning and Stackelberg reasoning provide possible mechanisms to explain the payoff-dominance principle, but the principle is not implied by the standard knowledge and rationality assumptions. Harsanyi and Selten (1988) introduced it into their general theory of equilibrium selection rather inelegantly as an axiom, though they did so with some reluctance (see their comments on pp. 355-63). In time, the fundamental assumptions of game theory may be amended to imply payoff dominance. But even that will not necessarily help us to understand strategic interactions in cooperative games in which binding and enforceable coalitions can be negotiated. When analysing a cooperative game, we can only hope that it has a core, because if it does not, then it may lack any determinate solution. Perhaps some games simply do not have uniquely rational solutions. If that is the case, it would be good to have a rigorous proof of it. Let us put on our wellington boots –there is work to be done.

## **ACKNOWLEDGEMENT**

Preparation of this article was facilitated by a period of study leave granted to me by the University of Leicester.

## **REFERENCES**

- Anand, P. (1990) 'Two types of utility: An experimental investigation into the prevalence of causal and evidential utility maximisation', *Greek Economic Review*, 12: 58-74.
- Aumann, R.J. (1976) 'Agreeing to disagree', *Annals of Statistics*, 4: 1236-9.
- Bacharach, M. (1987) 'A theory of rational decision in games' *Erkenntnis*, 27: 17-55.
- Bacharach, M. (1999) 'Interactive team reasoning: A contribution to the theory of co-operation', *Research in Economics*, 53: 117-47.
- Binmore, K. (1994) *Playing Fair: Game Theory and the Social Contract Volume I*, Cambridge, MA: MIT Press.
- Brams, S.J. (1976) *Paradoxes in Politics: An Introduction to the Nonobvious in Political Science*, New York: Free Press.
- Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton, NJ: Russell Sage Foundation.
- Campbell, R. and Sowden, L. (eds) (1985) *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem*, Vancouver: University of British Columbia Press.
- Colman, A.M. (1991) 'Crowd psychology in South African murder trials', *American Psychologist*, 46: 1071-9.
- Colman, A.M. (1995) *Game Theory and Its Applications in the Social and Biological Sciences*, 2nd edn, London: Routledge.
- Colman, A.M. (1997) 'Salience and focusing in pure coordination games', *Journal of Economic Methodology*, 4: 61-81.
- Colman, A.M. (in press) 'Cooperation, psychological game theory, and limitations of rationality in social interaction', *Behavioral and Brain Sciences*.
- Colman, A.M. and Bacharach, M. (1997) 'Payoff dominance and the Stackelberg heuristic', *Theory and Decision*, 43: 1-19.
- Colman, A.M. and Stirk, J.A. (1998) 'Stackelberg reasoning in mixed-motive games: An experimental investigation', *Journal of Economic Psychology*, 19: 279-93.
- Cooper, R.W., DeJong, D.V., Forsythe, R., and Ross, T.W. (1990) 'Selection criteria in coordination games: Some experimental results', *American Economic Review*, 80: 218-33.
- Crawford, V.P. and Haller, H. (1990) 'Learning how to cooperate: Optimal play in repeated coordination games', *Econometrica*, 58: 571-95.
- Deutsch, M. (1960) 'The effect of motivational orientation upon threat and suspicion', *Human Relations*, 13: 123-39.
- Dipboye, R.L. (1977) 'Alternative approaches to deindividuation', *Psychological Bulletin*, 85: 1057-75.
- Ellsberg, D. (1956) 'Theory of the reluctant duellist', *American Economic Review*, 46: 909-23.
- Fisher, R.A. (1959) *Smoking: The Cancer Controversy, Some Attempts to Assess the Controversy*, Edinburgh: Oliver and Boyd.
- Foddy, M., Smithson, M., Schneider, S., and Hogg, M. (eds) (1999) *Resolving Social Dilemmas: Dynamic, Structural, and Intergroup Aspects*, London: Psychology Press.
- Gillies, D.B. (1953) 'Some theorems on  $n$ -person games', Unpublished doctoral dissertation, Princeton University.

- Harsanyi, J.C. (1973) 'Games with randomly distributed payoffs: A new rationale for mixed-strategy equilibrium points', *International Journal of Game Theory*, 2: 1-23.
- Harsanyi, J.C. and Selten, R. (1988) *A General Theory of Equilibrium Selection in Games*, Cambridge, MA: MIT Press.
- Hume, D. (1739-1740/1978) *A Treatise of Human Nature*, 2nd edn, L.A. Selby-Bigge, ed., Oxford: Oxford University Press.
- Kagel, J.H. and Roth, A.E. (eds) (1995) *Handbook of Experimental Economics*, Princeton: Princeton University Press.
- Lewis, D.K. (1969) *Convention: A Philosophical Study*, Cambridge, MA: Harvard University Press.
- Luce, R.D. and Raiffa, H. (1957) *Games and Decisions: Introduction and Critical Survey*, New York: Wiley.
- Mehta, J., Starmer, C., and Sugden, R. (1994) 'Focal points in pure coordination games: An experimental investigation', *Theory and Decision*, 36: 163-85.
- Nash, J.F. (1950a) 'Equilibrium points in  $n$ -person games', *Proceedings of the National Academy of Sciences, USA*, 36: 48-9.
- Nash, J.F. (1950b) 'The bargaining problem', *Econometrica*, 18: 155-62.
- Nash, J.F. (1951) 'Non-cooperative games', *Annals of Mathematics*, 54: 286-95.
- Nozick, R. (1969) 'Newcomb's problem and two principles of choice', in N. Rescher (ed.) *Essays in Honor of Carl Hempel* (pp. 114-46), Dordrecht: D. Reidl.
- Nozick, R. (1974) 'Reflections on Newcomb's paradox', *Scientific American*, 230(3): 102-8.
- Nozick, R. (1993) *The Nature of Rationality*, Princeton, NJ: Princeton University Press.
- Park, J.R. and Colman, A.M. (2001) 'Team reasoning: An experimental investigation', paper presented at the V Conference of the Society for the Advancement of Economic Theory, Ischia, 2-8 July.
- Pruitt, D.G. and Kimmel, M.J. (1977) 'Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future', *Annual Review of Psychology*, 28: 363-92.
- Quattrone, G.A. and Tversky, A. (1984) 'Causal versus diagnostic contingencies: On self-deception and the voter's illusion', *Journal of Personality and Social Psychology*, 46: 237-48.
- Rapoport, A. (1962) 'The use and misuse of game theory', *Scientific American*, 207(6): 108-18.
- Rapoport, A. and Chammah, A.M. (1965) *Prisoner's Dilemma: A Study in Conflict and Cooperation*, Ann Arbor, MI: University of Michigan Press.
- Selten, R. (1965) 'Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit', *Zeitschrift für die gesamte Staatswissenschaft*, 121: 301-24, 667-89.
- Selten, R. (1975) 'Re-examination of the perfectness concept for equilibrium points in extensive games', *International Journal of Game Theory*, 4: 25-55.
- Sen, A.K. (1969) 'A game-theoretic analysis of theories of collectivism in allocation', in T. Majumdar (ed.), *Growth and Choice: Essays in Honour of U. N. Ghosal* (pp. 1-17), Calcutta: Oxford University Press.
- Shafir, E. and Tversky, A. (1992) 'Thinking through uncertainty: Nonconsequential reasoning and choice', *Cognitive Psychology*, 24: 449-74.
- Sugden, R. (1991) 'Rational bargaining', In M. Bacharach and S. Hurley (eds), *Foundations of*

*Decision Theory* (pp. 294-315), Oxford: Blackwell.

Sugden, R. (1993) 'Thinking as a team: Towards an explanation of nonselfish behavior', *Social Philosophy and Policy*, 10: 69-89.

Sugden, R. (1995) 'A theory of focal points', *Economic Journal*, 105: 533-50.

Van Lange, P.A.M., Liebrand, W.B.G., Messick, D.M., and Wilke, H.A.M. (1992) 'Social dilemmas: The state of the art', in W.B.G. Liebrand, D.M. Messick, and H.A.M. Wilke (eds), *Social Dilemmas: Theoretical Issues and Research Findings* (pp. 3-28), New York: Pergamon.

Von Neumann, J. and Morgenstern, O. (1944) *Theory of Games and Economic Behavior*, Princeton, NJ: Princeton University Press. [2nd ed., 1947; 3rd ed., 1953]

Williams, B. (1979) 'Internal and external reasons', in R. Harrison (ed.), *Rational Action: Studies in Philosophy and Social Science* (pp. 17-28), Cambridge: Cambridge University Press.

Zimbardo, P.G. (1969) 'The human choice: Individuation, reason, and order, vs deindividuation, impulse, and chaos', *Nebraska Symposium on Motivation*, 17: 237-307.