BEHAVIORAL AND BRAIN SCIENCES (2003) **26,** 139–198
Printed in the United States of America

Cooperation, psychological game theory, and limitations of rationality in social interaction

Andrew M. Colman
School of Psychology, University of Leicester, Leicester LE1 7RH, United Kingdom
amc@le.ac.uk **www.le.ac.uk/home/amc**

Abstract: Rational choice theory enjoys unprecedented popularity and influence in the behavioral and social sciences, but it generates intractable problems when applied to socially interactive decisions. In individual decisions, instrumental rationality is defined in terms of expected utility maximization. This becomes problematic in interactive decisions, when individuals have only partial control over the outcomes, because expected utility maximization is undefined in the absence of assumptions about how the other participants will behave. Game theory therefore incorporates not only rationality but also common knowledge assumptions, enabling players to anticipate their co-players' strategies. Under these assumptions, disparate anomalies emerge. Instrumental rationality, conventionally interpreted, fails to explain intuitively obvious features of human interaction, yields predictions starkly at variance with experimental findings, and breaks down completely in certain cases. In particular, focal point selection in pure coordination games is inexplicable, though it is easily achieved in practice; the intuitively compelling payoff-dominance principle lacks rational justification; rationality in social dilemmas is self-defeating; a key solution concept for cooperative coalition games is frequently inapplicable; and rational choice in certain sequential games generates contradictions. In experiments, human players behave more cooperatively and receive higher payoffs than strict rationality would permit. Orthodox conceptions of rationality are evidently internally deficient and inadequate for explaining human interaction. *Psychological game theory,* based on nonstandard assumptions, is required to solve these problems, and some suggestions along these lines have already been put forward.

Keywords: backward induction; Centipede game; common knowledge; cooperation; epistemic reasoning; game theory; payoff dominance; pure coordination game; rational choice theory; social dilemma

Were such things here as we do speak about?
Or have we eaten on the insane root
That takes the reason prisoner?
—*Macbeth* (I.iii.84)

1. Introduction

Is rational social interaction possible? This may seem a surprising question, given the Apollonian flavor of the contemporary behavioral and social sciences. Rational choice theory (RCT) is the cornerstone of neoclassical economics (Arrow et al. 1996; Elster 1986; Sugden 1991b). In political science, RCT began to mushroom after the publication of *Social Choice and Individual Values* (Arrow 1963) and transformed the discipline within a few decades (Friedman 1996; Green & Shapiro 1994; Ordeshook 1986). In sociology, Weber's (1922/1968) analyses of law and economics as models of rationality prepared the ground for the germination of RCT ideas half a century later (Abell 1991; Coleman & Fararo 1992; Hollis 1987; Moser 1990). Theories of behavioral ecology (Dawkins 1989; Krebs & Davies 1987), and particularly, the evolution of social behavior (Maynard Smith 1984), were revolutionized by the introduction of RCT-based game theory in the early 1970s (Maynard Smith & Price 1973); and even jurisprudence has been influenced by RCT (Raz 2000).

1.1. Rationality in psychology

In psychology, the picture is admittedly more complex. Since the publication of Freud's earliest metapsychological writings, and in particular his adumbration of the distinction between two principles of mental functioning, the

ANDREW M. COLMAN is Professor of Psychology at the University of Leicester, UK. He received his Ph.D. from Rhodes University in South Africa, and he taught at Cape Town and Rhodes Universities before emigrating to England in 1970. His publications include more than 120 articles and chapters, on cooperative reasoning in games and other topics, and several books, the most relevant of which are *Game Theory and its Applications in the Social and Biological Sciences* (2nd edn, 1995) and an edited volume entitled *Cooperation and Competition in Humans and Animals* (1982). His latest book is the Oxford *Dictionary of Psychology* (2001).

© 2003 Cambridge University Press 0140-525X/03 $12.50

reality principle and the *pleasure principle* (Freud 1911) – only the first of which he believed to be functionally rational – psychologists have paid particular attention to irrational aspects of thought and behavior. But psychologists have generally assumed, usually tacitly, that rationality is normal, whereas irrationality is, in some sense, abnormal or pathological.[1]

1.2. Bounded rationality

This article is not concerned merely with the accuracy of RCT in predicting human behavior. The concept of *bounded rationality* (Simon 1957) has been widely accepted and corroborated by experimental evidence. Our bounded rationality obliges us to use rough-and-ready rules of thumb (*heuristics*) that can lead to predictable errors and judgmental biases, many of which have been investigated empirically (Bell et al. 1988; Kahneman et al. 1982), but that allow us to solve problems quickly and efficiently (Gigerenzer & Goldstein 1996; Gigerenzer et al. 1999). For example, a simple rule of *win-stay, lose-change* can lead to the evolution of mutually beneficial cooperation in a group of players who are ignorant not only of the payoff structure of the game but even of the fact that they are involved with other players in a strategic interaction (Coleman et al. 1990).

1.3. Evolutionary game theory

Game-theoretic equilibrium points can thus be arrived at by entirely non-rational evolutionary processes. The basic concepts of game theory can be mapped to the elements of the theory of natural selection as follows. Players correspond to individual organisms, strategies to organisms' genotypes, and payoffs to the changes in their Darwinian fitness – the numbers of offspring resembling themselves that they transmit to future generations. In evolutionary game theory interpreted biologically, the players do not choose their strategies (genotypes) rationally or even deliberately, but different profiles of strategies lead to different payoffs, and natural selection mimics deliberate choice. Maynard Smith and Price (1973) introduced the concept of the *evolutionarily stable strategy* (ESS) to handle such games. It is a strategy with the property that if most members of a population adopt it, then no mutant strategy can invade the population by natural selection, and it is therefore the strategy that we should expect to see commonly in nature. An ESS is invariably a Nash equilibrium (see sect. 5.1 below), and therefore a type of game-theoretic solution; but not every Nash equilibrium is an ESS.

Evolutionary game theory deals with social as well as biological evolution. It has been studied intensively since the 1970s, and the theory is well understood (Hofbauer & Sigmund 1998; Samuelson 1997). Even purely analytic studies can solve problems and provide useful insights. A simple example with psychological relevance is an evolutionary model of Antisocial Personality Disorder, based on a multiplayer Chicken game, that provided an explanation for the low but stable prevalence of this disorder in widely diverse societies (Colman & Wilson 1997; see also Colman 1995b).

Evolutionary games have also been studied empirically (Maynard Smith 1984), and above all, computationally, sometimes by running strategies against one another and transmitting copies of these strategies to future generations according to their accumulating payoffs (Axelrod 1984; 1997, Chs. 1, 2; Nowak et al. 1995; Nowak & Sigmund 1993). Evolutionary game theory deals with non-rational strategic interaction driven by mindless adaptive processes resembling trial and error, and it is therefore not directly relevant to this article. Populations of insects, plants, and even computer programs can evolve to game-theoretic equilibrium points, and cooperation can evolve without rational decision making. This article, however, focuses on whether full rationality can be applied to social interaction.

1.4. Outline of the argument

When a decision involves two or more interactive decision makers, each having only partial control over the outcomes, an individual may have no basis for rational choice without strong assumptions about how the other(s) will act. This complicates the picture and leads to problems, in some cases even to the breakdown of the standard concept of rationality.[2] This article brings together a heterogeneous and disparate collection of arguments and evidence suggesting that rationality, conventionally defined, is not characteristic of human social interaction in general.

The remainder of the article is organized as follows. Section 2 outlines the nature of rationality and its formalization. Section 3 focuses more specifically on game theory, and section 4 on game theory's underlying assumptions. Section 5 argues that selection of focal points and of payoff-dominant equilibria is inexplicable according to game-theoretic rationality, and that a key solution concept for cooperative games is not always applicable. Section 6 is devoted to social dilemmas, in which rationality is self-defeating and human decision makers are paradoxically more successful than ideally rational agents. Section 7 deals with backward induction in sequential games where the standard concept of rationality appears incoherent. Section 8 introduces psychological game theory and outlines some nonstandard contributions designed to overcome these problems. Finally, section 9 draws the threads together and summarizes the conclusions.

2. Nature of rationality

What is rationality? Broadly speaking, it involves thinking and behaving reasonably or logically, and it comes in several guises (Manktelow & Over 1993). Rational beliefs are those that are internally consistent,[3] and rational arguments are those that obey the rules of logic. Rational preferences and decisions require more detailed explanation.

2.1. Rational preferences

Suppose a universe of alternatives includes a subset A of alternatives that are available in a particular decision context. Decision theorists generally assume that an agent's rational preferences obey the following conditions.

1. Completeness: For every pair of alternatives a_i and a_j in A, the agent either prefers a_i to a_j, or prefers a_j to a_i, or is indifferent between a_i and a_j.

2. Transitivity: Given alternatives a_i, a_j, and a_k in A, an agent who considers a_i to be at least as preferable as a_j, and a_j at least as preferable as a_k, considers a_i to be at least as preferable as a_k.

3. Context-free ordering: If an agent considers a_i to be at least as preferable as a_j in A, then that agent considers a_i to be at least as preferable as a_j in an enlarged set A' containing all the elements in A plus additional elements from the universe of alternatives.

These three conditions are collectively called the *weak ordering principle* (McClennen 1990, Ch. 2). We had to begin with a subset A to give meaning to the third condition, which would otherwise be implied by the first.[4] Given preferences that satisfy this tripartite principle, a rational decision maker always chooses a maximally preferable alternative (which may not be unique, hence "a" rather than "the"). The formalization of this in expected utility theory will be discussed in sections 2.3 to 2.6 below. Experimental evidence suggests that human decision makers frequently violate the second and third conditions (Doyle et al. 1999; Huber et al. 1982; Slovic & Lichtenstein 1983; Tversky 1969), although they tend to modify their intransitive preferences, at least, when their violations are pointed out to them.

2.2. Rational decisions

Rational decisions or choices are those in which agents act according to their preferences, relative to their knowledge and beliefs at the time of acting. This is *instrumental rationality* (or *means-end rationality*), and it can be traced back to the Scottish Enlightenment writings of David Hume (1739–1740/1978) and Adam Smith (1776/1910). Hume gave the most frequently quoted account of it in his *Treatise of Human Nature* (2.III.iii):

> 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. (Hume 1739–40/1978, pp. 415–16)

Hume conceived of reason as a faculty affording the means for achieving goals that are not themselves afforded by reason. Russell (1954) summed this up lucidly: "'Reason' has a perfectly clear and precise meaning. It signifies the choice of the right means to an end that you wish to achieve. It has nothing whatever to do with the choice of ends" (p. 8).

2.3. Expected utility theory

Formally, decisions that maximize *expected utility* (EU) are rational decisions. The theory of EU was first presented in axiomatic form by von Neumann and Morgenstern (1947) in an appendix to the second edition of *Theory of Games and Economic Behavior.* It is based on the weak ordering principle (see sect. 2.1 above), extended to gambles or lotteries among outcomes. It is assumed that a player can express a preference or indifference not only between any pair of outcomes, but also between an outcome and a gamble involving a pair of outcomes, or between a pair of gambles, and that the weak ordering principle applies to these preferences also.

This necessitates a further assumption, called the *independence principle* (McClennen 1990, Ch. 3). If g_1, g_2, and g_3 are any three gambles, and $0 < p \leq 1$, then g_1 is preferred to g_2 if and only if a gamble involving g_1 with probability p and g_3 with probability $1 - p$ is preferred to a gamble involving g_2 with probability p and g_3 with probability $1 - p$. From this independence principle, together with the weak ordering principle, it is possible to define a function $u(g)$ that assigns a numerical expected utility to every outcome and gamble, in such a way that the expected utility of a gamble is equal to the sum of the utilities of its components, weighted by their probabilities. It can then be proved that agents who maximize $u(g)$ are acting according to their preferences and are thus manifesting instrumental rationality. Von Neumann-Morgenstern utilities are measured on an interval scale, with an arbitrary zero point and unit of measurement, like temperature measured on a Fahrenheit or Celsius scale, and are therefore unique up to a strictly increasing linear (affine) transformation. This means that two utility scales u and u' represent the same preferences if a and b are arbitrary constants, $a > 0$, and $u' = au + b$. It follows that maximizing u' is equivalent to maximizing u. Harless and Camerer (1994) and Starmer (2000) have comprehensively reviewed EU theory and several alternative "non-expected utility" theories and related empirical findings (see also Camerer 1995; Fishburn 1988; Frisch & Clemen 1994; Hey & Orme 1994; Lea et al. 1987; Machina 1987; 1991; Sosa & Galloway 2000; Taylor 1996).

2.4. Subjective expected utility theory

In Bayesian game theory (initiated by Harsanyi 1967–1968), expected utilities are based on subjective probabilities rather than objective relative frequencies, and what is maximized is *subjective expected utility* (SEU). In SEU theory, utilities obey the axioms formulated by Savage (1954)[5] or one of the alternative axiom systems that have been proposed. Savage built on the axioms of von Neumann and Morgenstern (1947), who introduced the independence principle, and Ramsey (1931), who showed how to define subjective probabilities in terms of preferences among gambles. Rational decisions are those that maximize EU, whether objective or subjective.

2.5. Utility maximization

Utility maximization has a straightforward interpretation in individual decision making. The choice of alternative a_i is rational if the possible alternatives are $a_1, \ldots, a_m$, and there are foreseeable outcomes $c_1, \ldots, c_m$, such that a_1 leads reliably to $c_1, \ldots$, and a_m leads reliably to c_m, and no outcome has a higher utility for the decision maker than c_i. A choice is thus rational if no alternative yields a preferable outcome.

2.6. Expected utility maximization

According to the theory of *revealed preference,* popular with economists, a person who is observed to choose alternative a_i, and to reject a_j, is said to have revealed a preference of a_i over a_j and a higher utility for a_i than a_j – and choice behavior therefore maximizes expected utility tautologically.[6] If chance plays a part, and the choice of a_i leads not to a definite outcome c_i but to a foreseeable probability distribution over the set of outcomes, then a decision maker who chooses an alternative that maximizes the weighted average *expected utility* (EU) is acting rationally. But if the decision is interactive and the outcome is deter-

mined by two or more decision makers, then the interpretation of instrumental rationality is unclear because, except in special cases, an individual cannot maximize EU in any obvious way. In interactive decisions, EU maximization is undefined without further assumptions.

3. Game theory

The necessary assumptions are provided by game theory, the framework within which interactive decisions are modeled. This is a mathematical theory applicable to any social interaction involving two or more decision makers (*players*), each with two or more ways of acting (*strategies*), so that the outcome depends on the strategy choices of all the players, each player having well-defined preferences among the possible outcomes, enabling corresponding von Neumann-Morgenstern utilities (*payoffs*) to be assigned. The definition is inclusive, embracing as it does a wide range of social interactions.

3.1. Abstraction and idealization

A game is a mathematical abstraction functioning as an idealization of a social interaction. An actual interaction is invariably too complex and ephemeral to be comprehended clearly; thus it is replaced by a deliberately simplified abstraction in which the rules and basic elements (players, strategies, payoffs) are explicitly defined, and from which other properties can be inferred by logical reasoning alone. These inferences apply to the idealized game, not directly to the social interaction that it purports to model, and they are valid, provided that the reasoning is sound, whether or not the game models the original interaction accurately. But if it does not, usually because of faulty judgments about which features to ignore, then its relevance and usefulness are limited. To be both relevant and useful, a game must incorporate the important properties of the interaction and must also generate inferences that are not obvious without its help.

3.2. Normative theory

The primary objective of game theory is to determine what strategies rational players should choose in order to maximize their payoffs. The theory is therefore primarily *normative* rather than *positive* or *descriptive.* The founding game theorists stated this explicitly (von Neumann 1928, p. 1; von Neumann & Morgenstern 1944, pp. 31–33). So did Luce and Raiffa (1957), when they introduced game theory to social scientists:

> We feel that it is crucial that the social scientist recognize that game theory is not *descriptive,* but rather (conditionally) *normative.* It states neither how people do behave nor how they should behave in an absolute sense, but how they should behave if they wish to achieve certain ends. (p. 63, emphasis in original)

3.3. Positive theory

If game theory were *exclusively* normative, then it would have limited relevance to the (empirical) behavioral and social sciences, because a normative theory cannot be tested empirically, and evolutionary game theory (see sect. 1.3 above) would be pointless. Arguably, game theory 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. To err is human, and deviations from perfect rationality are inevitable, because of computational limitations or bounded rationality (see sect. 1.2 above), incomplete specification of problems (Berkeley & Humphreys 1982; Dawes 2000), or systematic irrationality (Stanovich & West 2000).[7] But none of this is inconsistent with the hypothesis that people try to act rationally. The addition of this hypothesis provides game theory with a secondary objective, to make testable predictions, and this justifies the thriving enterprise of experimental gaming.

The literature of experimental gaming (reviewed by Colman 1995a; Kagel & Roth 1995, Chs. 1–4; Pruitt & Kimmel 1977) testifies to the fruitfulness of empirical research within a broadly game-theoretic framework. Some important phenomena, such as the clash between individual and collective rationality (see sect. 6 below), cannot even be formulated clearly without the conceptual framework of game theory.

4. Standard assumptions

To give meaning to rational choice in games, it is necessary to introduce assumptions, not only about the players' rationality, but also about their knowledge. The following assumptions are fairly standard[8] and are often called *common knowledge and rationality* (CKR):

CKR1. The specification of the game, including the players' strategy sets and payoff functions, is *common knowledge* in the game, together with everything that can be deduced logically from it and from CKR2.

CKR2. The players are rational in the sense of expected utility (EU) theory (see sects. 2.3 to 2.6 above), hence they always choose strategies that maximize their individual expected utilities, relative to their knowledge and beliefs at the time of acting. (By CKR1 this too is *common knowledge* in the game.)

The concept of *common knowledge* was introduced by Lewis (1969, pp. 52–68) and formalized by Aumann (1976). A proposition is common knowledge among a set 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. Lewis originally wrote "ad infinitum" rather than "and so on" and commented that "this is a chain of implications, not the steps in anyone's actual reasoning" (p. 53). In fact, nothing is gained by carrying the knowledge beyond the *n*th degree when there are *n* players (Binmore 1992, pp. 467–72), and in some games players can reason to solutions with fewer degrees (Bicchieri 1993, Ch. 4). Even three or four degrees may seem impossibly demanding, but according to one interpretation, full common knowledge is an everyday phenomenon arising, for example, whenever a public announcement is made so that everyone knows it, knows that others know it, and so on (Milgrom 1981).[9] Common knowledge is crucially different from every player merely knowing a proposition to be true. The celebrated muddy children problem (Fagin et al. 1995, pp. 3–7) exposes this distinction dramatically (for a formal but simple proof, see Colman 1998, pp. 361–62).

4.1. Implications of the theory

The orthodox belief about the standard assumptions has been summed up by Binmore (1994a):

> Game theorists of the strict school believe that their prescriptions for rational play in games can be deduced, in principle, from one-person rationality considerations without the need to invent collective rationality criteria – provided that sufficient information is assumed to be common knowledge. (p. 142)

That is a fair statement of the belief that this article calls into question. The paragraphs that follow provide diverse reasons for doubting its validity. The implications of strict rationality for games is an important and intrinsically interesting problem. Aumann (2000) has argued that "full rationality is not such a bad assumption; it is a sort of idealization, like the ideas of perfect gas or frictionless motion; . . . no less valid than any other scientific idealization" (p. 139). In a survey of the foundations of decision theory, Bacharach and Hurley (1991) wrote: "Von Neumann and Morgenstern (1944) . . . set out to derive a theory of rational play in games from one of rational individual decision-making. Their successors have not deviated from the faith that this can be done" (pp. 3–4). But there are reasons to suspect that this faith may be misplaced.

5. Focal points and payoff dominance

Let us examine the implications of the CKR assumptions in the most trivial convention game that we may call Heads or Tails. Two people independently choose heads or tails, knowing that if they both choose heads or both tails, then each will receive a payoff of five units of utility, otherwise their payoffs will be zero. This is a *pure coordination game,* because the players' payoffs are identical in every outcome, and the players are motivated solely to coordinate their strategies. Figure 1 shows the payoff matrix.

Player I chooses between the rows, Player II between the columns, and the numbers in each cell represent the payoffs, the first conventionally being Player I's and the second Player II's, though in this game they are always equal. In the games discussed in this article, no harm comes from thinking of the payoffs as US dollars, pounds sterling, euros, or other monetary units. In general, this amounts to assuming that the players are *risk-neutral* within the range of payoffs in the game, so that utility is a strictly increasing linear function of monetary value, though that is immaterial in this trivial game.

		II	
		Heads	***Tails***
I	***Heads***	**5, 5**	**0, 0**
	Tails	**0, 0**	**5, 5**

Figure 1. Heads or Tails

5.1. Nash equilibrium

The players hope to coordinate on either (Heads, Heads) or (Tails, Tails). These are *Nash equilibria* or *equilibrium points.* In a two-person game, an equilibrium point is a pair of strategies that are *best replies* to each other, a best reply being a strategy that maximizes a player's payoff, given the strategy chosen by the other player.

If a game has a uniquely rational solution, then it must be an equilibrium point. Von Neumann and Morgenstern (1944, pp. 146–48) presented a celebrated *Indirect Argument* to prove this important result; Luce and Raiffa (1957, pp. 63–65) gave the most frequently cited version of it; and Bacharach (1987, pp. 39–42) proved it from formal axioms. Informally, the players are rational utility-maximizers (by CKR2). Any rational deduction about the game must (by CKR1) be common knowledge – Bacharach named this the *transparency of reason.* It implies that, if it is uniquely rational for Player I to choose Strategy *X* and Player II Strategy *Y,* then *X* and *Y* must be best replies to each other, because each player anticipates the other's strategy and necessarily chooses a best reply to it. Because *X* and *Y* are best replies to each other, they constitute an equilibrium point by definition. Therefore, if a game has a uniquely rational solution, then it must be an equilibrium point. Whether or not rational players can reason from the standard assumptions to an equilibrium solution is another matter altogether. When the logic of this problem was examined carefully, it became clear that the CKR assumptions are sometimes more than what is required and sometimes insufficient to allow players to reason to an equilibrium solution (Antonelli & Bicchieri 1994; Bicchieri 1993; Bicchieri & Antonelli 1995; Samet 1996).

5.2. Indeterminacy, refinements, and the core

Nash (1950a; 1951) formalized the equilibrium concept and proved that every finite game has at least one equilibrium point, provided that mixed strategies (probability distributions over the pure strategy sets) are taken into consideration. This does not always help a player to choose a strategy, as the game of Heads or Tails shows. In that game, (Heads, Heads) and (Tails, Tails) are equilibrium points, and there is also a mixed-strategy equilibrium in which each player chooses randomly with a probability of 1/2 assigned to each pure strategy (by tossing a coin, for example). But what *should* a rational player do? Any model of evolutionary game theory (see sect. 1.3 above) with a stochastic or noise component would converge on one or other of the pure-strategy equilibrium points, but rational choice remains indeterminate. This exposes a fundamental weakness of classical game theory, namely, its systematic indeterminacy.[10]

Various refinements of Nash equilibrium have been proposed to deal with the indeterminacy problem. The most influential is the *subgame-perfect* equilibrium, proposed by Selten (1965; 1975), but it and other refinements are merely palliative. The Holy Grail is a theory that invariably selects a single equilibrium point, but its status as a *solution* would rest on a dubious assumption of rational determinacy in games (see sect. 6.6 below).

Numerous solution concepts have been suggested for *cooperative games* – games in which players are free to negotiate coalitions based on binding and enforceable agree-

ments governing the division of a payoff. Nash (1950b) pioneered an approach involving the reformulation of cooperative games as non-cooperative ones and the search for equilibrium solutions in the reformulated games, but this *Nash program* ran up against the indeterminacy problem. The most fundamental and influential solution concept for cooperative games is the *core* (Gillies 1953). An outcome x of a cooperative game is said to dominate another outcome y if there is a potential coalition that has both the motive and the power to enforce x. The core of a cooperative game is the set of undominated outcomes. The core satisfies individual, coalition, and collective rationality, inasmuch as it includes only divisions of the payoff such that the players receive at least as much as they could guarantee for themselves by acting independently, every proper subset of the players receives at least as much as it could guarantee for itself by acting together, and the totality of players receives at least as much as it could guarantee for itself by acting collectively as a grand coalition, so that nothing is wasted. But there are many games in which no division satisfies all these requirements and the core is therefore empty. For example, if three people try to divide a sum of money among themselves by majority vote, then any proposed division can be outvoted by a two-person coalition with the will and the power to enforce a solution that is better for both of its members.[11] Rational social interaction, at least as defined by the core, is simply infeasible in these circumstances. Other solution concepts for cooperative games suffer from similar pathologies.

In the non-cooperative game of Heads or Tails, rational players are forced to choose arbitrarily, with a probability of successful coordination of 1/2 and an expected payoff of 2.5. Can they do better than that?

5.3. Focal points

Of course they can. Going beyond the mathematical properties of the game and delving into its psychology, if both players perceive heads to be more salient than tails, in other words if they both recognize (Heads, Heads) as a *focal point,* and if both believe this to be common knowledge, then both will unhesitatingly choose heads, and they will coordinate successfully. This was first pointed out by Schelling (1960, Ch. 3), who reported the results of informal experiments in which 86 percent of participants chose heads. This implies a probability of coordination of .86 × .86 or approximately 3/4, and hence an expected payoff of approximately 3.7 – a big improvement.

According to Schelling (1960), what enables players to focus on heads is the "conventional priority, similar to the convention that dictates A, B, C, though not nearly so strong" (p. 64) of heads over tails. Mehta et al. (1994a; 1994b) replicated his finding in England, where 87 percent of players chose heads. Both studies also included several more difficult pure coordination games, some with infinite strategy sets, in which players frequently coordinated on focal points without difficulty. For example, suppose that you have to meet a stranger at a specified place on a specified day but neither of you has been told the time of the meeting. What time would you choose to optimize your chances of coordinating? Most people focus unhesitatingly on 12 noon (Schelling 1960, p. 55).

5.4. Hume's example

The idea of a focal point can be traced back to a discussion by Hume (1739–40/1978, 3.II.iii) of a pure coordination game played by a German, a Frenchman, and a Spaniard, who come across three bottles of wine, namely Rhenish, Burgundy, and port, and "fall a quarrelling about the division of them" (pp. 509–10*n*). There are 27 ways of assigning three bottles to three people, or six permutations if each person gets exactly one bottle. Hume pointed out that the obvious focal point among these alternatives is to "give every one the product of his own country" (p. 510*n*).[12]

The focal point of Heads or Tails emerges from its representation within the *common language* shared by the players (Crawford & Haller 1990). Considered in the abstract, this game, or the problem of the unspecified meeting time, or Hume's problem of the three wines, has no focal point. To remove the common language, including the culturally determined labeling of strategies, is to filter out the focal points, reducing the prospects of coordination to chance levels.

5.5. Gilbert's argument

The salient focal points are obvious in Heads or Tails, the unspecified meeting time, and Hume's problem of the three wines. Nevertheless, it turns out that their selection cannot be justified rationally. Gilbert (1989b) showed that "if human beings are – happily – guided by salience, it appears that this is not a consequence of their rationality" (p. 61) and that "mere salience is *not* enough to provide rational agents with a reason for action (though it would obviously be nice, from the point of view of rational agency, if it did)" (p. 69, emphasis in original).

Gilbert's proof is easy to follow, though hard to swallow. The focal point of Heads or Tails is obviously (Heads, Heads), and to clarify the argument, let us assume that the players have previously agreed on this, so it is common knowledge. Under the CKR2 rationality assumption, Player I will choose heads, given any reason for believing that Player II will choose heads, to ensure a payoff of 5 rather than 0. But in the absence of any reason to expect Player II to choose heads, Player I has no reason to choose it or not to choose it. The fact that (Heads, Heads) is a focal point is not a valid reason for Player I to choose heads, because heads is best only if Player II chooses it also. Because the salience of (Heads, Heads) does not give Player I a reason to choose heads, it cannot give Player I a reason to expect Player II to choose heads. Both players are in exactly the same quandary, lacking any reason for choosing heads in the absence of a reason to expect the co-player to choose it. The argument goes round in circles without providing the players with any rational justification for playing their parts in the focal-point equilibrium, in spite of its salience and intuitive appeal.

This is an excellent example of the fundamental thesis of this article, that the concept of utility maximization cannot be applied straightforwardly to interactive decisions.

5.6. Payoff dominance

Gilbert's (1989b) argument applies even to games with structurally inbuilt *payoff-dominant* (or *Pareto-dominant*)

		II: *H*	II: *L*
I	*H*	6, 6	0, 0
	L	0, 0	3, 3

Figure 2. Hi-Lo Matching game

		II: *H*	II: *L*
I	*H*	6, 6	0, 7
	L	0, 0	3, 3

Figure 3. Modified Hi-Lo Matching game

equilibrium points. These games have focal points that do not depend on any common language, *pace* Crawford and Haller (1990). Payoff dominance is illustrated most simply in the Hi-Lo Matching game (Fig. 2). If both players choose *H,* each gains six units; if both choose *L,* each gains three units; otherwise neither gains anything. The two obvious equilibria are *HH,* with payoffs of (6, 6) and *LL,* with payoffs of (3, 3). (There is also a mixed-strategy equilibrium in which each player chooses 1/3 *H* and 2/3 *L,* with expected payoffs of 2 units each.) Rational players prefer *HH* to *LL,* because *HH* payoff-dominates *LL.* An equilibrium point payoff-dominates another if it yields a higher payoff to both players. It is obviously a structural focal point.

The *payoff-dominance principle* is the assumption that, if one equilibrium point payoff-dominates all others in a game, then rational players will play their parts in it.[13] Harsanyi and Selten's (1988) general theory of equilibrium selection is based on it, together with a secondary *risk-dominance principle,*[14] and most game theorists accept its intuitive force (e.g., Bacharach 1993; Crawford & Haller 1990; Farrell 1987; 1988; Gauthier 1975; Janssen 2001b; Lewis 1969; Sugden 1995; 2000). Empirical tests of the payoff-dominance principle have yielded mixed results (e.g., Cooper et al. 1990; van Huyck et al. 1990). But, astonishingly, a straightforward extension of Gilbert's (1989b) argument reveals that a player has a reason to choose *H* if and only if there is a reason to expect the co-player to choose *H,* and there is no such reason, because both players face the identical quandary. The fact that *HH* is the optimum equilibrium (indeed, the optimum outcome) for both players is not *ipso facto* a reason for Player I to expect Player II to choose *H,* because *H* is not a utility-maximizing choice for Player II in the absence of any reason to expect Player I to choose it, and vice versa (Casajus 2001; Colman 1997; Gilbert 1990; Hollis & Sugden 1993; Sugden 1991b). This is a startling failure of game theory. When one first appreciates the force of the argument, one feels like pinching oneself.

A common initial reaction is to try to justify the choice of *H* in Figure 2 on the grounds that "The best I can get from choosing *H* is better than the best I can get from choosing *L,* and the worst is no worse, therefore I should choose *H.*" To see the fallacy in this naive *maximax* reasoning, consider the slightly modified game shown in Figure 3. In this version, Strategy *L* gives Player II a higher payoff *whatever* Player I chooses, therefore a rational Player II will certainly choose it. By the transparency of reason, Player I will anticipate this and will therefore also choose *L,* hence the rational solution is unambiguously *LL.* But the maximax argument ("The best I can get from choosing *H* is better than the best I can get from choosing *L,* and the worst is no worse, therefore I should choose *H*") would still lead Player I to choose *H,* and that is manifestly absurd.

A more sophisticated fallacy is the attempt to justify choosing *H* in the Hi-Lo Matching game (Fig. 2) by assigning subjective probabilities to the co-player's strategies. The specific probabilities are immaterial, so let us suppose Player I assumes (perhaps by the Principle of Insufficient Reason) that Player II's strategies are equally probable. If this assumption were valid, then Player I would indeed have a reason (SEU maximization) to choose *H,* but a simple *reductio* proof exposes the error. By the transparency of reason, Player I's intention to choose *H* would be common knowledge and would induce Player II to choose the best reply, namely *H,* with *certainty,* contradicting Player I's initial assumption.

5.7. Coordination without rationality

Under the CKR knowledge and rationality assumptions, coordination by focal point selection ought to be impossible, yet it occurs quite frequently in everyday social interaction. Even in games with blindingly obvious payoff-dominant focal points, players have no rational justification for choosing the corresponding strategies. Orthodox game-theoretic rationality is powerless to explain these phenomena.

6. Social dilemmas

Social dilemmas are games in which individual and collective interests conflict. The simplest is the familiar two-person Prisoner's Dilemma game (PDG). The general *N*-player Prisoner's Dilemma (NPD), of which the PDG is a special case, was discovered simultaneously and independently by Dawes (1973), Hamburger (1973), and Schelling (1973). Social dilemmas have generated a vast amount of theoretical and empirical research (reviewed by Colman 1995a, Chs. 6, 7, 9; Dawes 1980; 1988, Ch. 9; Foddy et al. 1999; Ledyard 1995; Nozick 1993, pp. 50–59; Rapoport 1989, Chs. 12, 14; Schroeder 1995; van Lange et al. 1992; and van Vugt 1998; among others).

6.1. Self-defeating strategies

The peculiarity of the Prisoner's Dilemma game is that, of the two strategies available to each player, one is uniquely rational, yet each player fares better if both choose the other. To borrow a felicitous epithet from Parfit (1979; 1984, Ch. 1), rationality is *self-defeating* in the PDG, and in social dilemmas in general.

		II: *C*	II: *D*
I	*C*	**3, 3**	**1, 4**
	D	**4, 1**	**2, 2**

Figure 4. Prisoner's Dilemma game

6.2. Prisoner's Dilemma formalization

The PDG (Fig. 4) was discovered in 1950 by Flood and Dresher (Poundstone 1993, p. 8; Raifa 1992, p. 173). What defines it as a PDG are the relative rather than the absolute values of the payoffs, hence the numbers 4, 3, 2, and 1 are used for simplicity, though they are assumed to be utilities.

6.3. Lifelike interpretation

The name *Prisoner's Dilemma* comes from an interpretation involving two prisoners, introduced by Tucker in a seminar in the Psychology Department of Stanford University in 1950, the most familiar published version being Luce and Raiffa's (1957, pp. 94–97). The story is too well known to repeat, and the following alternative interpretation, based on an idea of Hofstadter's (1983), will help to fix the idea of a game as an abstract structure applicable to a potentially unlimited set of interactions.

Player I is keen to buy a packet of illegal drugs from Player II, and Player II is keen to sell it. They have agreed a price that suits them both, but because of the nature of the trade, it must take place without face-to-face contact. Player I promises to leave an envelope full of money in a dead-letter drop, such as a litter bin in a park. Player II promises to leave an envelope full of drugs in another, distant, dead-letter drop at the same time. Each player faces a choice between cooperating (leaving the promised envelope) or defecting (neglecting to leave it). If both players cooperate and choose *C*, then the payoffs are good for both (3, 3). If both defect and choose *D*, the payoffs are worse for each (2, 2). And if one player cooperates while the other defects, then the outcome is worst for the cooperator and best for the defector, hence the payoffs are (1, 4) or (4, 1), depending on who cooperates and who defects.

6.4. Ubiquity of social dilemmas

Many everyday two-person interactions have the strategic structure of the PDG. Rapoport (1962) discovered it in Puccini's opera *Tosca.* Lumsden (1973) showed empirically that the Cyprus conflict shared the preference structure of an indefinitely repeated PDG. The PDG is a standard model of bilateral arms races and duopoly competition. Many other two-person interactions involving cooperation and competition, trust and suspicion, threats, promises, and commitments are PDGs.

6.5. Strategic dominance

How should a rational player act in the PDG? There are two main arguments in favor of defecting (choosing *D*). They apply to the standard one-shot PDG. For indefinitely iterated PDGs, a folk theorem establishes a vast number of equilibrium points, including many leading to joint cooperation (see Binmore 1992, pp. 373–76, for a clear proof), and evolutionary experiments have reported high levels of cooperation (Axelrod 1984; 1997, Chs. 1, 2; Kraines & Kraines 1995; Nowak et al. 1995; Nowak & Sigmund 1993). The *finitely* iterated PDG presents a different problem altogether, to be discussed in section 7.1 below.

The most powerful reason for defecting in the one-shot PDG is *strategic dominance.* The *D* strategies are strongly dominant for both players inasmuch as each player receives a higher payoff by choosing *D* than *C* against *either* counterstrategy of the co-player. Player I receives a higher payoff by choosing *D* than *C* whether Player II chooses *C* or *D*, hence *D* is a strongly dominant strategy for Player I and, by symmetry, the same applies to Player II. It is in the interest of each player to defect *whatever the other player might do.*

It is generally agreed that a rational agent will never choose a dominated strategy. Dixit and Nalebuff (1991, p. 86) identified the avoidance of dominated strategies as one of the four basic rules of successful strategic thinking, and it has been proved that the only strategies in two-person games that can be *rationalized* – justified in terms of consistent beliefs about the co-player's beliefs – are those that survive a process of successively deleting strongly dominated strategies (Bernheim 1984; Pearce 1984).

Strategic dominance is a simplified version of the *sure-thing principle,*[15] first propounded by Savage (1951) and incorporated into his decision theory as an axiom, with the comment: "I know of no other extralogical principle governing decisions that finds such ready acceptance" (Savage 1954, p. 21). The strategic dominance principle, in its strong form, can be deduced from elementary axioms of game theory (Bacharach 1987), although the (weak) sure-thing principle cannot (McClennen 1983).

In individual decision making, the sure-thing principle seems intuitively compelling, except in certain pathological though interesting cases based on Simpson's paradox (Shafir 1993) or Newcomb's problem (Campbell & Sowden 1985), or in situations in which players' actions are not independent (Jeffrey 1983, pp. 8–10). But when the strategic dominance principle is applied to the PDG, the conclusion seems paradoxical, because if both players choose dominated *C* strategies, then each receives a higher payoff than if both choose dominant *D* strategies. The *DD* outcome resulting from the choice of dominant strategies is *Pareto-inefficient* in the sense that another outcome (*CC*) would be preferred by *both* players.

6.6. Argument from Nash equilibrium

The second major argument for defection focuses on the fact that *DD* is the PDG's only equilibrium point. It is obvious in Figure 4 that *D* is a best reply to *D* and that there is no other equilibrium point. From the Indirect Argument (see sect. 5.1 above), if the PDG has a uniquely rational solution, then, because it must be an equilibrium point, it must therefore be *this* equilibrium point.

It is often considered axiomatic that every game has a

uniquely rational solution that could, in principle, be deduced from basic assumptions (Harsanyi 1962; 1966; Harsanyi & Selten 1988; Weirich 1998). When it is expressed formally, this existence postulate is called the *principle of rational determinacy.* Nash (1950a; 1950b; 1951) assumed it tacitly at first, then in a later article (Nash 1953) introduced it explicitly as the first of seven axioms: "For each game . . . there is a unique solution" (p. 136). Although it is widely accepted, Bacharach (1987) pointed out that it remains unproven, and this blocks the inference that a game's unique equilibrium point must necessarily be a uniquely rational solution, because the game may have no uniquely rational solution. Sugden (1991a) presented several reasons for skepticism about the principle of rational determinacy. However, in the PDG, we know from the dominance argument that joint defection must be uniquely rational, and it is therefore paradoxical that irrational players who cooperate end up better off.

6.7. Experimental evidence

Joint defection is uniquely rational in the PDG. Binmore (1994a) devoted a long chapter (Ch. 3) to refuting fallacies purporting to justify cooperation. But as a prediction about human behavior, this fails miserably in the light of experimental evidence (reviewed by Colman 1995a, Ch. 7; Good 1991; Grzelak 1988; Rapoport 1989, Ch. 12; among others). In the largest published PDG experiment (Rapoport & Chammah 1965), almost 50 percent of strategy choices were cooperative, and even in experiments using one-shot PDGs, many players cooperate, to their mutual advantage. Game theory fails as a positive theory in the PDG, because human decision makers do not follow its rational prescriptions.

6.8. Three-player Prisoner's Dilemma

The simplest multi-player Prisoner's Dilemma, using payoffs of 4, 3, 2, and 1 for convenience once again, is the three-player NPD shown in Table 1. The first row of Table 1 shows that the payoff to each *C*-chooser is 3 if all three players choose *C,* and in that case the payoff to each (non-existent) *D*-chooser is undefined, hence the dash in the last column. In the second row, if two players choose *C* and the remaining player chooses *D,* then the payoff is 2 to each *C*-chooser and 4 to the *D*-chooser, and so on.

6.9. Defining properties

The defining properties of the NPD are as follows:

1. Each player chooses between two options that may be labeled *C* (cooperate) and *D* (defect).

Table 1. *Three-player Prisoner's Dilemma*

Number choosing *C*	Number choosing *D*	Payoff to each *C*-chooser	Payoff to each *D*-chooser
3	0	3	—
2	1	2	4
1	2	1	3
0	3	—	2

2. The *D* option is strongly dominant for each player: each obtains a higher payoff by choosing *D* than *C* no matter how many of the others choose *C.*

3. The dominant *D* strategies intersect in an equilibrium point that is Pareto-inefficient: the dominant *D* strategies are best replies to one another, but the outcome is better for every player if all choose their dominated *C* strategies.

The NPD in Table 1 has all three properties, and so does the two-person PDG in Figure 4, which can now be seen as a special case. The NPD is ubiquitous, especially in situations involving conservation of scarce resources and contributions to collective goods. Examples of defection include negotiating a wage settlement above the inflation rate, neglecting to conserve water during a drought or fuel during a fuel shortage, over-fishing, increasing armaments in a multilateral arms race, and bolting for an exit during an escape panic. In each case, individual rationality mandates defection regardless of the choices of the others, but each individual is better off if everyone cooperates.

6.10. More experimental evidence

Normative arguments for defection in the PDG – strategic dominance and unique Nash equilibrium – apply equally to the NPD. But experimental investigations since the mid-1970s, using both NPDs and strategically equivalent *commons dilemmas* (also called *resource dilemmas*) and *public goods dilemmas* (also called *free-rider problems*), have invariably found rampant cooperation. The proportion of cooperative choices tends to decrease as group size increases but remains substantial even in large groups. The experimental evidence has been reviewed by Colman (1995a, Ch. 9), Dawes (1980; 1988, Ch. 9), Foddy et al. (1999), Ledyard (1995), Rapoport (1989, Ch. 14), Schroeder (1995), and van Lange et al. (1992), among others. Ledyard's main conclusion was: "*Hard-nosed game theory cannot explain the data. . . .* If these experiments are viewed solely as tests of game theory, that theory has failed" (p. 172, emphasis in original).

6.11. Puzzling conclusions

Game-theoretic rationality requires rational players to defect in one-shot social dilemmas. This conclusion is puzzling in the light of experimental evidence showing widespread cooperation in two-player and multi-player social dilemmas. The evidence has failed to corroborate positive game theory. Even from a normative point of view, rationality is self-defeating in social dilemmas. Players who cooperate by choosing dominated *C* strategies end up with higher individual payoffs than players who follow game-theoretic rationality. Because utility maximization is the essence of game-theoretic rationality, this seems counter-intuitive and even paradoxical.

Social dilemmas are not the only games that generate robust experimental data starkly at odds with game theory. The more recently discovered Ultimatum game is beginning to upstage the PDG in the freak show of human irrationality. Orthodox rationality is manifestly ill-advised in the Ultimatum game. Space constraints forbid a detailed discussion of it here (see Thaler 1992, Ch. 3, for an excellent review).

7. Backward induction

Backward induction is a form of reasoning resembling mathematical induction, an established method of proving theorems. It was introduced into game theory by Zermelo (1912), who proved the first major theorem of game theory to the effect that every strictly competitive game like chess, in which the players have perfect information of earlier moves, has either a guaranteed winning strategy for one of the players or guaranteed drawing strategies for both.

7.1. Finitely iterated Prisoner's Dilemma

In a famous passage, Luce and Raiffa (1957, pp. 97–102) applied backward induction to the finitely iterated PDG – that is, the PDG repeated a finite number of times (n) by a pair of players. They proved that rational players who consider each other rational will defect on every move, provided that both know n in advance. The proof is easy. Suppose that the game is to be iterated twice ($n = 2$). Rational players know that defection is the rational strategy in a one-shot PDG (see sects. 6.5 and 6.6 above). There could be a reason for cooperating on the first round if it might influence the outcome of the second. But in the second round, there are no moves to follow and thus no future effects to consider; consequently, the second round is effectively a one-shot PDG, and its rational outcome is joint defection. Because the outcome of the second round is thus predetermined, and the first round cannot influence it, the first round is also effectively a one-shot PDG. Therefore, both players will defect on the first round as well as the second. This argument generalizes straightforwardly to any finite n, showing that rational players will defect on every round.

This conclusion is counterintuitive, because both players would receive higher payoffs if both behaved more cooperatively. Luce and Raiffa (1957) had difficulty accepting their own proof: "If we were to play this game we would *not* take the [*D*] strategy at every move!" (p. 100, emphasis in original). They offered no persuasive reason for disowning their proof, although subsequent experimental evidence corroborated their intuition by showing that intelligent players tend to cooperate (e.g., Andreoni & Miller 1993; Cooper et al. 1996; Selten & Stoecker 1986). What is typically observed in finitely iterated experimental PDGs with human players is a substantial proportion of cooperative choices, often exceeding 30 percent, and a sizeable minority even in the last round. Furthermore, high levels of cooperation are observed in evolutionary Prisoner's Dilemma games (e.g., Axelrod 1984; 1997, Chs. 1, 2; Nowak et al. 1995).

7.2. Chain-store game

Backward induction rose to prominence with Selten's (1978) introduction of the multi-player Chain-store game (subsequently discussed by Bicchieri 1989; 1993, pp. 192–94; Friedman 1991, pp. 190–93; Kreps & Wilson 1982a; Milgrom & Roberts 1982; Ordeshook 1986, pp. 451–62; Rosenthal 1981; and others). In outline, without quantifying payoffs, the game is as follows. A chain-store has branches in 20 cities, in each of which there is a local competitor hoping to sell the same goods. These potential challengers decide one by one whether to enter the market in their home cities. Whenever one of them enters the market, the chain-store responds either with aggressive predatory pricing, causing *both* stores to lose money, or cooperatively, sharing the profits 50–50 with the challenger.

Intuitively, the chain-store seems to have a reason to respond aggressively to early challengers in order to deter later ones. But Selten's (1978) backward induction argument shows that deterrence is futile. The argument begins at the final stage, when the twentieth challenger decides whether to enter the market. This challenger will enter, and the chain-store will respond cooperatively, because there is no future challenger to deter and therefore no rational reason for the chain-store to sacrifice profit by responding aggressively. Because the outcome of the final round is thus predetermined, the nineteenth challenger will enter the market on the penultimate round, and the chain-store will respond cooperatively, because both players know that the final challenger cannot be deterred and that the chain-store therefore has no reason to act aggressively on the current round. This argument unfolds backwards to the first move. Every challenger will enter the market when its turn comes, and the chain-store will always respond cooperatively.

Like Luce and Raiffa (1957) before him, Selten (1978) found his backward induction proof unpalatable. He admitted that, if he were to play the game in the role of the chain-store, he would play aggressively to deter subsequent competitors:

> I would be very surprised if it failed to work. From my discussions with friends and colleagues, I get the impression that most people share this inclination. In fact, up to now I met nobody who said that he would behave according to [backward] induction theory. My experience suggests that mathematically trained persons recognize the logical validity of the induction argument, but they refuse to accept it as a guide to practical behavior. (Selten 1978, pp. 132–33)

If deterrence is indeed futile in finitely iterated interactions, then the implications are momentous. Economic competition often involves finite sequences of challenges to the dominance of a market leader, and deterrence lies at the heart of military defense strategy and penology. If deterrence is potentially futile, then much commercial, military, and judicial strategy is irrational. On the other hand, the futility of deterrence in finite sequences of interactions could help to explain why animals often settle disputes by conventional displays rather than escalated fighting (Lazarus 1995; Maynard Smith 1976).

7.3. Centipede game

The paradoxical implications of backward induction emerge most vividly in the two-person Centipede game, introduced (among *obiter dicta* on the Chain-store game) by Rosenthal (1981) and subsequently named by Binmore (1987) after the appearance of its extensive-form graph. It has attracted much discussion (e.g., Aumann 1995; 1996; 1998; Basu 1990; Bicchieri 1993, pp. 91–98; Binmore 1994b; Bonanno 1991; Colman 1998; Hollis & Sugden 1993; Kreps 1990; Pettit & Sugden 1989; Sugden 1991b; 1992). A simple version – actually a tetrapod, but long enough to illustrate the basic idea – is shown in Figure 5.

The graph shows a sequence of moves, starting at the left. Players I and II alternate in choosing, at each decision node, whether to defect by moving down or to cooperate by moving across. If a player defects down, then the game stops at that point and the payoffs are shown in parentheses (in the

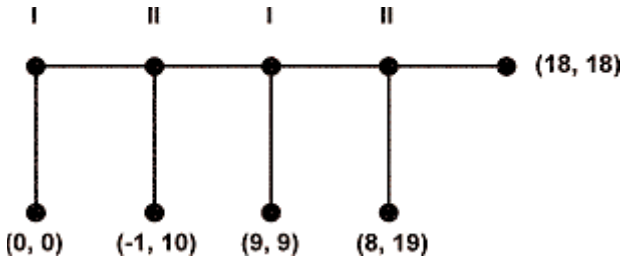

Figure 5. Centipede game

order I, II). Thus, if Player I defects down on the first move, then the game stops and both payoffs are zero. Whenever a player continues across, that player loses 1 unit and the other gains 10. Thus, if Player I continues across at the first decision node, losing 1 unit and adding 10 to Player II's payoff, and if Player II promptly defects down, then the game stops and the payoffs are −1 to Player I and 10 to Player II, and so on. If both players continue across at every decision node, then the game ends automatically after the fourth decision, with a payoff of 18 to each player.

7.4. Argument for defection

Players can earn large payoffs by cooperating, but backward induction shows, astonishingly, that Player I should defect down and stop the game on the first move, with zero payoffs to both players. Suppose the game has reached the final decision node. Player II can defect down and receive 19 or continue across and receive 18. A rational Player II will obviously defect down. Consequently, at the third decision node, a rational Player I who can anticipate Player II's reasoning will defect down to receive 9, rather than continue across and receive only 8 when Player II defects down on the following move. Extending this line of reasoning backwards, Player I will defect down at the first decision node, even in a Centipede with 100 feet and fabulous payoffs dangling from its head and shoulders, in spite of the fact that both players could enjoy these riches by playing more cooperatively.

This is an almost intolerably counterintuitive conclusion. In experimental Centipede games, most players behave far more cooperatively, the great majority continuing across at the first decision node, and a substantial minority even at the last (El-Gamal et al. 1993; Fey et al. 1996; Güth et al. 1997; McKelvey & Palfrey 1992). This cannot be explained by orthodox game-theoretic rationality with backward induction. Cooperative players, who earn significant monetary payoffs in experimental Centipede games, could legitimately ask rational utility maximizers, who come away with nothing: "If you're so rational, how come you ain't rich?" This seems a good question, given that rational decision making is, by definition, utility maximization.

7.5. Are the assumptions incoherent?

Player I, if rational, will defect down on the first move. According to the CKR2 rationality assumption, this means that, for Player I, the expected utility of defecting down must exceed the expected utility of continuing across. It follows that we must have proved that Player I expects Player II to respond to a cooperative opening move by defecting down immediately. If Player I assigned even a small probability – anything greater than 1/10 – to Player II's responding cooperatively, then to maximize expected utility, Player I would open with a cooperative move.

But is Player I's expectation rational? Perhaps not, for the following reason. If Player I were to open with a cooperative move, and if the backward induction argument is sound, then the CKR2 rationality assumption would have been violated. Player II would face something in conflict with that basic assumption, namely, an irrational co-player. It would then be impossible for Player II to respond rationally, because the theory would no longer apply. But if Player II's response is therefore unpredictable, how could Player I evaluate the expected utility of cooperating on the opening move? For Player I, the expected utility of this move seems indeterminate.

This leads to an impasse in which neither player can act rationally. Orthodox game-theoretic rationality seems to have broken down, and this raises a suspicion that the CKR assumptions may be incoherent. Cubitt and Sugden (1994; 1995) reached a similar conclusion from formal analyses in which they defined a concept of rationally justified play and examined its implications under the CKR assumptions, formulated as axioms, adding a further axiom according to which players never assign a zero probability to any of their co-players' rationally justifiable strategies being played. It should then have been possible to prove any strategy either rationally justifiable or unjustifiable, but it turned out that there are strategies that fall into neither category, even in a simple nonsequential game that they exhibited as a counterexample. Working along similar lines, Samuelson (1992) and Squires (1998) have cast further doubt on the coherence of the fundamental assumptions of game theory.

8. Psychological game theory

The arguments and evidence discussed above seem to imply that orthodox game theory cannot explain strategic interaction in widely disparate games. To deal with similar problems, Camerer (1997) proposed the label *behavioral game theory* for an approach that replaces descriptively inaccurate aspects of game theory with plausible explanations providing better predictions of empirical (especially experimental) observations. One of Camerer's examples of behavioral game theory is Rabin's (1993) *fairness equilibrium,* based on payoff transformations. According to this approach, a player's payoff increases by a fixed proportion α of a co-player's payoff if the co-player acts kindly or helpfully and decreases by α of the co-player's payoff if the co-player acts meanly or unhelpfully. In the Prisoner's Dilemma game (Fig. 4), for example, if $\alpha = 1/2$, so that each player's payoff increases by a half that of the co-player when the co-player cooperates, and decreases by half that of the co-player when the co-player defects, then joint cooperation (CC) emerges as a new equilibrium point, called a fairness equilibrium, with transformed payoffs of 4½ to each player. According to Camerer, this may help to explain cooperation in the Prisoner's Dilemma game.

Psychological game theory, as described in the following subsections, overlaps behavioral game theory but focuses specifically on nonstandard reasoning processes rather than other revisions of orthodox game theory such as payoff transformations. Psychological game theory seeks to mod-

ify the orthodox theory by introducing formal principles of reasoning that may help to explain empirical observations and widely shared intuitions that are left unexplained by the orthodox theory. An important forerunner is the work of Geanakoplos et al. (1989) on psychological games in which payoffs depend not only on players' actions, but also on their expectations. The forms of psychological game theory that will be discussed are team reasoning, Stackelberg reasoning, and epistemic and non-monotonic reasoning. Like behavioral game theory, psychological game theory is primarily descriptive or positive rather than normative, and it amounts to nothing more than a collection of tentative and ad hoc suggestions for solving the heterogeneous problems that have been highlighted in earlier sections.

8.1. Team reasoning

First, how do human players choose focal points in pure coordination games? In particular, what accounts for the intuitive appeal of payoff-dominant equilibria, such as *HH* in the Hi-Lo Matching game (Fig. 2), and how do human decision makers manage to coordinate their strategies on payoff-dominant equilibria? Two suggestions have been put forward to answer these questions, both involving departures from the CKR assumptions. The first is *team reasoning,* formulated by Gilbert for a variety of problems (1987; 1989a; 1989b; 1990; 2000); developed into a decision theory by Sugden (1993; 1995; 2000); and formalized in a stochastic form by Bacharach (1999). A team-reasoning player maximizes the objective function of the *set* of players by identifying a profile of strategies that maximizes their *joint* or *collective* payoff, and then, if the maximizing profile is unique, playing the individual strategy that forms a component of it.

Gilbert (1989a) showed how "a conclusion about what an individual should do can follow directly, without the interposition of any assumptions about what *that individual* wants or seeks. Indeed, no single *individual's* aims need be referred to" (p. 708, emphasis in original). This implies the existence of *collective preferences,* in addition to standard individual preferences. Sugden (2000) illustrated this concept with his family's preferences for spending a summer holiday: "When it comes to walks, we prefer walks of six miles or so. . . . But 'our' preferences are not exactly those of any one of us. My ideal walk would be somewhat longer than six miles" (p. 175). Gilbert (2000) examined the implications of collective preferences for rational choice theory.

Team reasoning is simple and intuitively compelling but profoundly subversive of orthodox decision theory and game theory, both of which rest on a bedrock of *methodological individualism.* It departs from the CKR2 rationality assumption, inasmuch as a team-reasoning player does not maximize individual EU but pursues collective interests instead. It is nevertheless easy to call to mind anecdotal evidence of joint enterprises in which people appear to be motivated by collective rather than individual interests. In team sports, military situations, joint business ventures, and even family outings, people sometimes choose to do what is best for "us," even when this collective interest does not coincide with their individual preferences. There is some experimental evidence that collective preferences can indeed arise in quite ordinary circumstances (Park & Colman 2001). Research based in social dilemmas has shown that mutually beneficial cooperation and "we-thinking" are enhanced by merely raising players' sense of group identity (Brewer & Kramer 1986; Dawes et al. 1988; 1990).

Team-reasoning players choose *HH* in the Hi-Lo Matching game (Fig. 2) and the modified version (Fig. 3), because the team's collective payoff, defined by the sum of the players' individual payoffs, is highest there. Thus, team reasoning offers a plausible explanation for the intuitive appeal of the payoff-dominance principle. It may even help to explain cooperation in social dilemmas, because a glance at Figure 4 and Table 1 confirms that, in those games, the collective payoff of the pair or set of players is maximized by joint cooperation, though it is arguably a weakness of team reasoning that it sometimes predicts such out-of-equilibrium outcomes.

The focus on cooperative rather than individual goals tends to foster a debilitating misconception among some social psychologists that team reasoning is merely a *social value orientation,* alongside individualism, altruism, competitiveness, and equality seeking, in the payoff-transformational theories of McClintock and Liebrand (1988), van Lange (1999), and van Lange and De Dreu (2001). In fact, it is impossible to accommodate team reasoning in these approaches, because they all rely implicitly on methodological individualism. Van Lange's (1999) integrative theory includes a model that resembles team reasoning superficially and may therefore be expected to predict coordination in the Hi-Lo Matching game. Van Lange defined *cooperation* as an individual motivation to maximize the outcome transformation function $OT = W_1$ (own payoff) + W_2 (co-player's payoff). The payoff matrix is transformed according to this function, and each player then proceeds with standard individualistic reasoning, using the transformed *OT* payoffs. This leads to cooperation in the Prisoner's Dilemma game, as van Lange pointed out, but it fails in the Hi-Lo Matching game and is not equivalent to team reasoning.

Application of van Lange's (1999) transformation to the Hi-Lo Matching game (Fig. 2) involves replacing each payoff with the sum of individual payoffs in the corresponding cell, and it produces the Bloated Hi-Lo Matching game shown in Figure 6. But this is strategically identical to the original game, and the coordination problem is back with a vengeance. The idea of forming linear combinations of own and other's payoffs was proposed, in a more general form than van Lange's, over a century ago by Edgeworth (1881/1967, pp. 101–102). Although it seems plausible and may help to explain some social phenomena (Camerer 1997, pp. 169–70), it cannot explain the payoff-dominance phe-

		II	
		H	***L***
I	***H***	**12, 12**	**0, 0**
	L	**0, 0**	**6, 6**

Figure 6. Bloated Hi-Lo Matching game

nomenon or incorporate team reasoning. Team reasoning is inherently non-individualistic and cannot be derived from transformational models of social value orientation.

8.2. Stackelberg reasoning

A second suggestion for solving the payoff-dominance puzzle is *Stackelberg reasoning,* proposed by Colman and Bacharach (1997). Stackelberg reasoning is a generalization of the "minorant" and "majorant" models used by von Neumann and Morgenstern (1944, pp. 100–101) to rationalize their solution of strictly competitive games. The basic idea 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 the players could anticipate each other's choices. This enables players to select payoff-dominant equilibria. For example, in the Hi-Lo Matching game (Fig. 2), a Stackelberg-reasoning player may deliberate as follows:

> Suppose my co-player could read my mind, or at least anticipate my strategy choice. If I chose *H,* my co-player would best-reply *H,* and I'd receive my optimal payoff of 6. If I chose *L,* then my co-player would best-reply *L,* and I'd receive 3. In this version of the game, I'd maximize my payoff by choosing *H.* My co-player knows all this, because we share a common faculty of reason, so in the actual game, I should choose *H* and expect my co-player to do likewise.

Formally, in any two-person game, Player *i* assumes that Player *j* can anticipate *i*'s choice. This implies that for every strategy s_i in *i*'s strategy set, Player *j* responds with a best reply $f(s_i)$. Player *i* expects this to happen and, if $f(s_i)$ is unique, chooses a payoff-maximizing best reply to it. Thus, Player *i,* seeking to maximize the payoff function H_i, chooses a strategy for which $\max_i H_i [s_i, f(s_i)]$ is attained. This is called Player *i*'s Stackelberg strategy.

In any game *G,* if the players' Stackelberg strategies are in Nash equilibrium, then *G* is *Stackelberg soluble,* and the corresponding equilibrium point is called a *Stackelberg solution.* Stackelberg reasoning acts as a strategy generator, and Nash equilibrium as a strategy filter.[16] Stackelberg reasoning mandates the choice of Stackelberg strategies only in games that are Stackelberg soluble. Colman and Bacharach (1997) proved that all common-interest games are soluble in this sense, and that in every game with Pareto-rankable equilibria, a Stackelberg solution is a payoff-dominant equilibrium. The Hi-Lo Matching game is the simplest example of this.

Stackelberg reasoning turns out to imply a form of evidential decision theory (Eells 1985; Horgan 1981; Jeffrey 1983; Nozick 1969; 1993, Ch. 2) that departs from both standard decision theory and causal decision theory. It involves maximizing *conditional* EU – conditioned on the strategy chosen – rather than *causal* EU, as in causal decision theory, or standard EU as in plain vanilla decision theory. Nozick's (1993, Ch. 2) careful discussion of Newcomb's problem suggests that evidential reasoning is characteristic of human thought in certain circumstances, and there is experimental evidence that it occurs (Anand 1990; Quattrone & Tversky 1984).

8.3. Evidence of Stackelberg reasoning

More specifically, Colman and Stirk (1998) reported experimental evidence of Stackelberg reasoning. Their 100 decision makers played all 12 ordinally nonequivalent, symmetric, 2×2 games, nine of which (including Prisoner's Dilemma and Stag Hunt) were Stackelberg soluble and three (including Battle of the Sexes and Chicken) were not. Players chose an overwhelming preponderance (78 to 98 percent) of Stackelberg strategies in the Stackelberg-soluble games but showed no significant biases and very small effect sizes in the non-Stackelberg-soluble games. A protocol analysis revealed that joint payoff maximization was a significantly more frequent reason for choice in the Stackelberg-soluble than the non-Stackelberg-soluble games. These results were replicated in a later study with 3×3 games (Colman et al. 2001). Taken together, these findings suggest that human decision makers may be influenced by Stackelberg reasoning, at least in simple games.

Both team reasoning and Stackelberg reasoning may help to explain the payoff-dominance phenomenon. In addition, Stackelberg reasoning can be shown to predict focal-point selection in pure coordination games in general (Colman 1997). Team reasoning may even offer a partial explanation for cooperation in social dilemmas. Any alternative explanations of these phenomena would have to invoke other nonstandard psychological game-theoretic processes that have yet to be discovered.

8.4. Epistemic and non-monotonic reasoning

The backward induction problem (sect. 7) requires an entirely different type of solution. In the Centipede game, backward induction, in conjunction with the CKR assumptions, seems to lead to an impasse in which neither player can act rationally. Epistemic reasoning – reasoning about knowledge and belief – in backward induction has been reexamined by Antonelli and Bicchieri (1994), Bicchieri (1993, Chs. 3, 4), Bonanno (1991), and Samet (1996), among others. Backward induction makes sense only if the players have no doubts about each other's rationality. To clear the logjam, perhaps common knowledge needs to be replaced by something like *common beliefs* (Monderer & Samet 1989) or *entrenched common beliefs* (Sugden 1992). Perhaps common knowledge of rationality, in particular, needs to be weakened to common belief. A proposition is a matter of common belief if each player believes it to be true, believes that the other player(s) believe(s) it to be true, and so on, and if each player continues to believe it only as long as it generates no inconsistency. Sugden showed that replacing common knowledge of rationality with entrenched common belief allows a player to evaluate the EU of cooperating, thus clearing the impasse, but the cure may be worse than the disease, as we shall see.

The appropriate formal apparatus for this type of analysis is *non-monotonic reasoning,* in which premises may be treated as default assumptions, subject to belief revision in the light of subsequent information. The replacement of common knowledge of rationality with common belief in rationality, as a default assumption, does not prevent the backward induction argument from going through (Colman 1998). The players remain instrumentally rational, and their default assumptions about their co-players' rationality need not be revised. Hence, in the Centipede game, Player I still defects down on the first move. The crucial difference is that it is no longer impossible for Player I to perform the necessary reasoning with subjunctive conditionals (*If I were to make a cooperative opening move, then . . .*) to judge how

Player II would respond to a deviation from the backward induction path.

Player I may calculate that a cooperative opening move would force Player II to abandon the default assumption that Player I is rational, attributing this to Player I's "trembling hand" (Selten 1975) or to some more serious cognitive limitation. Player II would choose a utility-maximizing reply in the light of these changed circumstances, defecting down at the second decision node if Player I was expected to respond by defecting down at the third. But Player II may reasonably expect Player I, who has already cooperated once, to do it again, given the chance. In that case, Player II would respond cooperatively, and if Player I anticipated this reasoning, that would provide a *rational* justification for choosing a cooperative opening move. A sequence of reciprocally cooperative moves could ensue, to the mutual benefit of the players.

This directly contradicts the conclusion of the backward induction argument. In place of an impasse, we now have a contradiction, suggesting either that the backward induction argument is flawed or that the revised assumptions are incoherent. Furthermore, it is debatable whether a rational Player I could make a cooperative move, as distinct from merely contemplating its implications, without violating the CKR2 rationality assumption. It is important to remember that the rationality assumption itself has not been weakened. What has made the difference is a weakening of the common *knowledge* of rationality assumption. This was necessary to clear the impasse and enable the players to evaluate the relevant expected utilities.

This discussion has no doubt muddied the waters, but it seems to have confirmed that, in some circumstances at least, strategic interaction is not rational in the official game-theoretic sense of the word.

9. Conclusions

Rationality has a clear interpretation in *individual* decision making, but it does not transfer comfortably to *interactive* decisions, because interactive decision makers cannot maximize expected utility without strong assumptions about how the other participant(s) will behave. In game theory, common knowledge and rationality assumptions have therefore been introduced, but under these assumptions, rationality does not appear to be characteristic of social interaction in general.

9.1. Failures of game-theoretic rationality

In pure coordination games, game-theoretic rationality cannot guide players to focal points, nor show why focal points are often intuitively obvious choices, nor explain how human players choose them in practice. Even in a game with an idiot-proof payoff-dominant equilibrium, game-theoretic rationality stubbornly fails to mandate its selection. Cooperative games, in which players can form coalitions sanctioned by binding and enforceable agreements, often have empty cores (see sect. 5.2 above), and rational social interaction is therefore impossible in such games, at least according to the leading solution concept of the core. In social dilemmas, game-theoretic rationality is self-defeating, and experimental players frequently violate it.

In certain sequential games, game theory yields powerfully counterintuitive prescriptions, but under closer examination the theory appears to generate contradictions, rendering it impotent as a guide to rational choice. Human experimental players tend to deviate from the backward induction path, behaving more cooperatively and earning higher payoffs as a consequence. These problems and paradoxes can be solved only by introducing forms of psychological game theory incorporating nonstandard types of reasoning.

9.2. Final evaluation

Game theory's contribution to the study of interactive decision making is incalculable, and it is indubitably one of the most powerful and successful theories in the behavioral and social sciences. But the project initiated by von Neumann and Morgenstern (1944) to characterize rationality as individual expected utility maximization has created profound problems for game theory, in both its normative and its positive interpretations. Although game theory has vastly increased our understanding of interactive decision making, and no alternative theory even comes close to challenging its power, the conception of rationality on which it rests appears to be internally deficient and incapable of explaining social interaction. Psychological game theory, some examples of which were outlined above, may help to provide a way forward.

ACKNOWLEDGMENTS

I am grateful to Ken Hughes, who contributed substantially to this article in email dialogue with me but declined to coauthor it. Thanks are due also to John Nash, who suggested the term "psychological game theory," to other participants at the Fifth SAET Conference in Ischia, and to Michael Bacharach and other members of the Oxford Experimental Economics Workshop for their helpful comments. I wish to dedicate this article to the memory of Michael Bacharach who died while it was in press. I am grateful also to Ken Binmore, Oliver Board, John Lazarus, and R. Duncan Luce for specific comments on an earlier draft of the article, preparation of which was facilitated by a period of study leave granted to me by the University of Leicester.

NOTES

1. According to a debatable behavioral interpretation of RCT (Herrnstein 1990), its central assumption is that organisms maximize reinforcement, and this "comes close to serving as the fundamental principle of the behavioral sciences" (p. 356). But, as Herrnstein pointed out, experimental evidence suggests that RCT, thus interpreted, predicts human and animal behavior only imperfectly.

2. Green and Shapiro (1994) did not touch on this most crucial problem in their wide-ranging critical review of the rational choice literature. Neither did any of the participants in the ensuing "rational choice controversy" in the journal *Critical Review,* later republished as a book (Friedman 1996).

3. This implies that rational beliefs have to respect the (known) evidence. For example, a person who has irrefutable evidence that it is raining and therefore knows (believes that the probability is 1) that it is raining, but also believes that it is fine, fails to respect the evidence and necessarily holds internally inconsistent beliefs.

4. I am grateful to Ken Binmore for pointing this out to me.

5. Kreps (1988) has provided an excellent summary of Savage's theory, although Savage's (1954) own account is brief and lucid.

6. To a psychologist, revealed preference theory explains too little, because there are other sources of information about preferences apart from choices, and too much, because there are other

factors apart from preferences that determine choices – see the devastating "rational fools" article by Sen (1978).

7. The possibility of systematic irrationality, or of demonstrating it empirically, has been questioned, notably by Broome (1991), Cohen (1981), and Stein (1996).

8. See, for example, Bicchieri (1993, Chs. 2, 3); Colman (1997; 1998); Cubitt and Sugden (1994; 1995); Hollis and Sugden (1993); McClennen (1992); Sugden (1991b; 1992). The common knowledge assumptions are sometimes relaxed in recent research (e.g., Aumann & Brandenburger 1995).

9. In other circumstances, experimental evidence suggests that human reasoners do not even come close to full common knowledge (Stahl & Wilson 1995).

10. The theory is determinate for every strictly competitive (finite, two-person, zero-sum) game, because if such a game has multiple equilibria, then they are necessarily equivalent and interchangeable, but this does not hold for other games.

11. See Colman (1995a, pp. 169–75) for a simple example of an empty core in Harold Pinter's play, *The Caretaker*.

12. Even Hume nods. Port comes from Portugal, of course.

13. Janssen's (2001b) principle of *individual team member rationality* is slightly weaker (it does not require equilibrium): "If there is a unique strategy combination that is Pareto-optimal, then individual players should do their part of the strategy combination" (p. 120). Gauthier's (1975) *principle of coordination* is slightly stronger (it requires both equilibrium and optimality): "In a situation with one and only one outcome which is both optimal and a best equilibrium . . . it is rational for each person to perform that action which has the best equilibrium as one of its possible outcomes" (p. 201).

14. If e and f are any two equilibrium points in a game, then e risk-dominates f if and only if the minimum possible payoff resulting from the choice of the strategy corresponding to e is strictly greater for every player than the minimum possible payoff resulting from the choice of the strategy corresponding to f. According to Harsanyi and Selten's (1988) risk-dominance principle, if one equilibrium point risk-dominates all others, then players should choose its component strategies. It is used when subgame perfection and payoff dominance fail to yield a determinate solution.

15. According to the sure-thing principle, if an alternative a_i is judged to be as good as another a_j in all possible contingencies that might arise, and better than a_j in at least one, then a rational decision maker will prefer a_i to a_j. Savage's (1954) illustration refers to a person deciding whether or not to buy a certain property shortly before a presidential election, the outcome of which could radically affect the property market. "Seeing that he would buy in either event, he decides that he should buy, even though he does not know which event will obtain" (p. 21).

16. I am grateful to Werner Güth for this insight.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Cooperation, evolution, and culture

Michael Alvard

Department of Anthropology, Texas A&M University, College Station, Texas 77843-4352. **alvard@tamu.edu** **http://people.tamu.edu/~alvard/**

Abstract: Rejecting evolutionary principles is a mistake, because evolutionary processes produced the irrational human minds for which Colman argues. An evolved cultural ability to acquire information socially and infer other's mental states (mind-reading) evokes Stackelberg reasoning. Much of game theory, however, assumes away information transfer and excludes the very solution that natural selection likely created to solve the problem of cooperation.

Colman rejects the relevancy of evolutionary game theory to his argument that rationality is not a general characteristic of human social interaction. Although the evolutionary process of natural selection is indeed mindless, as Colman notes, it is useful to consider that mindless evolutionary processes produced human minds. Human minds, and the behaviors that they produce, remain the focus of our interests. If people are less than rational in social interactions, as Colman suggests, it is because we evolved that way. The fact that rationality leads to inferior outcomes in social dilemmas, compared to alternative forms of reasoning, lends support to the idea that selection might have favored something other than strict rationality in our evolutionary past. Many of the ad hoc principles of psychological game theory introduced at the end of the target article might be deductively generated from the principles of evolutionary theory.

An evolutionary approach encourages the use of the comparative method. The ability of humans to cooperate to achieve common goals is nearly unique among animals and is perhaps matched in scope only by the social insects (Hill 2002). While insects accomplish their collectivity through rigid genetic rules, there is much to suggest that we are able to achieve our level of ultrasociality via cultural mechanisms (Richerson & Boyd 2001). Exactly how humans accomplish this is one of the key questions of the social sciences.

Researchers interested in the evolution of cultural abilities – culture is defined as the social transmission of information – should be particularly intrigued by the issues related to coordination that Colman raises. Among other advantages, cultural mechanisms provide people the ability to infer each other's mental states, to preferentially assort with others who have similar (or complementary) intentions or capabilities, and to reap the advantages of coordinated activities (Alvard, in press; Boyd & Richerson 1996; Tomasello 1999). Focal point selection is facilitated by cultural mechanisms that create shared notions among individuals. Colman hints at this himself when he says, "To remove . . . the culturally determined labeling of strategies, is to filter out the focal points" (target article, sect. 5.4, last para.). Having shared notions greatly enhances the ability to solve simple yet common and important coordination games.

The forms of psychological games Colman describes as alternatives to the classic game forms depend on psychological expectations. Stackelberg reasoning, for example, involves anticipating the other player's choices. Such reasoning requires a sophisticated theory of mind where others are viewed as intentional agents. It also suggests the related concept of mind-reading (Baron-Cohen 1995; Cheney & Seyfarth 1990). Not nearly as mysterious as it

sounds, though perhaps a uniquely human capability, mind-reading is the ability to reason about the otherwise unobservable mental states of others and make predictions about their behaviors based partly on the awareness that others are intentional agents with general goals similar to one's own. Colman uses the phrasing of mind-reading in his description of how a Stackelberg-reasoning player might deliberate.

It seems that cultural mechanisms solve cooperative problems so transparently, however, that many do not recognize them as solutions at all. Broadly speaking, communicating via spoken language can be construed as mind-reading. I can utter a sound and others can predict my intent based on that sound, unless they do not share my otherwise arbitrary association between sound and meaning. Part of the problem of classic analytic game theory revolves around the standard assumption that players do not speak to one another. This assumption is put into practice in experimental game research where subjects usually do not communicate during experiments. It seems that pre-game communication among subjects is such a simple solution to many games that researchers routinely disallow it in order for the "truly" interesting solutions to emerge (van Huyck et al. 1990). Although "cheap talk" solutions may seem trivial to game theoreticians, because all extant humans can easily communicate this way, from a comparative evolutionary perspective, such a solution is far from trivial. Although simple communication among players is often sufficient to generate complexly coordinated behaviors, speaking is anything but simple. Such a research design excludes the very solution that natural selection likely created to solve the problem. Experimental social games in which subjects are not allowed to speak to one another are a bit like sports competitions where subjects must compete with their legs shackled together.

Verbalizing intent may be feasible in small groups, but how do humans communicate expectations between members of large cooperative groups like those that characterize most human societies – ethnic groups, for example – in which many interactions are seemingly anonymous? How do fellows know that they share beliefs concerning behavior critical for coordination? How can individuals predict what others think and will do in such large groups? There are a number of options. One could attempt to learn, on one's own, the beliefs of all the potential cooperative partners. This could prove difficult, time-consuming, and error-prone (Boyd & Richerson 1995). In addition to speaking, however, humans use symbols and markers of group identity to transmit information that helps them make predictions about the otherwise unobservable mental states of others. McElreath et al. (2003) argue that group markers, such as speech or dress, function to allow individuals to advertise their behavioral intent so that individuals who share social norms can identify one another and assort for collective action. Although cheaters are a problem if interaction is structured like a prisoner's dilemma, McElreath et al.'s critical point is that group markers are useful if people engage in social interactions structured as coordination games. Colman notes the advantages of making predictions about the behavior of others based on information acquired culturally.

The great potential payoffs from successfully navigating real-life coordination games may have been part of the selective pressure favoring the evolution of language and culture. Coordination problems abound, and their solutions are facilitated when players have the ability to quickly acquire expectations about fellow players' behavior. Whether such adaptations are rational or not, ignoring the evolutionary mechanisms that produced these cognitive abilities is a mistake.

Humans should be individualistic and utility-maximizing, but not necessarily "rational"

Pat Barclay and Martin Daly

Department of Psychology, McMaster University, Hamilton, Ontario, L8S 4K1 Canada. **barclapj@mcmaster.ca** **daly@mcmaster.ca**
http://www.science.mcmaster.ca/Psychology/md.html

Abstract: One reason why humans don't behave according to standard game theoretical rationality is because it's not realistic to assume that everyone else is behaving rationally. An individual is expected to have psychological mechanisms that function to maximize his/her long-term payoffs in a world of potentially "irrational" individuals. Psychological decision theory has to be individualistic because individuals make decisions, not groups.

Game theoretical rationality in the service of personal profit maximization is not an adequate model of human decision-making in social bargaining situations. This proposition is a large part of Colman's thesis, and we have no quarrel with it. Does anyone? The point is proven whenever experimental subjects reject offers in Ultimatum Games, share the pot in Dictator Games, or cooperate in one-shot Prisoner's Dilemmas (e.g., Frank et al. 1993; Roth 1995). The idea that this simple game theoretical account is descriptive rather than normative is surely dead in experimental economics and psychology. Evolutionary models are an important exception, because they purport to describe what strategies will be selected for. However, in these models, the concept of "rationality" is superfluous, because the selection of superior strategies occurs by a mindless, competitive process (Gintis 2000).

One way in which rational choice theory (RCT) is problematic is the default expectation that all other players will behave "rationally." Individuals can be expected to occasionally make decisions that are not in accordance with predictions of RCT because of incomplete information, errors, concern for the welfare of others (such as friends or relatives), or manipulation by others. Also, individuals may be expected to act irrationally if that irrationality is more adaptive than rationality. For example, Nowak et al. (2000) show that the "irrational" behavior of demanding fair offers in the Ultimatum Game is evolutionarily stable if each individual has knowledge about what kind of offers each other individual will accept. Similarly, aggressive behavior or punishment, while not "rational" in the game theoretic sense, can evolve if the costs of being punished are high (Boyd & Richerson 1992), because the punished individual learns (potentially via operant conditioning) to desist from the behavior that brought on the punishment.

Given that others are sometimes not strictly rational, an instrumentally rational individual should reevaluate his/her situation and act accordingly (Colman hints at this in sect. 8.4). We argue that rationality should not even be the default assumption because individuals are repeatedly faced with evidence (from real life) that others are not always rational, and this affects the strategy that a profit-maximizing individual should take. For example, when playing an iterated Prisoner's Dilemma against what appears to be a conditional cooperator (such as a Tit-for-Tat player), a rational and selfish player should cooperate for a while. Even if the rational actor is cheated in later rounds, he/she has still done better than if he/she had never cooperated. Thus, a rational player should attempt to determine the likely responses of others, rather than assume that (despite past experience to the contrary) they will be "rational." Henrich et al. (2001) argue that when people play experimental games, they compare the games to analogous situations with which they have experience. If different people have had different experiences because of different backgrounds, then they will have different beliefs about how others will behave. Thus, in iterated games, each player may be acting rationally with respect to his/her past experience. Recent experiences have large effects on how people play experimental games (Eckel & Wilson 1998a), possibly because players use their experience in the games to update their beliefs of what others will do.

This does not explain behavior in one-shot games, but we would

not expect the human psyche to have evolved to deal with one-shot games. Under normal circumstances (and especially before the dawn of the global village), one can rarely (if ever) be absolutely certain that one will never interact with the same person again. Even if two individuals never interact again, others may observe the interaction (Alexander 1987). For this reason, humans may have internal rewards for acting cooperatively in repeated interactions, which evolved (or were learned) because those rewards caused people to cooperate and reap the benefits of mutual cooperation. These internal rewards (or nonstandard preferences such as a positive valuation of fairness, equity, or the well-being of individual exchange partners) would also cause them to act cooperatively in the novel case of one-shot interactions as well.

The target article's more contentious claim is that game theoretical rationality cannot be salvaged as a model of human decision-making in social situations by incorporating nonstandard preferences into the decision makers' utility functions. In section 8.1, Colman illustrates the problem of finding compromise solutions where individual preferences differ with Sugden's example of a family going for a walk, and asserts that tinkering with utility functions cannot explain their solubility. He insists that the "team reasoning" by which compromises are negotiated is an "inherently non-individualistic" process. However, we looked in vain for evidence or argument in support of these conclusions. It is, after all, individuals who ultimately make the choices in experimental games, so if "a team reasoning player" really seeks to maximize "joint or collective payoff," as Colman claims (sect. 8.1), then contra his own conclusion (sect. 9.1), this is evidence of nonstandard preferences, not of "nonstandard types of reasoning." Moreover, such a process of team reasoning cannot have general applicability to social dilemmas with divisible payoffs, because it is inconsistent with the evidence that experimental subjects will pay to punish other players (Fehr & Gächter 2000; Roth 1995). We do not understand the notion of a "psychological" theory that is "non-individualistic"; the individual organism is psychology's focal level of analysis.

We agree with Colman in saying that game theoretic rationality does not accurately describe human social behavior. However, he has not argued convincingly why expanding calculations of Expected Utility to include nonstandard preferences and rational responses to irrational behavior cannot salvage models of Expected Utility, so we would argue that such expanded models still may be effective at explaining human behavior. Evolutionary models can help generate hypotheses about what those nonstandard preferences are, and how we might expect people to respond to apparently irrational behavior.

ACKNOWLEDGMENTS

We would like to thank L. DeBruine and M. Wilson.

Neural game theory and the search for rational agents in the brain

Gregory S. Berns

Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA 30322. **gberns@emory.edu**
http://www.ccnl.emory.edu

Abstract: The advent of functional brain imaging has revolutionized the ability to understand the biological mechanisms underlying decision-making. Although it has been amply demonstrated that assumptions of rationality often break down in experimental games, there has not been an overarching theory of why this happens. I describe recent advances in functional brain imaging and suggest a framework for considering the function of the human reward system as a discrete agent.

The assumption of rationality has been under attack from several fronts for a number of years. Colman succinctly surveys the evidence against rationality in such ubiquitous decision games ranging from the Prisoner's Dilemma (PD) to Centipede and comes to the conclusion that standard game theory fails to explain much of human behavior, especially within the confines of social interactions. This is a startling conclusion, and not merely because the tools of game theory have become so enmeshed as an approach to decision-making and risk management. The startling implication is that almost every commonplace decision that humans make is socially constrained. Whether it is an explicit social interaction like the PD or an implicit social construct that underlies the decision to stay a few more hours at work versus spending time with family, social connectedness cannot be factored out of almost any meaningful human decision. If the assumption of rationality governs the application of game theoretic tools to understanding human decision-making, then its very failure in social domains brings into question its practical utility. Can the tools of game theory be applied within ad hoc frameworks like behavioral game theory, or psychological game theory? The reunification of psychological principles within economics is a necessary first step (Camerer 1999). However, like cognitive psychology before it, psychological principles also often fail to explain human behavior. Neuroscience offers yet another perspective on human behavior, and the application within economic frameworks has come to be called neural economics (Montague & Berns 2002; Wilson 1998).

One of the earliest applications of functional brain imaging, in this case functional magnetic resonance imaging (fMRI), to the evaluation of the neural basis of game theory was performed by McCabe and colleagues (McCabe et al. 2001). In a variant of Centipede, pairs of subjects played three types of games (trust, punish, and mutual advantage). As Colman points out, most players behave cooperatively in these types of games – an observation unexplainable by standard rational game theory. McCabe's results implicated a specific region of the medial prefrontal cortex that subserved this type of cooperative behavior. Perhaps because of the small sample size ($N = 6$), statistically significant results were not found for the noncooperators (i.e., "rational" players), and nothing could be said about the possible existence of rational agents in the brain.

In a recent study, our group used fMRI to examine the neural responses of one player in an all-female PD interaction (Rilling et al. 2002). The large sample size ($N = 36$) yielded reasonable power to detect a number of significant activations related to the different outcomes. At the simplest level, striatal activation was most strongly associated with mutual cooperation outcomes. When subjects played a computer, the striatal activation was greatly reduced, suggesting that the striatal activity was specifically modulated by the presence or absence of social context. This region of the striatum is of particular interest because it is the same region most closely associated with hedonic reward processes (Schultz et al. 1997). Activation of the ventral striatum, especially the nucleus accumbens, has been observed repeatedly in various forms of appetitive Pavlovian conditioning and drug administration – activity that is widely believed to be modulated by dopamine release (Robbins & Everitt 1992). The striatal activation observed with mutual cooperation most likely reflected the overall utility of that outcome in the context of the PD. The same region of striatum was also observed to be active during the decision-making phase of the experiment, but only when the subject chose to cooperate following her partner's cooperation in the previous round. This latter finding suggests that the striatum was not only encoding the actual utility of the outcome, but the expected utility during the decision-making phase. We do not yet know the exact relationship between reward-system activity and expected utility (modified or not), but the mesolimbic reward system appears to be a promising candidate for a "rational agent" within the brain.

Based on the results of the PD experiment, our group realized that it would be desirable to monitor brain activity in both players simultaneously. The rationale is that by monitoring the activity in the reward pathways of both players in a two-player game, one should have a direct assay of the player's expected utility functions

without resorting to revealed preference. Under rational assumptions, these functions should be about the same. In a proof of principle experiment, based loosely on the game of matching pennies, we described the methodology necessary to conduct such simultaneous imaging experiments, which we have termed, "Hyperscanning" (Montague et al. 2002). In this first experiment, we did not undertake an assessment of utility functions, but it is worth pointing out that the methodology is generalizable to *N*-person interactions.

There is good reason to be hopeful that neuroscientific methods, especially functional brain imaging, will help resolve the apparent paradoxes between rational game theory and both behavioral and psychological variants. By looking inside the brain, it becomes possible to identify specific neuronal clusters that may be operating near different equilibria. Recent work suggests that neurons in the lateral intraparietal area encode expected utilities during probabilistic reward paradigms in monkeys (Glimcher 2002; Gold & Shadlen 2001). In humans, correlates of utility, as predicted by prospect theory, have been found in discrete elements of the reward pathway (Breiter et al. 2001). Taken together, these early neuroscientific enquiries suggest that game theoretic principles are very much viable predictors of neuronal behavior. The interaction of different pools of neurons in the brain may result in phenotypic behavior that appears to be irrational, but it is possible that the rational agents are the neurons, not the person.

ACKNOWLEDGMENTS

The author is supported by grants from the National Institutes of Health (K08 DA00367, RO1 MH61010, and R21 DA14883).

Evolution, the emotions, and rationality in social interaction

David J. Butler

Department of Economics, University of Arizona, Tucson, AZ, 85721 and Department of Economics, University of Western Australia, Nedlands, WA 6009, Australia. **dbutler@eller.arizona.edu**

Abstract: Although Colman's criticisms of orthodox game theory are convincing, his assessment of progress toward construction of an alternative is unnecessarily restrictive and pessimistic. He omits an important multidisciplinary literature grounded in human evolutionary biology, in particular the existence and function of social emotions experienced when facing some strategic choices. I end with an alternative suggestion for modifying orthodox game theory.

Colman has brought together an impressive collection of arguments to demonstrate both serious weaknesses and failures of orthodox game-theoretic rationality. But to address these problems he offers only some "tentative and ad hoc suggestions" (sect. 8, para. 2) from psychological game theory. Although I strongly endorse his criticisms of orthodox game theory and agree that the new reasoning principles he describes have a part to play, I think his discussion of "where next" neglects some important ideas from a recent and exciting multidisciplinary literature.

Because of the newness of this research and its multidisciplinary origins, we must piece together some apparently disparate strands of thought in order to glimpse the beginnings of an alternative to orthodox game-theoretic rationality. One reason why Colman's "destruction" work is much more comprehensive and convincing than his subsequent "construction" is his early distinction between the nonrational "mindless" (sect. 1.3, para. 3) strategic interaction of evolutionary game theory, and the rational strategic interaction of human agents. He argues the former is not of interest for his views on rationality, but I will argue that this dichotomy severely restricts the variety of new ideas that he can consider.

To understand human decision-making in social interactions, we should keep in mind that both humans and their decision-making apparatus are themselves products of natural selection. There is a growing consensus behind the "social animal" hypothesis (e.g., Barton & Dunbar 1997; Cartwright 2000), which maintains that the selection pressures among humans were primarily an intraspecies phenomenon. In successive generations, reproductive success went to those with the best grasp of the complexities of the "social chess" that was a constant theme of tribal life. In this view, underlying the anomalous cooperation observed in both experimental and real world social dilemmas is an innate predisposition, not for unconditional cooperation, but for some form of reciprocity. Indeed, there is a now a significant literature in experimental and theoretical economics on reciprocity models (see Sethi & Somanathan 2003 for a recent survey).

Trivers (1985) argued that reciprocal altruism in humans evolved by molding our emotional responses to the cost/benefit calculus of social exchange; among these emotions are both cooperative and punitive sentiments. In a recent study, Price et al. (2002) demonstrate that "punitive sentiments in collective action contexts have evolved to reverse the fitness advantages that accrue to free riders over producers." Indeed, punitive sentiments must go hand in hand with a preparedness to risk cooperation if cooperation is to survive the process of natural selection.

There is also a growing recognition that contrary to the standard model of rational choice, "gut feelings experienced at the moment of making a decision, which are often quite independent of the consequences of the decision, can play a critical role in the choice one eventually makes" (Loewenstein et al. 2001). For example, they refer to the work of the neuroscientist Damasio (1994), who shows how our ability to choose rationally is intertwined with our ability to experience emotional reactions to the choices we face. Damasio calls these reactions "somatic markers" and argues: "Nature appears to have built the apparatus of rationality (the cerebral cortex) not just on top of the apparatus of biological regulation (the limbic system), but also from it and with it" (p. 128). A more human rationality may also allow for heterogeneity of choices, in recognition of the differing intensities with which the social (and other) emotions are experienced by different people in the deliberation process.

Although neither Damasio nor Loewenstein and colleagues directly address the social emotions, we can easily extend their arguments to the context of strategic interaction, where the emotions that need incorporating for a descriptive theory are the cooperative and punitive sentiments behind reciprocity. We might even go further and argue for their incorporation into normative models, as well. This is because our emotional responses to choices that place our individual and collective interests in opposition embody adaptive knowledge that helped win many games of "social chess." These somatic responses may help us to extract the long run benefits of cooperation.

There is also now direct evidence that a somatic response specific to human strategic interactions exists. Recent work by Rilling et al. (2002), using fMRI scans on subjects playing prisoner's dilemma games, found that an individual's brain activation patterns when the playing partner was identified as a human differ from when the partner was identified as a computer. They conclude "that (the relevant activation patterns) may relate specifically to cooperative social interactions with human partners." It seems that human players rely more on a common knowledge of *humanity* in strategic interaction than a common knowledge of *rationality* as conventionally understood.

The finding of Rilling and colleagues also highlights the importance of the description or "framing" of the game for our choices. Loewenstein and colleagues also noted, for choice under risk, that these factors become important when we incorporate emotions experienced when choosing, in contrast to the purely cognitive evaluations of the standard model that are supposedly context independent. This implies we can no longer expect game theoretic models to satisfy description invariance if a change in the description (e.g., that the other player is a person or a program) is implemented.

Colman discusses "Stackelberg reasoning" and "team thinking," and he mentions (sect. 8.1, para. 3) that the collective preferences of team reasoning can be triggered by the acceptance of a group identity in certain contexts. But he doesn't explain where these alternative reasoning methods come from, how they survive, or how, if cooperation in social dilemmas is sensitive to the cost/benefit ratio, we might "trade-off" the different reasoning methods in some meta-reasoning process. Hamilton's (1964) "kin-selection," Trivers' (1971) "reciprocal altruism," and Alexander's (1987) "indirect reciprocity" models might at least offer a way to think about answering these questions.

If we wish to incorporate the social emotions triggered by a strategic choice into our models, how might we proceed? Hollis and Sugden (1993) explained (p. 28) why our attitudes toward consequences cannot be simply "bundled in" with the existing utilities of a game. A more plausible path then may be to alter the weighting we attach to the consequences, along the lines of the "rank dependent" transformation of the cumulative probability distribution, which has worked so well among the alternatives to expected utility theory (see Starmer 2000). In this way, some plausible improvements to orthodox game theory might be developed, as has already happened to expected utility theory in choice under risk.

Behavioral game theory: Plausible formal models that predict accurately

Colin F. Camerer

Division of Social Sciences, California Institute of Technology, Pasadena, CA 91125. **camerer@hss.caltech.edu** **http://hss.caltech.edu/~camerer**

Abstract: Many weaknesses of game theory are cured by new models that embody simple cognitive principles, while maintaining the formalism and generality that makes game theory useful. Social preference models can generate team reasoning by combining reciprocation and correlated equilibrium. Models of limited iterated thinking explain data better than equilibrium models do; and they self-repair problems of implausibility and multiplicity of equilibria.

Andrew Colman's wonderful, timely, and provocative article collects several long-standing complaints about game theory. Part of the problem is that game theory has used lots of applied math and little empirical observation. Theorists think that deriving perfectly precise analytical predictions about what people will do (under differing assumptions about rationality) from pure reasoning is the greatest challenge. Perhaps it is; but why is this the main activity? The important uses of game theory are prescriptive (e.g., giving people good advice) and descriptive (predicting what is likely to happen), because good advice (and good design of institutions) requires a good model of how people are likely to play. It is often said that studying analytical game theory helps a player understand what might happen, vaguely, even if it does not yield direct advice. This is like saying that studying physics helps you win at pool because the balls move according to physical laws. A little physics probably doesn't hurt, but also helps very little compared to watching other pool players, practicing, getting coaching, studying what makes other players crumble under pressure, and so on.

While Colman emphasizes the shortcomings of standard theory, the real challenge is in creating new theory that is psychological (his term) or "behavioral" (my earlier term from 1990; they are synonymous). Models that are cognitively plausible, explain data (mostly experimental), and are as general as analytical models, have developed very rapidly in just the last few years. Colman mentions some. Others are described in my book (Camerer 2003).

An important step is to remember that games are defined over utilities, but in the world (and even the lab) we can usually only measure pecuniary payoffs – status, territory, number of offspring, money, and so forth. The fact that people cooperate in the prisoner's dilemma (PD) is *not* a refutation of game theory per se; it is a refutation of the joint hypothesis of optimization (obeying dominance) and the auxiliary hypothesis that they care only about the payoffs we observe them to earn (their own money). The self-interest hypothesis is what's at fault.

Several new approaches to modeling this sort of "social preferences" improve on similar work by social psychologists (mentioned in sect. 8.1), because the new models are designed to work across games and endogenize when players help or hurt others. For example, in Rabin's fairness theory, player A treats another player's move as giving herself (A) a good or bad payoff, and forms a judgment of whether the other player is being nice or mean. Players are assumed to reciprocate niceness and also meanness.

Rabin's model is a way to formalize *conditional* cooperation – people cooperate if they expect others to do so. This provides a way to anchor the idea of "team reasoning" in methodological individualism. In experiments on group identity and cooperation, a treatment (like subjecting subjects to a common fate or dividing them into two rooms) or categorization (whether they like cats or dogs better) is used to divide subjects into groups. In the Rabin approach, PD and public goods games are coordination games in which players are trying to coordinate on their level of mutual niceness or meanness.

Experimental identity manipulations can be seen as correlating devices that tell subjects which equilibrium will be played, that is, whether they can expect cooperation from the other players or not (which is self-enforcing if they like to reciprocate). This explanation is *not* merely relabeling the phenomenon, because it makes a sharp prediction: A correlated equilibrium requires a publicly observable variable that players commonly know. If identity is a correlating device, then when it is not commonly known, cooperation will fall apart. For example, suppose members of the A team ("informed A's") are informed that they will play other A's, but the informed As' partners will not know whether they are playing A's or B's. Some theories of pure empathy or group identification predict that who the other players think they are playing won't matter to the informed A's because they just like to help their teammates. The correlated equilibrium interpretation predicts that cooperation will shrink if informed A's know that their partners don't know who they are playing, because A's only cooperate with other A's *if they can expect cooperation by their partners.* So there is not necessarily a conflict between an individualist approach and team reasoning: "Teamness" can arise purely through the conjunction of reciprocal individual preferences and observable correlating variables, which create shared beliefs about what team members are likely to do. What those variables are is an interesting empirical matter.

Another type of model weakens the mutual consistency of players' choices and beliefs. This might seem like a step backward but it is not – in fact, it solves several problems that mutual consistency (equilibrium) *creates.* In the cognitive hierarchy (CH) model of Camerer et al. (2002), a Poisson distribution of discrete levels of thinking is derived from a reduced-form constraint on working memory. Players who use 0 levels will randomize. Players at level $K > 0$ believe others are using 0 to K to 1 levels. They know the normalized distribution of lower-level thinkers, and what those others do, and best respond according to their beliefs. The model has one parameter, τ, the average number of levels of thinking (it averages around 1.5 in about a hundred games). In the CH model, every strategy is played with positive probability, so there are no incredible threats and odd beliefs after surprising moves. Once τ is fixed (say 1.5), the model produces an exact statistical distribution of strategy frequencies – so it is *more* precise in games with multiple equilibria, and is generally *more* empirically accurate than equilibrium models. The model can explain focal points in matching games if level-0 subjects choose what springs up. The model also has "economic value": If subjects had used it to forecast what others were likely to do, and best responded to the model's advice, they would have earned substantially more (about

a third of the economic value of perfect advice). Nash equilibrium, in contrast, sometimes has negative economic value.

Beliefs, intentions, and evolution: Old versus new psychological game theory

Jeffrey P. Carpenter and Peter Hans Matthews

Economics Department, Middlebury College, Middlebury, VT 05753.
jpc@middlebury.edu peter.h.matthews@middlebury.edu
http://community.middlebury.edu/~jcarpent/
http://community.middlebury.edu/~pmatthew/

Abstract: We compare Colman's proposed "psychological game theory" with the existing literature on psychological games (Geanakoplos et al. 1989), in which beliefs and intentions assume a prominent role. We also discuss experimental evidence on intentions, with a particular emphasis on reciprocal behavior, as well as recent efforts to show that such behavior is consistent with social evolution.

Andrew Colman's target article is a call to build a new, psychological, game theory based on "nonstandard assumptions." Our immediate purpose is to remind readers that the earlier work of Geanakoplos et al. (1989), henceforth abbreviated as GPS, which the target article cites but does not discuss in detail, established the foundations for a theory of "psychological games" that achieves at least some of the same ends. Our brief review of GPS and some of its descendants – in particular, the work of Rabin (1993) and Falk and Fischbacher (2000) – will also allow us to elaborate on the connections between psychological games, experimental economics, and social evolution.

The basic premise of GPS is that payoffs are sometimes a function of both actions *and* beliefs about these actions, where the latter assumes the form of a subjective probability measure over the product of strategy spaces. If these beliefs are "coherent" – that is, the information embodied in second-order beliefs are consistent with the first-order beliefs, and so on – and this coherence is common knowledge, then the influence of second (and higher) order beliefs can be reduced to a set of common first-order beliefs. That is, in a two-player psychological game, for example, the utilities of A and B are functions of the strategies of each and the beliefs of each about these strategies. A psychological Nash equilibrium (PNE) is then a strategy profile in which, given their beliefs, neither A nor B would prefer to deviate, and these first-order beliefs are correct. If these augmented utilities are continuous, then all normal form psychological games must have at least one PNE.

The introduction of beliefs provides a natural framework for modeling the role of intentions in strategic contests, and this could well prove to be the most important application of GPS. It is obvious that intentions matter to decision-makers – consider the legal difference between manslaughter and murder – and that game theorists would do well to heed the advice of Colman and others who advocate a more behavioral approach.

For a time, it was not clear whether or not the GPS framework was tractable. Rabin (1993), which Colman cites as an example of behavioral, rather than psychological, game theory, was perhaps the first to illustrate how a normal form psychological game could be derived from a "material game" with the addition of parsimonious "kindness beliefs." In the standard two-person prisoner's dilemma (PD), for example, he showed that the "all cooperate" and "all defect" outcomes could *both* be rationalized as PNEs.

As Rabin (1993) himself notes, this transformation of the PD is not equivalent to the substitution of altruistic agents for self-interested ones: the "all defect" outcome, in which each prisoner believes that the other(s) will defect, could not otherwise be an equilibrium. This is an important caveat to the recommendation that we endow economic actors with "nonstandard reasoning processes," and prompts the question: What observed behavior will the "new psychological game theory" explain that an old(er) GPS-inspired one cannot? Or, in narrower terms, what are the shortcomings of game theoretic models that incorporate the role of intentions, and therefore such emotions as surprise or resentfulness?

The answers are not obvious, not least because there are so few examples of the transformation of material games into plausible psychological ones, and almost all of these share Rabin's (1993) emphasis on kindness and reciprocal behavior. It does seem to us, however, that to the extent that Colman's "nonstandard reasoning" can be formalized in terms of intentions and beliefs, there are fewer differences between the old and new psychological game theories than at first it seems.

There is considerable experimental evidence that intentions matter. Consider, for example, Falk et al. (2000), in which a first mover can either give money to, or take money away from, a second mover, and any money given is tripled before it reaches the second mover, who must then decide whether to give money back, or take money from, the first mover. Their analysis suggests that there is a strong relationship between what the first and second movers do: in particular, the more the first mover gives (takes), the more the second mover takes (gives) back.

Falk et al. (2000) find that first mover giving (taking) is interpreted as a friendly (unfriendly) act, and that these intentions matter. Without the influence of beliefs or intentions on utilities, there would be a single Nash equilibrium in which the first mover takes as much as possible because she "knows" that the second has no material incentive to retaliate. Although this behavior can also be supported as a PNE, so can that in which the first mover gives and expects a return and the second mover understands this intention and reciprocates. When the experiment is changed so that the first mover's choice is determined randomly, and there are no intentions for the second mover to impute, the correlation between first and second mover actions collapses. We see this as evidence that beliefs – in particular, intentions – matter, but also that once these beliefs have been incorporated, a modified "rational choice framework" is still useful.

Building on both GPS and Rabin (1993), Dufwenberg and Kirchsteiger (1998) and Falk and Fischbacher (2000) derive variations of Rabin's (1993) "fairness equilibrium" for extensive form games, with results that are consistent with experimental evidence. The simplest of these is the ultimatum game, in which a first mover offers some share of a pie to a second mover who must then accept or reject the proposal. With kindness functions similar to Rabin's (1993), Falk and Fischbacher (2000) show that the ultimatum game has a unique PNE that varies with the "reciprocity parameters" of proposer and responder. Furthermore, this equilibrium is consistent with the observations that the modal offer is half the surplus, that offers near the mode are seldom rejected, that there are few of the low offers that are consistent with the subgame perfect equilibrium, and that most of these low offers are rejected.

This result does *not* tell us, though, whether this outcome is consistent with the development of reciprocal intentions or norms over time, or, in other words, whether social evolution favors those with "good intentions." To be more concrete, suppose that the proposers and responders in the ultimatum game are drawn from two distinct populations and matched at random each period, and that these populations are heterogeneous with respect to intention. Could these intentions survive "selection" based on differences in material outcomes? Or do these intentions impose substantial costs on those who have them?

There are still no definitive answers to these questions, but the results in Binmore et al. (1995), henceforth abbreviated as BGS, hint that prosocial intentions will sometimes survive. BGS consider a "miniature ultimatum game" with a limited strategy space and show there are two stable equilibria within this framework. The first corresponds to the subgame perfect equilibrium – proposers are selfish, and responders accept these selfish offers – but in the second, proposers are fair and a substantial share of responders would turn down an unfair offer. Furthermore, these dy-

namics can be rationalized as a form of social or cultural learning: BGS emphasize the role of aspirations, but evolution toward fair outcomes is also consistent with imitation (Björnerstedt & Weibull 1996). It is tempting, then, to interpret the second BGS outcome as a Falk and Fischbacher (2000) "fairness equilibrium."

All of this said, we share most of Colman's concerns with standard game theoretic arguments, and suspect that psychological game theorists, both old and new, will have much to contribute to the literature.

ACKNOWLEDGMENTS

We thank Corinna Noelke and Carolyn Craven for their comments on a previous draft.

To have and to eat cake: The biscriptive role of game-theoretic explanations of human choice behavior

William D. Casebeer[a] and James E. Parco[b]

[a]*Department of Philosophy, United States Air Force Academy, Colorado Springs, CO 80840;* [b]*American Embassy, Tel Aviv, 63903 Israel.*

william.casebeer@usafa.af.mil **james.parco@usafa.af.mil**

http://www.usafa.af.mil/dfpfa/CVs/Casebeer.html

http://parco.usafa.biz

Abstract: Game-theoretic explanations of behavior need supplementation to be descriptive; behavior has multiple causes, only some governed by traditional rationality. An evolutionarily informed theory of action countenances overlapping causal domains: neurobiological, psychological, and rational. Colman's discussion is insufficient because he neither evaluates learning models nor qualifies under what conditions his propositions hold. Still, inability to incorporate emotions in axiomatic models highlights the need for a comprehensive theory of functional rationality.

The power and beauty of von Neumann and Morgenstern's *Theory of Games and Economic Behavior* (1944) and Luce and Raiffa's *Games and Decisions* (1957) lie in their mathematical coherence and axiomatic treatment of human behavior. Once rational agents could be described mathematically, game theory provided a far-reaching normative model of behavior requiring an assumption of common knowledge of rationality. This assumption (in addition to the often unstated requirement that a player fully understand the game situation) is subsumed under the phrase "the theory assumes rational players" (Luce & Raiffa 1957). But we know that, descriptively speaking, this is not always the case. The literature has clearly shown that not only are these (mathematically required) assumptions often too strong to be met in practice, but also that the "rational actor theory" (hereafter RAT) is underspecified in that it cannot effectively accommodate emotions. But does this constitute a failure of RAT? We think not.

Nevertheless, we agree with Colman's larger point that we need a "psychological game theory," or rather, a neurobiologically informed theory of decision-making. This is not because of the spectacular failure of game theoretic assumptions in any particular experiment, but rather stems from an ecumenical and fully naturalizable worldview about the causes of, and norms governing, human behavior. Choice-driven behavior is a function of multiple, highly distributed brain subsystems that include affect and emotion. For example, in the domain of moral judgment, good moral cognition is driven by a variety of brain structures, only some involved in ratiocination as traditionally construed (Casebeer & Churchland 2003). Even the most ardent RAT enthusiast recognizes that if your *explanandum* is all human behavior, your *explanans* will be more comprehensive than adverting to RAT alone.

Thus, we question the usefulness of Colman's ad hoc refinements for prescriptions of behavior in interactive decision-making, primarily because he has neither (1) qualified his theory as to when and under what conditions it applies, nor (2) provided an account for learning in games (beyond simple Stackelberg reasoning). For example, Colman uses the two-player centipede game as a primary domain in which he justifies his theory. However, recent evidence experimentally investigating three-player centipede games (Parco et al. 2002) directly contradicts it. Parco et al. extended the McKelvey and Palfrey (1992) study to three players using small incentives (10 cents for stopping the game at the first node, and $25.60 for continuing the game all the way to the end) and obtained similar results, soundly rejecting the normative equilibrium solution derived by backward induction. However, when the payoffs of the game were increased by a factor of 50 (and each player thus had the opportunity to earn $7,680), the results were markedly different. Although initial behavior of both the low-pay and high-pay conditions mirrored that of the McKelvey and Palfrey study, over the course of play for 60 trials, behavior in the high-pay treatment converged toward the Nash equilibrium and could be well accounted for using an adaptive reinforcement-based learning model. Furthermore, as noted by McKelvey and Palfrey (1992) and later by Fey et al. (1996), in all of the centipede experiments that were conducted up until then, there were learning effects in the direction of equilibrium play. Colman's oversight of the extant learning in games literature and his brief account for the dynamics of play through Stackelberg reasoning is insufficient. Learning in games manifests itself in a variety of processes quite different from simple Stackelberg reasoning (see Camerer & Ho 1999; Erev & Roth, 1998). For example, Rapoport et al. (2002) document almost "magical" convergence to the mixed-strategy equilibrium over 70 trials without common knowledge or between-trial feedback provided to subjects. Neither traditional game theory nor Colman's model can account for such data.

Generally speaking, Colman does little to improve prescriptions for human behavior both within and outside of the subset of games he has described; his paper is really a call for more theory than a theory proper. RAT's difficulty in dealing with emotions serves as proof-of-concept that we need a more comprehensive theory. Humans are evolved creatures with multiple causes of behavior, and the brain structures that subserve "rational" thought are, on an evolutionary timescale, relatively recent arrivals compared to the midbrain and limbic systems, which are the neural mechanisms of affect and emotion. Ultimately, our goal should be to formulate an explanation of human behavior that leverages RAT in the multiple domains where it is successful, but that also enlightens (in a principled way) as to when and why RAT fails. This more comprehensive explanation will be a neurobiological cum psychological cum rational theory of human behavior.

The problems game-theoretic treatments have in dealing with the role of emotions in decision-making serve to underscore our point. There are at least two strategies "friends of RAT" can pursue: (1) attempt to include emotions in the subjective utility function (meaning you must have a mathematically rigorous theory of the emotions; this is problematic), or (2) abandon RAT's claim to be discussing proximate human psychology and, instead, talk about how emotions fit in system-wide considerations about long-term strategic utility (Frank 1988). The latter approach has been most successful, although it leaves RAT in the position of being a distal explanatory mechanism. The proximate causes of behavior in this story will be locally arational or possibly irrational (hence the concerns with emotions). How would "new wave RAT" deal with this? One contender for a meta-theory of rationality that can accommodate the explanatory successes of RAT, yet can also cope with their failure in certain domains, is a functional conception of rationality. The norms that govern action are reasonable, and reason-giving for creatures that wish to be rational, insofar as such norms allow us to function appropriately given our evolutionary history and our current environment of action (Casebeer 2003).

We acknowledge that RAT will require supplementation if it is to fully realize its biscriptive explanatory role of predicting human action and providing us with a normative yardstick for it. Utility theory must incorporate neurobiological and psychological deter-

minants, as well as the rational, if game theory is to become as descriptively appealing as it is normatively.

Experience and decisions

Edmund Fantino and Stephanie Stolarz-Fantino

Department of Psychology, University of California–San Diego, La Jolla, CA 92093-0109. **efantino@ucsd.edu sfantino@psy.ucsd.edu**

Abstract: Game-theoretic rationality is not generally observed in human behavior. One important reason is that subjects do not perceive the tasks in the same way as the experimenters do. Moreover, the rich history of cooperation that participants bring into the laboratory affects the decisions they make.

Colman reviews many instances of game playing in which human players behave much more cooperatively and receive larger payoffs than permitted by conceptions of strict rationality. Specifically, he points out that although "Game-theoretic rationality requires rational players to defect in one-shot social dilemmas" (sect. 6.11), experimental evidence shows widespread cooperation. We agree that strict rationality does not accurately portray or predict human behavior in interactive decision-making situations. Particularly problematic are predictions made on the basis of backward induction. The Chain-store and Centipede games are good examples. In each case, backward induction makes it appear that the likely last move is inevitable, rather than one of a number of possible outcomes, as it must appear to the participant. In any case, it is unlikely that participants would reason backwards from the conclusion, even if such reasoning made sense. For example, Stolarz-Fantino et al. (2003) found that students were more likely to demonstrate the conjunction effect (in which the conjunction of two statements is judged more likely than at least one of the component statements) when the conjunction was judged before the components, than when it was judged after them. Further, if people easily reasoned backward from likely end-states, they should be more adept at demonstrating self-control (preferring a larger, delayed reward to a smaller, more immediate reward) than in fact they are (see discussion in Logue 1988).

Colman proposes "Psychological game theory" as a general approach that can be argued to account for these deviations. We agree that this is a promising approach, although it is a fairly broad and nonspecific approach as presented in the target article. We would add a component to Psychological game theory that appears to be relevant to the types of problems discussed: the pre-experimental behavioral history of the game participants. We are studying various types of irrational and nonoptimal behavior in the laboratory (e.g., Case et al. 1999; Fantino 1998a; 1998b; Fantino & Stolarz-Fantino 2002a; Goodie & Fantino 1995; 1996; 1999; Stolarz-Fantino et al. 1996; 2003) and are finding a pronounced effect of past history on decision-making (a conclusion also supported by Goltz' research on the sunk-cost effect, e.g., Goltz 1993; 1999). One example will suffice.

A case of illogical decision-making is base-rate neglect, first developed by Kahneman and Tversky (1973) and discussed often in this journal (e.g., Koehler 1996). Base-rate neglect refers to a robust phenomenon in which people ignore or undervalue background information in favor of case-specific information. Although many studies have reported such neglect, most have used a single "paper-and-pencil" question with no special care taken to insure attentive and motivated subjects. Goodie and Fantino wondered if base-rate neglect would occur in a behavioral task in which subjects were motivated and in which they were exposed to repeated trials. We employed a matching-to-sample procedure (MTS), which allowed us to mimic the base-rate problem quite precisely (Goodie & Fantino 1995; 1996; 1999; Stolarz-Fantino & Fantino 1990). The sample in the MTS task was either a blue or green light. After sample termination, two comparison stimuli appeared: these were always a blue and a green light. Subjects were instructed to choose either. We could present subjects with repeated trials rapidly (from 150 to 400 trials in less than a one-hour session, depending on the experiment) and could readily manipulate the probability of reinforcement for selecting either color after a blue sample and after a green sample. Consider the following condition (from Goodie & Fantino 1995): Following either a blue sample or a green sample, selection of the blue comparison stimulus is rewarded on 67% of trials, and selection of the green comparison stimulus is rewarded on 33% of trials; thus, in this situation the sample has no informative or predictive function. If participants responded optimally, they should have come to always select blue, regardless of the color of the sample; instead they focused on sample accuracy. Thus, after a green sample, instead of always choosing blue (for reward on 67% of trials) they chose the (matching) green comparison stimulus on 56% of trials (for a 48% rate of reward). This continued for several hundred trials. In contrast, Hartl and Fantino (1996) found that pigeons performed optimally, ignoring the sample stimulus when it served no predictive function. They did not neglect base-rate information.

What accounts for pigeons' and people's differing responses to this simple task? We have speculated that people have acquired strategies for dealing with matching problems that are misapplied in our MTS problem (e.g., Stolarz-Fantino & Fantino 1995). For example, from early childhood, we learn to match like shapes and colors at home, in school, and at play (e.g., in picture books and in playing with blocks and puzzles). Perhaps, this learned tendency to match accounts for base-rate neglect in our MTS procedure. If so, Goodie and Fantino (1996) reasoned that base-rate neglect would be eliminated by using sample and comparison stimuli unrelated to one another (line orientation and color). In this case, base-rate neglect was indeed eliminated. To further assess the learning hypothesis, Goodie and Fantino (1996) next introduced an MTS task in which the sample and comparison stimuli were physically different but related by an extensive history. The samples were the words "blue" and "green"; the comparison stimuli were the colors blue and green. A robust base-rate neglect was reinstated. Ongoing research in our laboratory is showing that pigeons with sufficient matching experience (where matching is required for reward) can be induced to commit base-rate neglect. These and other studies have led us to conclude that base-rate neglect results from preexisting learned associations.

How might learned associations account for nonoptimal decisions in the Prisoner's Dilemma Game (PDG)? Rationality theory argues that the selfish response is optimal. But we have been taught since childhood to be unselfish and cooperative. For many of us, these behaviors have been rewarded with praise throughout our lives (see the discussion of altruism in Fantino & Stolarz-Fantino 2002b; Rachlin 2002). Moreover, actual deeds of unselfish and cooperative behavior are often reciprocated. Why then should these behaviors not "intrude" on the decisions subjects make in the laboratory? Viewed from this perspective, there is nothing surprising about the kinds of behavior displayed in PDG. Indeed, such behavior is variable (many subjects cooperate, many defect), as one would expect from the variable behavioral histories of the participants.

A critique of team and Stackelberg reasoning

Herbert Gintis

Emeritus Professor of Economics, University of Massachusetts, Northampton, MA 01060; External Faculty, Santa Fe Institute, Santa Fe, NM. **hgintis@comcast.net http://www-unix.oit.umass.edu/~gintis**

Abstract: Colman's critique of classical game theory is correct, but it is well known. Colman's proposed mechanisms are not plausible. Insufficient reason does what "team reasoning" is supposed to handle, and it applies to a broader set of coordination games. There is little evidence ruling out more traditional alternatives to Stackelberg reasoning, and the latter is implausible when applied to coordination games in general.

Colman's critique of classical game theory is correct, but it is well known. He misses one critique that I consider to be among the most telling. If two "rational" players play a game with a unique, strictly mixed strategy equilibrium, neither player has an incentive to play using this equilibrium strategy, because in a true one-shot game, there is absolutely no reason to randomize. It is easy to explain why one would prefer that one's opponent not know which action we will take, and it is possible to work this up into a full-fledged justification of randomizing. But in a true one-shot, your opponent knows nothing about you, so even if you choose a pure strategy, you do no worse than by randomizing. The evolutionary game-theoretic justification is that in a large population of agents meeting randomly and playing the game in each period, in equilibrium a fraction of the population will play each of the pure strategies in proportion to that strategy's weight in the mixed-strategy Nash equilibrium.

Indeed, most of the problems with classical game theory can be handled by evolutionary/behavioral game theory, and do not need models of "nonstandard reasoning" (Gintis 2000). For instance, in a pure coordination game with a positive payoff-dominant equilibrium, and the payoffs to noncoordinated choices zero, evolutionary game theory shows that each pair of coordinated choices is a stable equilibrium, but if there are "trembles," then the system will spend most of its time in the neighborhood of the payoff-dominant equilibrium (Young 1993).

As Colman notes, many of the empirical results appearing to contradict classical game theory, in fact contradict the assumption that agents are self-regarding. In fact, agents in many experimental situations care about fairness, and have a propensity to cooperate when others cooperate, and to punish noncooperators at personal cost, even when there can be no long-run personal material payoff to so doing. For an analysis and review of the post-1995 studies supporting this assertion, see Gintis 2003.

Evolutionary game theory cannot repair all the problems of classical game theory, because evolutionary game theory only applies when a large population engages in a particular strategic setting for many periods, where agents are reassigned partners in each period. We still need a theory of isolated encounters among "rational" agents (i.e., agents who maximize an objective function subject to constraints). Colman proposes two such mechanisms: team reasoning and Stackelberg reasoning. I am not convinced that either is a useful addition to the game-theoretic repertoire.

Concerning "team reasoning," there is certainly much evidence that pregame communication, face-to-face interaction, and framing effects that increase social solidarity among players do increase prosocial behavior and raise average group payoffs, but this is usually attributed to players' placing positive weight on the return to others, and increasing their confidence that others will also play prosocially. But these are nonstandard *preference* effects, not nonstandard *reasoning* effects. Choosing the payoff-maximum strategy in pure coordination games, where players receive some constant nonpositive payoff when coordination fails, is most parsimoniously explained as follows. If I know nothing about the other players, then all of my strategies have an equal chance of winning, so personal payoff maximization suggests choosing the payoff maximum strategy. Nothing so exotic as "team reasoning" is needed to obtain this result. Note that if a player *does* have information concerning how the other players might choose, an alternative to the payoff-maximum strategy may be a best response.

Moreover, "team reasoning" completely fails if the pure coordination game has nonconstant payoffs when coordination is not achieved. Consider, for instance, the following two-person game. Each person chooses a whole number between 1 and 10. If the numbers agree, they each win that amount of dollars. If the numbers do not agree, they each lose the larger of the two choices. For example, if one player chooses 10, and the other chooses 8, they both lose ten dollars. This is a pure coordination game, and "team reasoning" would lead to both players choosing 10. However, all pure strategies are evolutionary equilibria, and computer simulation shows that the higher numbers are less likely to emerge when the simulation is randomly seeded at the start (I'll send interested readers the simulation program). Moreover, if an agent knows nothing about his partner, it is easy to show, using the Principle of Insufficient Reason, that 2 and 3 have the (equal and) highest payoffs. So if an agent believes that partners use the same reasoning, he will be indifferent between 2 and 3. By the same reasoning, if one's partner chooses 2 and 3 with equal probability, then the payoff to 3 is higher than the payoff to 2. So 2 is the "rational" choice of "ignorant" but "rational" agents.

Colman argues that there is strong evidence supporting Stackelberg reasoning, but he does not present this evidence. Some is unpublished, but I did look at the main published article to which he refers (Colman & Stirk 1998). This article shows that in 2×2 games, experimental subjects overwhelmingly choose Stackelberg solutions when they exist. However, a glance at Figure 1 (p. 284) of this article shows that, of the nine games with Stackelberg solutions, six are also dominance-solvable, and in the other three, any reasoning that would lead to choosing the payoff-maximum strategy (including the argument from insufficient reason that I presented above), gives the same result as Stackelberg reasoning. So this evidence does not even weakly support the existence of Stackelberg reasoning. I encourage Colman to do more serious testing of this hypothesis.

I find the Stackelberg reasoning hypothesis implausible, because if players used this reasoning in pure coordination games, it is not clear why they would not do so in other coordination games, such as Battle of the Sexes (in this game, both agents prefer to use the same strategy, but one player does better when both use strategy 1, and the other does better when both use strategy 2). Stackelberg reasoning in this game would lead the players never to coordinate, but always to choose their preferred strategies. I know of no experimental results using such games, but I doubt that this outcome would be even approximated.

How to play if you must

Hans Haller

Department of Economics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. **haller@vt.edu**

Abstract: Beyond what Colman is suggesting, some residual indeterminacy of Nash equilibrium may remain even after individual rationality is amended. Although alternative solution concepts can expand the positive scope (explanatory power) of game theory, they tend to reduce its accuracy of predictions (predictive power). Moreover, the appeal of alternative solutions may be context-specific, as illustrated by the Stackelberg solution.

Analysis of a strategic or noncooperative game presumes that the players are committed to participate. Normative analysis then aims at an unambiguous recommendation of how to play the game. If the analyst and the players adhere to the same principles of rationality, then the players will follow the recommendation; indeed, the players can figure out how to play without outside help. But can they? Like Colman, I shall refrain from elaborating on bounded rationality.

Andrew Colman presents the argument of Gilbert, that common knowledge of individual rationality does not justify the use of salient (exogenous, extrinsic) focal points to resolve indeterminacy. Nor does it justify endogenous or intrinsic focal points based on payoff dominance or asymmetry. This argument is in line with the critique by Goyal and Janssen (1996) of Crawford and Haller's heuristic principle, to stay coordinated once coordination is obtained. It applies as well to folk theorem scenarios, as in the infinitely repeated Prisoner's Dilemma Game (PDG): None of the multiple equilibria is distinguished on the grounds of individual rationality alone. The argument shows that principles other than individual rationality have to be invoked for equilibrium selection

à la Harsanyi and Selten (1988), and for rationalizing focal points *à la* Kramarz (1996) or Janssen (2001b). Yet, even the addition of several compelling principles need not result in a unique solution for every game. For instance, in the Battle of the Sexes game, the equilibrium in mixed strategies is ruled out by payoff dominance, and there is no obvious way to select between the two equilibria in pure strategies. It seems that there always remains some residual indeterminacy – unless it is stipulated by law how to play certain games. Thus, the ambitious goal of orthodox game theory, broadly defined, to identify a unique solution for each game, has been almost, but not completely, reached.

But do players play as they should? As the author of the target article observes, it takes a further bridging hypothesis of weak rationality – that people try to act rationally – to turn the normative theory into a positive one. Then, as a rule, the recommendations of normative theory are treated as predictions. On a more fundamental level, the common knowledge and rationality (CKR) assumptions may be tested. Although I agree that the literature on experimental gaming testifies to the fruitfulness of empirical research, I would add that empirical research in industrial organization tends to rely on natural rather than laboratory experiments. This is worth noting, because economics, and in particular industrial economics, has been the main area of applied game theory and has immensely contributed to the development and proliferation of game-theoretical modeling.

Obviously, one would not necessarily observe the predicted outcome, if the participants played a game that was different from the one specified by the analyst or experimentalist. This would be the case if the monetary payoffs, or hypothetical payoffs according to the instructions, did not represent the subjects' preferences. Such instances are altruism or fairness considerations not accounted for in the original payoff functions. In such a case, the "neoclassical repair kit" can be applied, to use a popular, albeit somewhat derogatory, term: After a payoff transformation or, more generally, substitution of suitable utility functions for the original payoff functions, the data no longer reject the model. Thus, although the original model proved numerically mis-specified, the theory at large has not been rejected.

Yet, there are plenty of instances where the specified payoffs do represent player preferences, and orthodox and not-so-orthodox game theory is rejected in laboratory experiments. The first response to discrepancies between theory and evidence would be to perform further experiments, to corroborate or reevaluate the earlier evidence. After all, the immediate response to reports of cold fusion was additional experimentation, not a rush to revise theory. It appears that deliberate attempts at duplication are rare and poorly rewarded in experimental gaming. Still, certain systematic violations of individual rationality are abundant, like playing one's strictly dominated strategy in a one-shot PDG and the breakdown of backward induction in a variety of games.

In response to concerns rooted both in theory and evidence, game theory has become fairly heterodox. The recent developments suggest an inherent tension between the goals of explaining additional phenomena and of making more specific predictions (Haller 2000). Less stringent requirements on solutions can help explain hitherto unexplained phenomena. In the opposite direction, the traditional, or if you want, orthodox literature on equilibrium refinements and equilibrium selection has expended considerable effort to narrow the set of eligible equilibrium outcomes, to make more accurate predictions. Apart from the tradeoff mentioned, achieving a gain of explanatory power at the expense of predictive power, novel solution concepts may be compelling in some contexts and unconvincing under different but similar circumstances. One reason is that many experiments reveal a heterogeneous player population, with a substantial fraction evidently violating individual rationality, and another non-negligible fraction more or less conforming to orthodoxy. This raises interesting questions; for example, whether the type of a player is time-invariant or not.

Among the host of tentative and ad hoc suggestions falling under the rubric of psychological game theory, Stackelberg reasoning can explain specific payoff dominance puzzles, but yields detrimental outcomes when applied to other classes of Stackelberg solvable games. For instance, in a Cournot duopoly with zero costs and linear demand, the Stackelberg solution yields the perfectly competitive outcome, which is payoff-dominated by the Cournot-Nash outcome. Hence, the Stackelberg solution illustrates that the appeal of alternative solutions may be context-specific. Incidentally, a Stackelberg solution is a special case of a conjectural variation equilibrium. The latter concept can be traced back to Bowley (1924). It introduces a quasidynamic element into a static game. It has been utilized in models of imperfect competition and strategic trade from time to time, and has seen a revival recently. Despite its appeal, this modeling approach has been frequently dismissed on the grounds that it makes ad hoc assumptions and constitutes an unsatisfactory substitute for explicit dynamics.

Colman's article is thought-provoking and touches on several of the most pressing challenges for game theory, without pretending to be comprehensive or definitive. It will be fascinating to see which new theoretical concepts will emerge to address these challenges, and which ones will last.

What's a face worth: Noneconomic factors in game playing

Peter J. B. Hancock[a] and Lisa M. DeBruine[b]

[a]*Department of Psychology, University of Stirling, Stirling, FK9 4LA United Kingdom;* [b]*Department of Psychology, McMaster University, Hamilton, Ontario L8S 4K1, Canada.* **pjbh1@stir.ac.uk** **debruilm@mcmaster.ca**
http://www.stir.ac.uk/psychology/staff/pjbh1
http://homepage.mac.com/debruine/

Abstract: Where behavior defies economic analysis, one explanation is that individuals consider more than the immediate payoff. We present evidence that noneconomic factors influence behavior. Attractiveness influences offers in the Ultimatum and Dictator Games. Facial resemblance, a cue of relatedness, increases trusting in a two-node trust game. Only by considering the range of possible influences will game-playing behavior be explained.

Whenever a game is played between two people, there are many potential motives for particular forms of behavior. One player may wish to impress or defer to the other. One may feel vindictive towards or sorry for the other player. Such motivations and others, in various combinations, can add many layers of complexity to a game-theoretic analysis of the payoffs. Where people behave in an apparently irrational manner, it is possible that their perception of the payoff does not equate to the economic one because of these other factors. Players may also use cues to predict the behavior of playing partners. For example, images of smiling partners are trusted more than those who are not smiling (Scharlemann et al. 2001).

The Ultimatum Game is one where behavior defies a simple payoff analysis (e.g., Thaler 1988). One player (the proposer) can allocate some proportion of a sum of money to the second player (the responder), who may accept or refuse the offer. If the offer is refused, the money is returned and neither player gets anything. Usually the game is played single-shot, where the players do not know or even see each other. A payoff analysis suggests that any offer should be accepted, but in typical western societies anything less than about 35% is refused. This is usually explained as enforcement of "fair play" by the responder. In the related Dictator Game, the second player has no choice. Now, the first player is free to offer nothing, but in practice, usually does make some offer. It appears that something inhibits purely selfish behavior. The situation is more complicated when the players know something of each other, as the other kinds of factors mentioned above may affect decisions.

Attractiveness is one of these factors. Apart from being desirable in its own right, the halo effect causes many assessments of another, such as their intelligence and character, to be estimated more highly. Thus, Solnick and Schweitzer (1999) found that more was expected of attractive faces. Joergensen and Hancock (2001) reported an Ultimatum Game where proposers saw a picture of the responder. Offers were higher to faces rated as attractive, echoing results from Solnick and Schweitzer (1999), but with a stronger effect for attractive women. The correlation between rated attractiveness and offer level was 0.83. However, the effect of attractiveness was transient, it disappeared in a second round of the game following information about who had refused low offers. Hancock and Ross (2002) investigated the Dictator Game with similar results: The correlation between offer levels within the game and independently rated attractiveness was 0.91.

These experiments use anonymous faces, but what effect might the perception that someone is related to you have? Hamilton's theory of kin selection (Hamilton 1964) suggests that people should be favorably disposed toward relatives. Any gene promoting altruistic behavior can influence its own success by benefiting those most likely to share a copy of itself. Thus, a gene causing altruism will be favored if the benefit (b) to the recipient multiplied by the relatedness[1] (r) between the altruist and recipient is greater than the cost (c) to the altruist.

Given this logic, cues of relatedness between individuals may change the payoffs attributed to different behaviors. DeBruine (2002) explored behavior in a two-person, two-node sequential trust game (after Eckel & Wilson 1998b, and related to the Centipede game described by Colman). The first player can decide either not to trust the second, in which case both get a sure payoff of $3, or to trust the second player with the decision. The second player can decide between selfish behavior, keeping $5 and giving $2 to player one, or unselfish behavior, allocating $4 to each. Given no information about the second player, the first player's expected payoff is $3 for either choice, so a rational player should be indifferent. However, if the other player is a relative with relatedness of 0.5,[2] then the structure of the game is changed to a choice between a sure payoff of $4.50 ($3 to self plus 0.5 times $3 to the other) and a risky payoff with an expected value of $5.25 ($3 to self plus 0.5 times $4.50 to the other). In addition, assessment of the second player's trustworthiness may bias the expected payoff of trusting.

DeBruine (2002) digitally morphed images of the other player to manipulate one possible cue of relatedness, facial resemblance. Players in this trust game chose the riskier option more often when the image of the opponent had been morphed to resemble the first player than when the image had been morphed to resemble an unknown person. This is consistent with an increase in either the expected payoff of trusting or the expected probability of trustworthy behavior. Analysis of the responses indicated that independently rated attractiveness of the second player did not influence behavior in this situation, although current research by DeBruine indicates that resemblance to self increases the attractiveness of faces (also see Penton-Voak et al. 1999).

In an unpublished study, DeBruine randomized the computer-generated second players' responses in the trust game. Players were less likely to trust opponents in games immediately after they had been cheated than in games after the opponent was unselfish. This echoes findings by Eckel and Wilson (1998a) that the previous opponent's response influences the current choice, even when the current opponent is a different person. Within the sets of games played after either a selfish or unselfish response, players were still more likely to trust faces morphed to resemble themselves.

In any social situation, people evaluate others. We have shown that even a static photograph of another player can cause significant differences to behavior in simple games. A playing partner's attractiveness may introduce noneconomic motivations to the game or change the player's predictions of the partner's behavior. The perception that someone may be related to you introduces further complications, because of the shift of possible inclusive fitness payoffs. The inclusion of such factors will broaden the scope of a psychological game theory.

ACKNOWLEDGMENTS

We wish to thank M. Daly, P. Joergensen, K. Ross, I. Penton-Voak, C. Taylor, and M. Wilson.

NOTES

1. Relatedness refers not to the total proportion of genes shared, but to those shared by identical descent. In this case, r is the probability that the recipient shares a gene for altruism with the altruist.

2. The conclusion holds for any $r > 0$ with this particular game payoff structure.

Rational belief and social interaction

Daniel M. Hausman

Department of Philosophy, University of Wisconsin-Madison, Madison, WI 53706-1474. **dhausman@wisc.edu**
http://philosophy.wisc.edu/hausman

Abstract: Game theory poses problems for modeling rational belief, but it does not need a new theory of rationality. Experimental results that suggest otherwise often reveal difficulties in testing game theory, rather than mistakes or paradoxes. Even though the puzzles Colman discusses show no inadequacy in the standard theory of rationality, they show that improved models of belief are needed.

The theory of rational choice takes choice to be rational when it tracks preferences. Preferences are rational when they are, in a precise sense, consistent. When there is risk or uncertainty, preferences and hence choices depend on beliefs; and neither preference nor choice is rational unless belief is rational. Rational beliefs must conform to the calculus of probabilities. When they do, and preferences satisfy relevant consistency and technical axioms, then preferences can be represented by expected utilities, and choice is rational if and only if it maximizes expected utility.

Expected utility maximization is defined whenever rational beliefs and preferences are defined. The fact that an interaction is strategic by itself causes no problem. If I am playing the pure coordination game in Colman's Figure 1 and believe that the other player will play Tails, then I should choose Tails, too.

But suppose I have no beliefs about what strategies other players will choose other than those that I can deduce from (1) beliefs about the other players' preferences over the payoffs, (2) beliefs about the other players' beliefs concerning both the game and its players, and (3) beliefs about the other players' rationality. All of these beliefs of mine have to be rational – that is, they have to be consistent with the calculus of probabilities – but this constraint permits many different sets of beliefs to count as rational. One way of developing game theory that greatly narrows the set of rational beliefs has been to assume that the players are all perfectly rational, that they all share the same subjective prior probabilities, that they have complete knowledge of the extensive form of the game, and that all of this is common knowledge. Call this "the total rationality representation" (TRR). TRR is not required by the standard theory of rational belief, and the fact that it leads to surprising and sometimes arguably paradoxical results is no indictment of the standard theory.

Colman also objects that game theory employing TRR may be uninformative. In the Hi-Lo Matching game of Figure 2, the theory fails to predict and recommend that players choose strategy H. As Colman correctly points out, if TRR requires common prior point probabilities, then no argument can be given for the rationality of playing H. But the remedy here is just a mild relaxation of the idealizations. If one does not require that the players have point priors, then Player I can believe that the probability that Player II will play *H* is not less than one-half, and also believe that

	II		
I		C	D
	C	\$3,\$3	\$1,\$4
	D	\$4,\$1	\$2,\$2

Figure 1 (Hausman). A prisoner's dilemma game form

Player II believes the same of Player I. Player I can then reason that Player II will definitely play *H*, update his or her subject probability accordingly, and play *H*. The problem lies with one idealized development of the standard view of rational belief, not with the view itself.

Many of the purported paradoxes Colman discusses yield to similar, though more complicated treatment. But some of the purported paradoxes are not paradoxes at all, and some of the apparent experimental disconfirmations are dubious. Consider first the standard single-shot prisoner's dilemma (PDG). Mutual defection is the uniquely rational outcome. Colman takes this to be paradoxical and to show that rationality is self-defeating, on the grounds that mutual cooperation is better for both players. In addition, he cites evidence showing that many experimental subjects, in fact, cooperate.

Although rationality is indeed *collectively* self-defeating in a PDG, there is no paradox or problem with the theory of rationality, and the apparently disconfirming data Colman cites are questionable. Consider the following game form (Fig. 1), which represents a PDG if the two players care only about their own monetary payoffs.

Mutual cooperators do better than mutual defectors. But the benefit comes from the choice the other player makes, not from one's own choice. (Remember this is a simultaneous play one-shot game in which I and II choose independently.) Unlike the finite iterated prisoner's dilemma or the centipede game, mutual cooperators cannot taunt mutual defectors, "If you're so rational, how come you ain't rich?" because the defectors can reply, "Because I wasn't lucky enough to be playing against a fool."

In addition, the apparently disconfirming experimental evidence is dubious, because cooperating subjects facing a game form like the one in Figure 1 might not be playing a PDG. To know what game they are playing one needs to know their preferences. For example, unless II prefers the outcome where II gets \$4 and I gets \$1 to the actual outcome of \$3 each, II was not playing a PDG. For those who do not have these preferences, the interaction depicted in Figure 1 is not a prisoner's dilemma. Similar remarks apply to the tetrapod in Colman's Figure 5. If the numbers represented dollars, many people would prefer the outcome where both get \$18 and player I's trust is rewarded, to the outcome where II gets \$19 and I gets \$8. The numbers in Figure 5 are, of course, supposed to represent utilities rather than dollars, but the common view, that the recommendation to play down on the first move is absurd, may reflect a common refusal to believe that these numbers correctly represent the preferences.

A great deal remains to be done to figure out how to represent rational beliefs. Wonderful controversy still rages. But one should not thereby conclude, as Colman does, that "the conception of rationality on which it [game theory] rests appears to be internally deficient" (target article, sect. 9.2). His essay does not address the treatment of rational preference, and the problems Colman explores concerning rational belief show, at most, the limitations of specific modeling choices, rather than a deficiency in basic concepts.[1]

NOTE

1. I do not, in fact, think that the standard theory of rationality is unproblematic (see e.g., my 1992 book, Chs. 2, 12, 13), but the difficulties I see are independent of those that Colman alleges.

The limits of individualism are not the limits of rationality

Susan Hurley

PAIS, University of Warwick, Coventry CV4 7AL, United Kingdom.
susan.hurley@warwick.ac.uk **www.warwick.ac.uk/staff/S.L.Hurley**

Abstract: Individualism fixes the unit of rational agency at the individual, creating problems exemplified in Hi-Lo and Prisoner's Dilemma (PD) games. But instrumental evaluation of consequences does not require a fixed individual unit. Units of agency can overlap, and the question of which unit should operate arises. Assuming a fixed individual unit is hard to justify: It is natural, and can be rational, to act as part of a group rather than as an individual. More attention should be paid to how units of agency are formed and selected: Are the local processes local or nonlocal? Do they presuppose the ability to understand other minds?

I disagree with little that Colman says about the limitations of orthodox rational choice theory, but wonder why he doesn't say more to challenge individualism as their source, and why he omits references to trailblazers such as Regan (1980) and Howard (1988).

In 1989, I argued that Hi-Lo and Prisoner's Dilemma games (PDs) exemplify the limits of individual rationality. In Hi-Lo, individuals have the same goals, yet individual rationality fails to guarantee them the best available outcome. In PDs, individuals have different goals, and individual rationality guarantees an outcome worse for all than another available outcome. These problems stem not from nature of individuals' goals, or the instrumental character of rationality, but from individualism about rationality, which holds the unit of rational agency exogenously fixed at the individual (cf. Hurley 1989).

Activity by a given unit of agency has consequences, calculated against a background of what occurs outside that unit, and can be evaluated instrumentally. Such consequentialist evaluation does not require the unit whose activity is evaluated to be fixed at the individual. Larger units of agency can subsume smaller ones, and consequentialist evaluation can apply to different units, with different results. We can think of individuals as composed of persons-at-times (or in other ways, involving multiple personalities); similarly, we can think of collective agents as composed of persons. In both cases, lower-level rationality (or irrationality) may coexist with, or even explain, higher-level irrationality (or rationality). For example, we understand from social dilemmas and social choice theory how a group can behave irrationally as a unit, although the agents composing it are individually rational. Intrapersonal analogues of social dilemmas may explain some forms of individual irrationality. Conversely, agents can behave irrationally as individuals, yet their actions fit together so that the group they compose behaves rationally (Hutchins 1995, pp. 235ff).

Individualism requires the individual to do the individual act available that will have the best expected consequences, given what other individuals are expected to do. Given others' expected acts, an individual agent has certain possible outcomes within her causal power. The best of these may not be very good, and it may be indeterminate what others are expected to do. But a group of individuals acting as a collective agent can have different possible outcomes within its causal power, given what agents outside the group are expected to do. A collective agent may be able to bring about an outcome better than any that the individual agent can bring about – better for that individual, inter alia. If so, the issue is not just what a particular unit of agency should do, given others'

expected acts, but also *which* unit should operate. The theory of rationality has yet to endogenize the latter question; Bacharach calls this "an important lacuna" (1999, p. 144; but cf. Regan 1980).

The assumption of a fixed individual unit, once explicitly scrutinized, is hard to justify. There is no theoretical need to identify the unit of agency with the source of evaluations of outcomes; collective agency does not require collective preferences. Although formulations of team reasoning may assume team preferences (see target article, sect. 8.1), what is distinctive about collective agency comes into sharper relief when it is made clear that the source of evaluations need not match the unit of agency. As an individual, I can recognize that a wholly distinct agent can produce results I prefer to any I could bring about, and that my own acts would interfere. Similarly, as an individual I can recognize that a collective agent, of which I am merely a part, can produce results I prefer to any I could bring about by acting as an individual, and that my doing the latter would interfere. Acting instead in a way that partly constitutes the valuable collective action can be rational. Not only can it best serve my goals to tie myself to the mast of an extended agent, but rationality itself can directly so bind me – rather than just prompt me to use rope.

Acting as part of a group, rather than as an individual, can also be natural. Nature does not dictate the individual unit of agency. Persons can and often do participate in different units, and so face the question of which unit they *should* participate in. Moreover, the possibility of collective agency has explanatory power. For example, it explains why some cases (e.g., Newcomb's Problem and Quattrone & Tversky's voting result) of supposedly evidential reasoning have intuitive appeal, while others (e.g., the smoking gene case) have none (Hurley 1989, Ch. 4; 1991; 1994).[1]

If units of agency are not exogenously fixed, how are units formed and selected? Is centralized information or control required, or can units emerge as needed from local interactions? At what points are unit formation and selection rationally assessable? I cannot here offer a general view of these matters, but highlight two important issues.

First, are the relevant processes local or nonlocal? Regan's version of collective action requires cooperators to identify the class of those intending to cooperate with whomever else is cooperating, to determine what collective action by that group would have the best consequences (given noncooperators' expected acts), and then play their part in that collective action. This procedure is nonlocal, in that cooperators must type-check the whole class of potential cooperators and identify the class of cooperators before determining which act by that group would have the best consequences. This extensive procedure could be prohibitive without central coordination. The problem diminishes if cooperators' identities are preestablished for certain purposes, say, by their facing a common problem, so preformed groups are ready for action (see Bacharach 1999).

A different approach would be to seek local procedures from which potent collective units emerge. Flexible self-organization can result from local applications of simple rules, without central coordination. Slime mold, for example, spends most of its life as separate single-celled units, but under the right conditions these cells coalesce into a single larger organism; slime mold opportunistically oscillates between one unit and many units. No headquarters or global view coordinates this process; rather, each cell follows simple local rules about the release and tracking of pheromone trails.

Howard's (1988) Mirror Strategy for one-off PDs may allow groups of cooperators to emerge by following a simple self-referential local rule: Cooperate with any others you encounter who act on this very same rule. If every agent cooperates just with its copies, there may be no need to identify the whole group; it may emerge from decentralized encounters governed by simple rules. Evidently, rules of cooperation that permit groups to self-organize locally have significant pragmatic advantages.

Both Regan's and Howard's cooperators need to perceive the way one another thinks, their methods of choice. Which choices their cooperators make, depends on which other agents are cooperators, so cooperation must be conditioned on the *methods* of choice, not the choices, of others. If method-use isn't perfectly reliable, however, cooperators may need to be circumspect in assessing others' methods and allow for the possibility of lapses (Bacharach 1999).

These observations lead to the second issue I want to highlight: What is the relationship between the processes by which collective agents are formed and selected, and the ability to understand other minds? Does being able to identify with others as part of a unit of agency, require being able to identify with others mentally? Psychologists ask: What's the functional difference between genuine mind-reading and smart behavior-reading (Whiten 1996)? Many social problems that animals face can be solved merely in terms of behavior-circumstance correlations and corresponding behavioral predictions, without postulating mediating mental states (see Call & Tomasello 1999; Heyes & Dickinson 1993; Hurley 2003; Povinelli 1996). What kinds of problems also require understanding the mental states of others?

Consider the kinds of problems that demonstrate the limitations of individualistic game theory. When rational individuals face one another, mutual behavior prediction can break down in the ways that Colman surveys; problem-solving arguably requires being able to understand and identify with others mentally. If cooperators need to know whether others have the mental processes of a cooperator before they can determine what cooperators will do, they must rely on more than unmediated associations between circumstances and behavior. Collective action would require mind-reading, not just smart behavior-reading. Participants would have to be mind-readers, and be able to identify, more or less reliably, other mind-readers.

NOTE

1. It is widely recognized that Prisoners' Dilemma can be interpreted evidentially, but less widely recognized that Newcomb's Problem and some (but not all) other cases of supposed evidential reasoning can be interpreted in terms of collective action.

Coordination and cooperation

Maarten C. W. Janssen

Department of Economics, Erasmus University, 3000 DR, Rotterdam, The Netherlands. **janssen@few.eur.nl** **www.eur.nl/few/people/janssen**

Abstract: This comment makes four related points. First, explaining coordination is different from explaining cooperation. Second, solving the coordination problem is more important for the *theory* of games than solving the cooperation problem. Third, a version of the Principle of Coordination can be rationalized on individualistic grounds. Finally, psychological game theory should consider how players perceive their gaming situation.

Individuals are, generally, able to get higher payoffs than mainstream game-theoretic predictions would allow them to get. In coordination games, individuals are able to coordinate their actions (see e.g., Mehta et al. 1994a; 1994b; Schelling 1960) even though there are two or more strict Nash equilibria. In Prisoner's Dilemma games, individuals cooperate quite often, even though mainstream game theory tells that players should defect. In this comment, I want to make four points. First, it is important to distinguish the cooperation problem from the coordination problem. Second, from the point of view of developing a *theory* of games, the failure to explain coordination is more serious than the failure to explain cooperation. Third, the Principle of Coordination, used to explain why players coordinate, can be rationalized on individualistic grounds. One does not need to adhere to "we thinking" or "Stackelberg reasoning." Finally, psychological game theory may gain predictive power if it takes into account how players perceive their gaming situation.

The problem of *coordination* is different from the problem of *cooperation.* In a cooperation problem, as the *one-shot* Prisoner's Dilemma, players have a dominant strategy, which is *not* to cooperate, and one may wonder why people deviate from their dominant strategy and *do* cooperate. To explain cooperation, one has to depart from the axiom of individual rationality. This is not the case for the problem of coordination. In a coordination problem, there are two or more Nash equilibria in pure strategies and the issue is that individual rationality considerations are *not sufficient* to predict players' behavior. To explain coordination, an approach that supplements the traditional axioms of individual rationality may be taken.

In a truly *one-shot* Prisoner's Dilemma, where the payoffs are formulated such that players care only about their individual payoffs, I find it hard to find reasons (read: to explain) why people cooperate. Of course, I don't want to deny the empirical evidence, but the dominant strategy argument seems to me very appealing and difficult to counteract. If people choose to cooperate, they must be in one way or the other boundedly rational. I think the *theory* of games should not just explain how people in reality behave when they play games. It should also have an answer to the question *why,* given their own preferences, they behave in a certain way. The weakest form this requirement can take is that, given a theoretical prediction people understand, it is in their own interest not to deviate from the prediction. In other words, a theory of games should be reflexive. The problem with a theory of games which says that players cooperate is that "smart students" don't see any reason why it is beneficial to do so. If the theory makes a prediction, then it should be in the interest of the players to make that prediction come true.

In coordination problems, the concept of Nash equilibrium is too weak, as it does not give players a reason to choose one out of several alternatives. Gauthier (1975), Bacharach (1993), Sugden (1995), and Janssen (2001b) make use of (a version of) the Principle of Coordination to explain coordination. Janssen (2001a) develops a relatively simple framework that rationalizes the uniqueness version of this Principle. The basic idea is that each player individually forms a plan, specifying for each player how to play the game, and which conjecture to hold about their opponent's play. Individual plans should satisfy two axioms. *Individual rationality* says that a plan be such that the sets of strategies that are motivated by the plan must be best responses to the conjectures that are held about the other player's play. *Optimality* requires that players formulate optimal plans, where a plan is optimal if the maximum payoff both players get if they follow this plan is larger than the minimum payoff both players would get according to any alternative plan satisfying the individual rationality axiom.

If there is a unique strict Pareto-efficient outcome, then there is a unique plan satisfying Individual Rationality and Optimality how to play the game. To see the argument, consider the following game (Table 1).

It is clear that a plan where every player conjectures the other to play *L,* and where both players actually choose *L,* is a plan that satisfies Individual Rationality and, moreover, is better for both players than any other plan. As the plan is uniquely optimal, both players *thinking individually* formulate the same plan, and they will choose to do their part of it.

Note that the above approach is different from "we thinking" as discussed by Colman, as no common preferences are specified.

Table 1 (Janssen). *A game of pure coordination with a uniquely efficient equilibrium*

	L	*R*
L	*2,2*	*0,0*
R	*0,0*	*1,1*

Table 2 (Janssen). *A game of pure coordination without a uniquely efficient equilibrium*

	Blue	*Blue*	*Red*
Blue	*1,1*	*0,0*	*0,0*
Blue	*0,0*	*1,1*	*0,0*
Red	*0,0*	*0,0*	*1,1*

Also, no coach is introduced who can make recommendations to the players about how to coordinate their play, as in Sugden (2000, p. 183).

This approach, by itself, cannot explain coordination in a game where two players have to choose one out of three (for example, two blue and one red) objects and where they get awarded a dollar if they happen to choose the same object. Traditionally, game theory would represent this game in the following "descriptively objective" matrix (Table 2).

Intuitively, the players should pick the red object, but the Principle of Coordination advocated here, by itself, cannot explain this intuition.

Psychological game theory may, in addition to the elements mentioned by Colman, also further investigate Bacharach's (1993) suggestion, and investigate how people describe the game situation to themselves (instead of relying on some "objective" game description). By using the labels of the strategies, individuals may describe the above game as being a game between picking a blue and a red object, where the chance of picking the *same* blue object, given that both pick a blue object, is equal to a half. Given such a description, there is (again) a unique plan satisfying Individual Rationality and Optimality.

Which is to blame: Instrumental rationality, or common knowledge?

Matt Jones and Jun Zhang

Department of Psychology, University of Michigan, Ann Arbor, MI 48109-1109. **mattj@umich.edu** **junz@umich.edu**
http://umich.edu/~mattj

Abstract: Normative analysis in game-theoretic situations requires assumptions regarding players' expectations about their opponents. Although the assumptions entailed by the principle of common knowledge are often violated, available empirical evidence – including focal point selection and violations of backward induction – may still be explained by instrumentally rational agents operating under certain mental models of their opponents.

The most important challenge in any normative approach to human behavior is to correctly characterize the task the person is presented with. As Colman points out, the normative analysis of game settings provided by instrumental rationality is incomplete; information must be included about the opponent. We argue here that the common knowledge of rationality (CKR) axioms, which are meant to extend normative analysis to game theory, actually limit the rationality attributed to subjects. When players are allowed to reason about their opponents, using more information than just that provided by CKR2, we find that the major phenomena cited as evidence against rational choice theory (RCT) – focal point selection and violations of backward induction arguments – can be predicted by the resulting normative theory. This line of reasoning follows previous research in which supposed suboptimalities in human cognition have been shown to be adaptive given a more fully correct normative analysis (e.g., Anderson &

Schooler 1991; Anderson et al. 1997; Flood 1954; Jones & Sieck, in press; Oaksford & Chater 1996; Schacter 1999).

The difficulty with the CKR axioms is that they require players to reason about their opponents entirely a priori, based only on the assumptions of rationality and common knowledge, while ignoring all other potential sources of information. A more faithful model of rational choice would allow players to utilize all the knowledge available to them, including general knowledge about human behavior or specific knowledge about the opponent gained from previous interactions (e.g., earlier moves). For example, the fact that the conventional priority of Heads over Tails leads to the phenomenon of focal point selection should realistically be available to each player as information for use in predicting the opponent's choice. Thus, all that is needed is a simple intuitive understanding of human behavior for a subject to infer correctly (and rationally) that the opponent is likely to choose the focal option. Instrumental rationality then dictates that the player chooses that option as well. Similar reasoning applies to payoff dominance in the case of the Hi-Lo matching game.

Relaxing the restrictions provided by CKR on players' models of their opponents can also explain violations of the prescriptions of backward induction arguments. If Player II's model of Player I admits alternatives to perfect rationality, then an initial cooperative move by Player I will simply lead to an update of II's beliefs about I (rather than generating a logical impasse). This sort of updating can be formalized using a Bayesian framework, in which each player has probabilistic prior beliefs about the opponent (perhaps peaked around rationality, but nonzero elsewhere), which are determined by prior experience with the opponent or with people in general. Even if the prior expectation were heavily biased towards strict rationality, an initial cooperative move by Player I would force Player II's model to favor other possibilities, for example, that Player I always plays Tit-For-Tat. This could lead to Player II cooperating on step 2, in turn giving Player I justification for cooperating on step 1.

The preceding arguments have shown how failures of CKR can be remedied by more complete normative analyses that preserve the assumption of instrumental rationality, that is, optimality of actions as conditioned on the model of the opponent. The question of rationality in game scenarios then shifts to the rationality of that model itself (inductive rationality). In the case of focal point selection, we have offered no specific mechanism for the inductive inference regarding the opponent's likely choice, as based on general experience with human behavior. We merely point out that it is perfectly consistent with the assumption of inductive rationality (although it has no basis in CKR). (Ironically, the same empirical fact that is cited as evidence against RCT – namely, focal point selection – actually corroborates the rationality of people's inductive inferences.)

The stance taken in our discussion of backward induction, whereby people are rational yet they entertain the possibility that others are not, presents a subtler problem. What must be remembered here is that, as a positive theory, RCT only claims that people try to act rationally (target article, sect. 3.3), and that the idealization of perfect rationality should give qualitatively correct predictions. Of course, in reality, people do err, and subjects are aware of this fact. Therefore, in forming expectations about their opponents' actions, subjects are open to the possibility of errors of reasoning by the opponent. Furthermore, as one progresses further back in the chain of reasoning entailed by backward induction, the expectation of such errors compounds. Thus, the framework proposed here can be viewed as idealizing rationality at the zero level, but not at higher orders of theory-of-mind reasoning.

Our thesis, that people follow instrumental rationality but anchor it on their model of the opponent, is supported by Hedden and Zhang's (2002) recent investigation of the order of theory-of-mind reasoning employed by subjects in three-step sequential-move games. On each trial, subjects, who controlled the first and third moves, were asked first to predict the response of the opponent (a confederate who controlled the second move) and their own best choice on the first move. Initially, subjects tended to predict myopic choices by the opponent, corresponding to level 0 reasoning (level 1 was optimal for the opponent). Accordingly, subjects' own actions corresponded to the level 1 strategy, rather than the level 2 strategy prescribed by CKR. However, after sufficient experience with an opponent who played optimally, 43% of subjects came to consistently predict the opponent's action correctly, and altered their own behavior to the level 2 strategy. Although the remaining subjects failed to completely update their mental model of the opponent, errors of instrumental rationality (discrepancies between the action chosen and that dictated by the expectation of the opponent's response) remained low and approximately constant throughout the experiment for both groups. These results support the claim that violations of the predictions of CKR can be explained through scrutiny of player's models of their opponents, without rejecting instrumental rationality, and suggest that further investigations of rational choice in game situations must take into account the distinction between instrumental and inductive rationality.

Analogy in decision-making, social interaction, and emergent rationality

Boicho Kokinov

Central and East European Center for Cognitive Science, Department of Cognitive Science and Psychology, New Bulgarian University, Sofia, 1618 Bulgaria. **bkokinov@nbu.bg**
http://www.nbu.bg/cogs/personal/kokinov

Abstract: Colman's reformulation of rational theory is challenged in two ways. Analogy-making is suggested as a possible candidate for an underlying and unifying cognitive mechanism of decision-making, one which can explain some of the paradoxes of rationality. A broader framework is proposed in which rationality is considered as an emerging property of analogy-based behavior.

Rationality has long been shown to fail as a descriptive theory of human decision-making, both at the individual and social levels. In addition, Colman presents strong arguments that rationality also fails as a normative theory for "good" decision-making – "rational" thinking does not produce optimal behavior in social interaction and even acts against the interests of the individual in some cases. Fortunately, human beings often act against the postulates of rationality and achieve better results than prescribed by the theory. Therefore, Colman concludes that "rationality" has to be redefined by extending it with additional criteria for optimization, such as the requirement for maximizing the "collective" payoff, or with additional beliefs about the expected strategies of the coplayers. He does not clarify how and when these additional criteria are triggered or where the common beliefs come from.

We are so much attached to the notion of rationality that we are always ready to repair it, but not to abandon it. The theory of rationality is, in fact, a formalization of a naive theory of human thinking. This naive theory makes it possible to predict human behavior in most everyday situations in the same way as naive physics makes it possible to predict natural phenomena in everyday life. However, no one takes naive physics so seriously as to claim that it provides "the explanation" of the world. Moreover, even refined and formalized versions of this naive theory, like Newtonian mechanics, are shown not to be valid; and more complicated and counterintuitive theories at the microlevel, like quantum mechanics, have been invented. On the contrary, rationality theory is taken seriously, especially in economics, as an explanation of human behavior.

Instead of extending rationality theory with additional socially oriented rules, it may be more useful to make an attempt to build a multilevel theory that will reveal the implicit and explicit cognitive processes involved in decision-making. These underlying cog-

nitive mechanisms produce decisions, which are sometimes "individually rational," sometimes "collectively rational," and sometimes "not rational at all." Because these mechanisms have been evolved and developed to assure human survival, they will, most of the time, produce results that are "rational" or "optimal" from some point of view – this is what makes rationality a good naive theory. However, this does not mean that people explicitly follow the rules of maximization prescribed by the theory.

Colman proposes an eclectic collection of ad-hoc strategies (team reasoning, Stackelberg reasoning, epistemic, and nonmonotonic reasoning), which are all different forms of explicit deductive reasoning. Deduction can certainly play a role in decision-making, but it is not enough to explain it. Recent studies revealed that analogy-making is a more basic mechanism of human thinking, which is present from early infancy and is used ubiquitously in everyday life (Gentner et al. 2001). Analogy-making is a process of perceiving one situation (target) in terms of another (base), thereby preserving the system of relations among elements and transferring knowledge from the base to the target. Arguments have been presented that deduction is in fact based on analogy, and a special form of it (Halford 1993; Kokinov 1992). Markman and Moreau (2001) have reviewed the evidence that analogy plays an important role in perceiving and framing the decision situation, as well as in comparison of the alternatives. Moreover, analogy may be used both explicitly and implicitly (Kokinov & Petrov 2001; Markman & Moreau 2001). Thus, analogy may play a unifying role in describing the mechanisms of decision-making.

Analogy-making may explain the paradoxes of using the focal points described by Colman. They are easily perceivable and analogous to focal points in other games. Therefore, it is natural to expect people to use them again and again if previous experience of using a focal point has been successful. Similar arguments may be applied to social dilemmas and trust games. If another player has used a certain strategy in a previous case, I may expect him or her to behave the same way in an analogous situation, and thus have a prediction for his or her behavior.

Analogies may be applied at various levels: Analogies to previous cases of decision-making in the same game or analogies to games with similar structure; analogies to cases of social interaction with the same individual or to cases of social interactions with individuals who are considered analogous (i.e., are in similar relations to me, like family or team members). Thus, even a novice in a particular game can still use his or her previous experience with other games.

Analogy can explain the "deviations" from the prescribed "rational" behavior and the individual differences among players. If a player has an extensive positive experience of cooperative behavior (i.e., many successful cases of benefiting from acting together), and if the current game is found to be analogous to one of these cases, then he or she might be expected to act cooperatively (even if this is not the optimal strategy). On the contrary, if the game reminds the player of a previous case of betrayal or fraud, then defection strategy should be expected.

In summary, analogy may play a crucial role in a future theory of decision-making. Instead of explaining rationality with rules for utility maximization, which people follow or break, we may explain human behavior by assuming that decisions are made by analogy with previous cases (avoid strategies that were unsuccessful in analogous situations and re-use strategies that were successful). Thus, utility maximization is an emergent property that will emerge in most cases, but not always. In this view, rationality is an emergent phenomenon, and rational rules are only a rough and approximate explanation of human behavior.

Wanted: A reconciliation of rationality with determinism

Joachim I. Krueger

Department of Psychology, Brown University, Providence, RI 02912.
joachim_krueger@brown.edu
http://www.brown.edu/departments/psychology/faculty/krueger.html

Abstract: In social dilemmas, expectations of reciprocity can lead to fully determined cooperation concurrent with the illusion of choice. The choice of the dominant alternative (i.e., defection) may be construed as being free and rational, but only at the cost of being incompatible with a behavioral science claiming to be deterministic.

The conspicuous failure of orthodox game theory is its inability to account for cooperative behavior in noniterated social dilemmas. Colman outlines a psychological revision of game theory to enhance the predictability of hitherto anomalous behavior. He presents the Stackelberg heuristic as a form of evidential reasoning. As Colman notes, evidential reasoning is assumed to lead respondents to shun the dominating alternative in Newcomb's problem and in decisions to vote. In the prisoner's dilemma game (PDG), however, Stackelberg reasoning leads to defection (Colman & Stirk 1998). Thus, Stackelberg reasoning appears to be neither evidential nor parsimonious in this domain. After all, players can select the dominating alternative in the PDG without making any predictions of what their opponents will do. How, then, can evidential reasoning lead to cooperation?

The logic of the PDG is the same as the logic of Newcomb's problem (Lewis 1979). Just as players may expect that their choices will have been predicted by Newcomb's savvy demon, they may expect that their choices in the PDG will most likely be matched by their opponent's choices (unless the rate of cooperation is exactly 50%). The issue is whether this statistical realization gives cooperators (or one-boxers, in Newcomb's case) license to lay claim to being rational.

Orthodox game theorists insist on defection, because a player's cooperation cannot make an opponent's cooperation more likely. Evidentialists, however, claim that cooperation may be chosen without assuming a causal effect on the opponent's choice. Only the assumption of conditional dependence is needed. If nothing is known about the opponent's choice, conditional dependence is obvious *after* a player committed to a choice. By definition, most players choose the more probable alternative, which means that the choices of two independent players are more likely to be the same than different (Krueger 1998). Because time is irrelevant, it follows that it is more likely that two players *will* make the same, instead of different, choices. In the extreme case, that players expect their responses to be reciprocated without fail, their dilemma devolves into a choice between mutual cooperation and mutual defection. As mutual cooperation offers the higher payoff, they may choose cooperation out of self-interest alone.

Evidentialist reasoning is distasteful to the orthodox mind because it generates two divergent conditional probabilities that cannot both be correct (i.e., p[opponent cooperation/own cooperation] $>$ p[opponent cooperation/own defection]). Choosing the behavior that is associated with the more favorable prospect then smacks of magical thinking. But causal assumptions enter at two levels: at the level of the investigator and at the level of the participant. Investigators can safely assume that players' efforts to influence their opponents are pointless. Players, however, may *think* they can exert such influence. Although this expectation is irrational, it does not invalidate their cooperative choices. Note that investigators can also subscribe to a more plausible causal argument, which holds that both players' choices result from the same set of latent variables. These variables, whatever they may be, produce the proportions of cooperation found in empirical studies. Players who realize that one option is more popular than the other, but do not know which, can *discover* the popular choice by observing their own. The fact that they may have an experience of

unfettered choice, and perhaps even hope to influence their opponents, is quite irrelevant (Wegner 2002).

The burgeoning literature on social dilemmas suggests that individual behavior in these situations presents a more poignant dilemma to the investigators than to the participants. However modest their predictive successes may be, experimental studies of social behavior rest on a bedrock assumption of determinism. In this spirit, experimentalists assume that individuals' judgments and decisions are fully determined (Bargh & Ferguson 2000). It is ironic that research participants who are cast into the PDG or confronted with Newcomb's problem can satisfy norms of rationality only by denying any determining effect on their own behavior that would make them act like most others.[1] They are enjoined to choose defection without drawing any inference as to what this might say about their opponents' choices. Evidentialists, in contrast, can maintain a deterministic outlook without being perplexed. They need only assume that cooperators choose "as if" they were free.

Incidentally, players working on the assumption that their own choices will likely be reciprocated are also comfortable with common-interest games. They do well without experiencing the puzzlement of orthodox game theorists and even without resorting to von Stackelberg's best-bet heuristic. Perhaps more importantly, evidential reasoning preserves *methodological individualism* in common-interest games. Collective preferences, as entailed by team spirit, are unnecessary. A game in which players are paid only if their choices do not match, however, would be a true puzzle to the evidentialist and the orthodox alike. Even team spirit, no matter how lofty its intent, cannot overcome this hurdle.

NOTE

1. In iterated PDGs, the assumption of determinism is more apparent than in one-shot games. Players' choices are assumed to be controlled by the design of the game (i.e., the experimenters) and by each other's choices in preceding rounds (e.g., Rachlin 2002).

Let's cooperate to understand cooperation

John Lazarus

Evolution and Behaviour Research Group, Psychology, School of Biology, University of Newcastle, Newcastle upon Tyne, NE2 4HH United Kingdom.
j.lazarus@ncl.ac.uk
http://www.ncl.ac.uk/biol/staff/john_lazarus.html

Abstract: The importance of understanding human cooperation urges further integration between the relevant disciplines. I suggest ideas for bottom-up and top-down integration. Evolutionary psychology can investigate the kinds of reasoning it was adaptive for humans to employ. Disciplines can learn from each other's approaches to similar problems, and I give an example for economics and evolutionary biology.

Understanding the factors that facilitate and constrain human cooperation is of the greatest importance. I suggest here ways in which disciplines with a convergent interest in cooperation might fruitfully interact, with an emphasis on theoretical modelling.

Colman describes "nonstandard forms of reasoning" that help to explain irrational social decisions. Psychological game theory should employ the methods of evolutionary psychology (Tooby & Cosmides 1992) to determine both the kinds of social problems that early humans were selected to solve, and the kinds of reasoning that were adaptive to employ. Such an analysis of social problems has shown that human reasoning is well-designed for cheater detection, for example (Cosmides & Tooby 1992). An evolutionary analysis of kinds of reasoning could start with team reasoning (target article, sect. 8.1), for which two potential adaptive explanations seem worth pursuing. Team reasoning might be favoured where cooperation benefits the group, or where maximizing collective payoff raises one's reputation and thus brings future rewards (Milinski et al. 2002). Evolutionary game theory is the tool for analyzing the evolutionary fate of competing modes of reasoning.

Knowledge of social decision-making in dyads and small, unstructured groups is a starting point for understanding cooperation at the higher levels of structured groups, firms, institutions, communities, and states (cf. Hinde 1987). Table 1 (see overleaf) lists disciplines sharing an interest in cooperation, indicating their interests, methods, and levels of analysis; it is not exhaustive (e.g., nothing on military strategy). Its purpose is to indicate the multidisciplinary nature of cooperation, to encourage further interdisciplinary work (following, e.g., Axelrod 1984; 1997; Frank 1988), and to act as a reference point for the following proposals in this direction.

Colman shows that there is much to be done before we understand cooperative decision-making at the lowest level, although understanding should be advanced by reference to the social psychological foci in Table 1. To bring greater psychological reality to decision theory in the structured groups of institutions and societies, game theory models and psychological game theory findings should be combined with the decision-making models of economics and related disciplines (Table 1; see also Axelrod 1997).

This bottom-up approach should be complemented by psychological game theory adopting top-down insights gained from analyses of real-life economic behaviour. Decision-making in these real-life contexts may reflect evolved predispositions, and may tap motivations at work even in the economically elementary scenarios of the psychological laboratory. For example, studies of the way in which communities govern their own use of common pool resources (CPRs), such as grazing pastures (Ostrom 1990), may reveal evolved influences on cooperative decision-making, and even evolved modes of reasoning, because the hunting and gathering activities of early humans also have CPR properties. Successful CPR decisions are characterized by: a clear in-group/out-group distinction; resource provision in proportion to need and sharing of costs in proportion to ability to pay; and graded punishments for the greedy (Ostrom 1990). Whether these characteristics apply to decision-making in other kinds of cooperative relationship is open to evolutionary psychological and empirical analysis. It would be valuable to know whether cooperation was rational and evolutionarily stable in CPR scenarios.

In addition to bottom-up and top-down integration, different disciplines can surely learn from each other's approaches to similar problems. I close with an example. In economics, a common pool resource is "subtractable," because resources removed by one person are unavailable for others. In contrast, a pure public good (e.g., a weather forecasting system) is "nonsubtractive" in that its use by one person leaves it undiminished for others (Ostrom 1990, pp. 31–32). In evolutionary biology, parental investment in offspring is of two kinds, "shared" and "unshared," respectively, the identical concepts just described from economics. Food for the young must be shared among them, whereas parental vigilance for predators is enjoyed by all simultaneously. Modelling in the evolutionary biology case has examined the influence of the number of users on the optimal allocation of investment, and on conflict between producer (parent) and user (offspring) (Lazarus & Inglis 1986). Could economists use these results? Have economists produced similar results that evolutionary biologists should know about?

ACKNOWLEDGMENTS

I am grateful to the "Society and Modernization" seminar group at the University of Newcastle for broadening my horizons.

Table 1 (Lazarus). *Approaches to cooperation*

Discipline	Focus	Levels of Analysis	Methods	Sample References
Ethology	Cooperation in the animal kingdom	Dyad Group	Field work; Laboratory experiments	Dugatkin 1997
Evolutionary biology	Biological and cultural evolution	Dyad Group	Game theory; simulation and analytical modelling; evolutionary stability	Axelrod & Hamilton 1981; Boyd & Richerson 1991; Roberts & Sherratt 1998
Artificial intelligence	Artificial societies; computing applications; trust	Group Network	Agent-based simulation modelling; complexity theory	Andras et al. 2003; Axelrod 1997; Schillo et al. 2000
Psychology				
Evolutionary psychology	Evolutionary origin, adaptive biases, brain modularity	Dyad Group	Laboratory experiments	Cosmides & Tooby 1992
	Commitment and the emotions	Dyad Group	Evolutionary theory; laboratory experiments	Frank 1988
Psychological game theory	Rationality; biases in decision-making; framing effects; influences on cooperation	Dyad Group	Game theory; laboratory experiments	Milinski et al. 2002; Colman, target article
Developmental psychology	Moral development	Dyad Group	Laboratory experiments and natural observations; Cross-cultural comparison	Kohlberg 1984
Social psychology	Egoistic or altruistic motivation?	Dyad Group	Laboratory experiments	Batson 1987; Cialdini et al. 1987
	Empirical notions of reciprocity; equity; desert and fairness	Dyad Group	Questionnaire studies; evolutionary psychology	Buunk & Schaufeli 1999; Charlton 1997; Wagstaff 2001
	Cooperation within and between groups	Group(s)	Laboratory experiments; field work	Feger 1991; Rabbie 1991
	Trust	Group	Field work, discourse	Hardin 2002; Kramer & Tyler 1996
Anthropology	Social exchange; social hunting	Dyad Group	Field work	Kaplan & Hill 1985; Kelly 1995
Sociology	Trust	Group	Discourse	Hardin 2002
Economics	Trust	Group	Field work	Kramer & Tyler 1996
	Common pool resources	Common resource group	Game theory; field work; historical studies	Ostrom 1990
	Public choice	Public goods	Economic decision theory	Margolis 1982; van den Doel & van Velthoven 1993
Political philosophy	Collective action	Community State	Game theory	Taylor 1987
	Distributive justice	Community State	Discourse	Rawls 1999
	Trust	Community State	Discourse	Hardin 2002
Ethics	Moral behavior	Dyad Group	Metaethics; Cross-cultural comparison	Arrington 1998; Yeager 2001

Game theory need not abandon individual maximization

John Monterosso[a] and George Ainslie[b]

[a]*Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California at Los Angeles, Los Angeles, CA 90024;* [b]*Department of Psychiatry, Coatesville VA Medical Center, Coatesville, PA 19320.* **jmont@ucla.edu** **george.ainslie@med.va.gov**

Abstract: Colman proposes that the domain of interpersonal choice requires an alternative and nonindividualistic conception of rationality. However, the anomalies he catalogues can be accounted for with less radical departures from orthodox rational choice theory. In particular, we emphasize the need for descriptive and prescriptive rationality to incorporate recursive interplay between one's own choices and one's expectation regarding others' choices.

Colman proposes that an alternative conception of rationality is required to account for human interaction, and he provides some suggestions in this direction. What he specifically sees the need to give up is "methodological individualism" – the premise that "rational play in games can be deduced, in principle, from one-person rationality considerations" (Binmore 1994a, quoted in target article, sect. 4.1, para. 1). We think the anomalies he catalogues can be accounted for without abandoning this foundational principle of deterministic behavioral science.

First, the prevailing payoffs in experimental games are not the same as the specified payoffs. Social interactions are rife with invisible contingencies that are impossible to bring under full experimental control. Human beings are fundamentally social creatures, which entails the presence of powerful interpersonal motivations, too numerous to list. Otherwise, anomalous play, such as rejecting a low offer in a one-shot ultimatum game or cooperating in a one-round prisoner's dilemma game, is sensible if we allow that the dollars offered do not exhaust the prevailing payoffs. Colman discusses this type of proposal (Camerer's "behavioral game theory," Rabin's fairness equilibrium), but he concludes it is not enough to account for all the phenomena he presents. We agree, and furthermore, do not think that the subjects' many motives, beyond maximizing the specified matrix outcomes, are orderly enough to inspire any useful addition to game theory (such as adding X points to particular cells); discrepancies between the specified payoffs and the prevailing payoffs will always be noise in the experiment, the friction that distorts the ideal physics lab.

However, permitting the free use of probability estimates of other's choices should be enough to let "methodological individualism" both describe and prescribe rationality to the extent that subjects *are* motivated by the specified matrices of the game. Of particular importance in explaining otherwise anomalous play is the subject's use of her own inclinations and behavior as test cases that inform her expectations regarding what others will do. In the kinds of situations game theorists care about, it is neither descriptively tenable *nor prescriptively effective* to require individuals to finalize their assessments of what others will do prior to considering what they will do. Instead, we think that a rational individual is simultaneously engaging in both computing expectation of what the other player will be motivated to do and contemplating what she herself should do, and each process informs the other. In the absence of specific information about one's counterpart, what better basis is there to predict her behavior than via one's own response to the situation?

Colman describes such a recursive process in characterizing one attempt to develop a game-theoretic rationale for the Pareto-dominant H-H solution in the Hi-Lo game. In this account, Player I assumes by default that Player II's strategies are equally probable. She thus concludes she should choose H, because the probability-weighted sum is higher. But this, Colman adds, violates rational choice theory. "By the transparency of reason, Player I 's intention to choose *H* would be common knowledge and would induce Player II to choose the best reply, namely *H*, with *certainty*, contradicting Player I's initial assumption" [i.e., of equal probability of moves] (sect. 5.6, para. 4). While such recursion may violate game theory's constraints, we think it is descriptively accurate, prescriptively rational, and it does not entail abandoning methodological individualism.

The recursion between someone's own perceived choices and their expectations about the choices of others is easier to see in a *discoordination* variant of the Hi-Lo game, in which players get to keep their choice (in some monetary unit) if and only if they chose differently from each other. With no a priori expectation regarding what Player II will choose, Player I's first-order inclination is to choose the high amount, following the same logic as above. But seeing the similarity of her counterpart's predicament, she may expect her to have thought the same way, giving her the second-order inclination that she must go for the lower amount to get anything. But then again, if she thinks her counterpart is a similarly sophisticated sort, she might get the feeling that her counterpart went through the same thought process, thus giving her the third-order inclination that maybe she *should* therefore go for *H*. The more similar she thinks her counterpart to be to herself, the more dizzying the potential for iteration, and the less likely there will be a probable solution.

The recursive prediction model has the advantage that it also provides intertemporal bargaining within the individual person. In situations that involve resisting temptation, individuals cannot be certain of their own future choices. The need to choose in the present, with an eye to the precedent this choice will set for the future (e.g., whether or not I am sticking to my diet), places people in a situation analogous to a repeated prisoner's dilemma (PD) game, as we have argued elsewhere (Ainslie 2001, pp. 90–104; Ainslie & Monterosso 2003; Monterosso et al. 2002). Briefly, the danger that future selves will see past violations of a resolution as a reason to violate it, in turn, is similar to the danger that one player's defection will cause the other(s) to defect. But, in this bargaining, a person may propose a choice to herself ("I'll have an ice cream"), then put herself in the shoes of her future self to evaluate it retrospectively ("I'll have gone off my diet"), then revise her current choice in light of this evaluation ("I'll have a muffin instead"), and evaluate this ("no"), and propose again ("a *bran* muffin") at some length before making a single concrete choice. Choices may turn out to divide along salient features, just as in the Hi-Lo game, not because of their intrinsic payoff, but because they make intertemporal cooperation more likely. Intertemporal bargaining theory predicts the emergence of both positive and negative features that have been ascribed to willpower. It generates an internal version of Adam Smith's "unseen hand" without assuming an innate faculty of self-control.

Second-order indeterminacy

Marco Perugini

Department of Psychology, University of Essex, Colchester, CO4 3SQ United Kingdom. **mperug@essex.ac.uk**
http://privatewww.essex.ac.uk/~mperug

Abstract: Psychological game theory, as defined by Colman, is meant to offer a series of solution concepts that should reduce the indeterminacy of orthodox game theory when applied to a series of situations. My main criticism is that, actually, they introduce a second-order indeterminacy problem rather than offering a viable solution. The reason is that the proposed solution concepts are under-specified in their definition and in their scope.

Colman looks at game theory from a psychological perspective. In the first part of his article, he convincingly argues about the limitations of orthodox game theory, especially when applied to social interactions. The examples are well chosen and the case is well built. This is an important contribution that might help us to focus, once and for all, on these important issues. However, Colman's suggestion of psychological game theory as a way forward to overcome the severe limitations of orthodox game theory in ex-

plaining social interactions is not entirely convincing. The spirit behind this attempt should be praised, yet psychological game theory as defined and exemplified by Colman does not offer a truly viable solution. The key problem is that the suggested solutions are theoretically under-specified, quite limited in scope, and lead to a second-order indeterminacy.

To illustrate my point I will focus on the concept of "team reasoning." What is so special about team reasoning that cannot be said about other ways of reasoning? For example, one might define "altruistic reasoning," "individualistic reasoning," "fairness reasoning," "reciprocity reasoning," and so on, in the same kind of holistic way as the definition is offered for "team reasoning." It is easy to find examples of games that can be solved using some of these concepts; although they can be solved promptly also via "team reasoning," the intuition is that it would not necessarily be the best solution concept. By best solution concept I mean a concept that is intuitively compelling and likely to be empirically supported with actual behavioral data.

I will present two examples of games. For the first example, let's consider all modified coordination games for two players with asymmetrical payoffs. Let's consider this asymmetric coordination game with the following payoffs and choices (Fig. 1):

As for every coordination game, a standard analysis would show two Nash equilibria (*H*, *H* and *L*, *L*), and the issue would be how to select one of the two. Applying a team reasoning would single out *H*, *H* as the best equilibrium. Would this be a compelling solution? I doubt it. If I were Player I, I would think twice before choosing *H*. By applying "fairness reasoning" or "Reciprocity reasoning," I could anticipate that Player II would like *L*, *L* much more than *H*, *H* (or, put differently, dislike much more the inequality of payoffs resulting from *H*, *H*). I would therefore anticipate that the other player would play *L*, and as a consequence I would decide to play *L*. On the other hand, if I were to apply "altruistic reasoning" or "individualistic reasoning," for opposite reasons I should come to the conclusion that Player II will play *H*, and hence so would I. The problem is threefold: First, we can list a series of reasoning concepts besides "team reasoning"; second, psychological game theory, as defined by Colman, would offer no tools to select among these different reasoning concepts; and third, the solution concept which would be the best for a player, depends on his expectations about the other player's type.

The second example is perhaps even more intriguing.[1] The Ultimatum Game (UG) is a well-known paradigm that has been the subject of several studies in experimental economics and in social psychology. The UG is a very simple game whereby two players bargain over a given monetary endowment. The first player proposes a division of the endowment and the second player can either accept or refuse it. If she refuses it, both players end up with nothing. Orthodox game theory predicts that the first player will propose a small amount for the second player (e.g., 99% for self vs. 1% for other) and the second player will accept the proposal. However, several experimental studies have found systematic deviations from these predictions (e.g., Guth 1995; Thaler 1988). It is well established that a consistent portion of second players would reject low offers (e.g., 25% or lower) even though this means that both players end up with nothing. What about team reasoning? A team-reasoning second player should never reject any offer, because from the perspective of a second player the strategy that maximizes the joint payoff is to accept any offer. In fact, for every offer, the alternative would be to reject it, which is always dominated in terms of joint payoffs, given that it implies no payoff for both players. Therefore, a team reasoning second player would be equally likely to accept a 1/99 or a 50/50 split. The intriguing conclusion is that a team-reasoning player often will behave exactly as dictated by orthodox game theory, even in those situations where our intuition would suggest we do otherwise.

		I: H	I: L
II	H	100, 10	0, 0
	L	0, 0	9, 9

Figure 1 (Perugini). Example of a coordination game with asymmetric payoffs.

Equally problematic are those cases where team reasoning offers different predictions from orthodox game theory. Take social dilemmas. Of course, social dilemmas can be solved by using team reasoning, but this is equally true for several of the nonstandard solution concepts that I have sketched previously. I wonder how well a team reasoning concept would fare when compared with other nonstandard solution concepts across a comprehensive range of social dilemmas. To sum up, I am not convinced that team reasoning can be a good solution to much more than the specific example of the Hi-Lo matching game with symmetrical payoffs illustrated by Colman. But then, why should it not be named "matching reasoning" instead?

These examples illustrate my main problem with Colman's suggestions: Concepts such as team reasoning must be defined more precisely, which ultimately means that it will be necessary to specify the payoffs involved, how they are transformed, and under which conditions each solution concept primarily applies. The preceding examples have made clear that an important parameter is the symmetry of the payoffs for the players: Everything else being equal, the more asymmetrical the payoffs, the less likely is that team reasoning can offer a compelling solution for all players. But this implies that the team reasoning concept should specify what level of asymmetry is acceptable to the players, which ultimately means to specify some function of weighting the payoffs involved. Only in this way can the solution concepts pass more stringent theoretical and empirical tests. The alternative would be to have a storage bin full of loose ad-hoc reasoning concepts that can be used post-hoc for different situations, but without any rule that specifies when and why they should be adopted. In other words, ironically, the lack of a reason for choosing, which was the main point behind many of Colman's sharp criticisms on the indeterminacy of orthodox game theory, will strike back with a vengeance. Without specifying the concepts more precisely – given that they can explain or predict only some interactions and not others, and that alternative nonstandard concepts can be compellingly applied in several circumstances – we will be left without any reason why to apply a given nonstandard psychological solution concept in the first place.

ACKNOWLEDGMENT

Preparation of this commentary was supported by RPF grant DGPC40 from the University of Essex.

NOTE

1. I owe this example to Tim Rakow.

Chance, utility, rationality, strategy, equilibrium

Anatol Rapoport

Department of Psychology, University of Toronto, Toronto, Ontario M5S 3G3, Canada. **anatol.rapoport@utoronto.ca**

Abstract: Almost anyone seriously interested in decision theory will name John von Neumann's (1928) Minimax Theorem as its foundation, whereas Utility and Rationality are imagined to be the twin towers on which the theory rests. Yet, experimental results and real-life observations seldom support that expectation. Over two centuries ago, Hume (1739–40/1978) put his finger on the discrepancy. "Reason," he wrote "is, and ought to be the slave of passions, and can never pretend to any other office than to serve and obey them." In other words, effective means to reach specific goals can be prescribed, but not the goals. A wide range of experimental results and daily life behavior support this dictum.

In November 1945, a conference of mathematicians was held at the Museum of Science and Industry in Chicago. A robot was displayed in the entrance hall, inviting the visitors to play a game of tic-tac-toe. Needless to say, regardless of who made the first move, every game ended either in a draw or in a win for the robot. Many were impressed. Today, an exhibit of this sort would be unthinkable, except, possibly in a children's section.

The term "game," in the context of interacting actors with usually different goals, was introduced by von Neumann and Morgenstern (1944) in their seminal treatise *Theory of Games and Economic Behavior.* It will surprise many that von Neumann did not recognize chess as a "game" in the sense that he used the term.

"Chess is not a game," von Neumann told Jacob Bronowski, who worked with him during World War II (Poundstone 1992, p. 6). He meant that there is a "correct" way to play the game – although no one presently knows what it is – and that the game should therefore be trivial, in much the same sense as tic-tac-toe is trivial to players aware of a "correct" strategy.

It turned out that the inspiration for game theory was not chess or any parlor game, which can be shown to have one or more "correct" ways to play it, but poker instead, where it is not possible to guess with certainty the choice of strategy of one's opponent(s).

According to Luce and von Winterfeldt (1994), "[M]ost people in real situations attempt to behave in accord with the most basic (conventional) rationality principles, although they are likely to fail in more complex situations." The "situations" are not mentioned, but one can surmise that "utility" is (perhaps tacitly) represented in them by a linear function of some "good" (money, survivors), and that expected utilities are either given (EU) or subjectively assumed (SEU).

This observation tends to imply that normative (prescriptive) decision theory has a positive role to play along with recent empirical descriptive approaches, which seek to gain understanding of how people actually make decisions in a great variety of situations. Yet, already in the late eighteenth century, maximization of expected utility was shown to lead to absurd results in the so-called Petersburg Paradox. A fair coin is thrown. The gambler wins $n2^{n-1}$ rubles if "heads" appears n times before the first "tails." The gambler's expected gain is infinite. Daniel Bernoulli (1738/1954), to whom the invention of the game is attributed, modified the rule, whereby expected utility increased logarithmically rather than linearly with n. This still made the expected gain, and thus a "rational" maximum stake, enormous, and hence unacceptable to any "rational" player. Indeed, as long as expected gain increases monotonically with n, a rule can be devised to make the expected gain enormous. No "rational" gambler can be expected to pay anywhere near it for the privilege of playing the game once.

Passing from gambling to two-or-more-person games, we encounter similar difficulties with prescriptive decision theory. Especially impressive are paradoxes resulting from backward induction. Consider the Prisoner's Dilemma game played a large known number of times. In a single play, defection by both players is a minimax outcome, which, according to von Neumann, is the only rational one. In a long sequence of plays, however, one might suppose that repeated cooperation (CC) might emerge, as each player forestalls the other's "revenge" for defection. Nevertheless, the last "rational" outcome ought to be double defection (DD), because no retaliation can follow. Given this conclusion, the next to the last play also ought to be (DD), and so on down to the first play.

When Flood and Dresher, discoverers of the Prisoner's Dilemma game (Poundstone 1992), reported to von Neumann that in a long sequence of plays of the game, the outcome was not at all a solid string of DD's, as predicted by the minimax theorem, the great mathematician did not take the result of the admittedly informal experiment seriously. Subsequent experiments, however, showed that, especially in long repeated plays, substantial runs of CC are a rule rather than an exception. Even in single plays by total strangers, frequent CC outcomes have been observed (Rapoport 1988).

Colman cites backward induction in "Centipede," a weirdly designed multi-move game in which both players could win fabulous sums if they tacitly agreed to cooperate after the first play. Nevertheless, backward induction would dictate stopping after the first play, whereby both would receive zero. In contrast, backward induction in R. Selten's "Chain Store" game prescribes CC throughout. The inventor of the game writes that, in the role of the chain store, he would not play as prescribed and presumably would get more money (Selten 1978). Luce and Raiffa (1957) also preferred to violate the backward induction prescription in finitely repeated Prisoner's Dilemma, thus avoiding the only minimax equilibrium of this game.

It turns out that these frequent failures of rational choice theory to prescribe acceptable actions in gambling or game-like situations can be traced to two often tacitly implied assumptions, namely, rejection of so called "evidential" decision theory (Joyce 1999) and independence of decisions of individual players.

Suppose a believer in predestination (Calvinist, Presbyterian) is asked why, if his ultimate abode is fixed, he leads a sober and chaste life. Why doesn't he drink, gamble, chase women, and so on while he can? He might answer, "Since God is just, I can assume that I am among the saved, because I live as I live." He considers his frugal and chaste life as "evidence" that he has been saved and he cherishes this feeling (Joyce 1999).

Asked why he bothers to vote in a general election, seeing that his vote can't possibly make a difference in the result, Herr Kant replies, "I vote, because I would like everyone to vote and because it makes me feel that I have done my duty as a citizen." In the wilderness, Dr. Z has one dose of a life-saving medicine. If given to Mr. X, it will save his life with a probability of 0.9; if given to Mr. Y it has a probability of 0.95. Maximization of expected utility demands that she give the medicine to Mr. Y. But Dr. Z tosses a fair coin to decide. She doesn't want to "play God."

The concept of rationality in classical prescriptive decision theory has three weak spots: individualism, decision independence, and the minimax equilibrium dogma. "Individualism" in this context means "egoism." To avoid the pejorative connotation, Wicksteed (1933) called it "non-tuism." Decision independence is dropped in evidential decision theory. "However I decide, so will my co-players, since there is no reason to suppose that they think not as I do." This perspective dethrones the Nash equilibrium from its role as a *sine qua non* condition of rational choice.

In spite of his generous appreciation of game-theoretic contributions to decision theory, Colman effectively pronounces the end of prescriptive theory founded on the orthodox paradigm, and discusses the promising dawn of inductive experimental-psychological approaches.

Why not go all the way

Richard Schuster

Department of Psychology, University of Haifa, Haifa 31905, Israel.
schuster@psy.haifa.ac.il

Abstract: "Psychology Game Theory" grafts social-process explanations onto classical game theory to explain deviations from instrumental rationality caused by the social properties of cooperation. This leads to confusion between cooperation as a social or individual behavior, and between *ultimate* and *proximate* explanations. If game theory models explain the existence of cooperation, different models are needed for understanding the proximate social processes that underlie cooperation in the real world.

Colman's provocative paper reminds me of a familiar scenario in science. A popular doctrine is under stress but refuses to die. Instead, it is amended again and again in a vain attempt to forge an accommodation with a new reality. A good example is the assumption that individual self-interest, which can explain the evolution of cooperation, must also underlie the *behavior* of cooperating in the real world. Colman characterizes this assumption (from Hume) as "instrumental rationality." The stress comes from

the emerging reality (cf. Palameta & Brown 1999) that humans, in a variety of interactive decision games, prefer to cooperate more than predicted from instrumental rationality alone. Colman's proposal is to explain the cooperation bias with "Psychological game theory," a set of "formal principles of nonrational reasoning" linked to the social context of cooperation in which the actions and motives of other participants may not be completely known. This commentary will suggest that Colman's allegiance to the basic paradigm of game theory does not allow his theory to go far enough when addressing the social reality of cooperation under free-ranging conditions.

The root of the problem seems to be the familiar mistake of confounding *ultimate* versus *proximate* explanations. If game theoretical models are designed to provide an ultimate explanation for the existence of cooperation, such models only create obscurity and confusion when they are stretched to incorporate what happens when live subjects are used in the kinds of experiments spawned by game theory. Setting aside the evolution of cooperation and altruism at the level of groups (e.g., Sober & Wilson 1998), natural selection at the level of individuals is necessarily selfish. But as Colman realizes (see also Chase 1980; Dennett 1995), classical game theory was never meant to be about behavior. Evolution is a process that guarantees maximization of individual fitness precisely because it is mindless and therefore "nonrational," leading to higher payoffs in the absence of intentional choosing. It is this nonrational aspect of cooperation that is grist for the mill of game theory, whether in behavioral ecology (e.g., Mesterton-Gibbons & Dugatkin 1992), the social sciences (e.g., Luce & Raiffa 1957), or economics (e.g., Arrow 1963).

When humans or animals actually choose whether or not to cooperate, they are no longer "mindless," in the sense that their choices are no longer immune from the influence of proximate psychological processes evoked by the presence and behaviors of others (Boesch & Boesch 1989; Roberts 1997; Schuster 2001; 2002; Schuster et al. 1993). The existence of such processes invites the question of how best to study them. By grafting psychological game theory onto classical game theory, the result is a peculiar kind of hybrid explanation that is part social and part individual, part proximate and part ultimate. On the one hand, cooperation is retained as an individual behavior, resting on "a bedrock of *methodological individualism*" (cf. target article, sect. 8.1, emphasis Colman's). This is obvious from the basic design of laboratory experiments in which social interaction is deliberately minimized or totally absent. Instead, anonymous players are physically isolated in separate cubicles and individually compensated according to how all players have chosen between strategic options.

On the other hand, humans playing such games show a persistent bias towards cooperating, despite the impoverished social conditions. The experimental conditions leave the experimenter with few social variables to manipulate. And the theoretician is left with the less-than-enviable task of speculating about why decision games with human subjects might generate too much cooperation. One proposal is to suggest traits such as "group-identity" or "social value orientation," which can vary across subjects. Colman's proposal is a set of unobservable intervening variables, such as "focal points," "collective preferences," and "team reasoning." Still another possibility is to speculate that humans, despite the physical isolation, are nevertheless influenced by the possibility that outcomes also depend on others (e.g., Forsythe et al. 1994). A player might then opt for cooperation because it is intrinsically reinforcing (Frank 1988; Schuster 2001; 2002; Sober & Wilson 1998), or merely to avoid the embarrassment of meeting opponents who were hurt by defection. At the end of the day, all such explanations remain plausible, but highly speculative, with little likelihood that they could be disentangled using the nonsocial experimental conditions associated with game theory.

To better understand the social dimensions of cooperation, and their impact on instrumental rationality, one alternative is to study examples such as team sports, military strategies, and collaborative business ventures under free-ranging conditions, where cooperation is intrinsically social (see Dugatkin 1997 for a review of animal examples). The social properties of cooperation can then be separately analyzed *before, during* and *after* engaging in a bout of cooperation (Schuster 2001; 2002). Instead of the anonymity among partners favored by game theorists, real-life "players" are usually familiar from group membership and/or prior encounters. Cooperators are therefore likely to have preferred partners (Dugatkin 1995). With interaction unrestricted, social influences also impact on how cooperation is performed, including the ability to coordinate actions, use signals, and develop a division-of-labor based on different and complementary roles (Boesch & Boesch 1989; Schuster et al. 1993; Stander 1992). There is also the possibility that the cooperators themselves are affected by working together in ways that affect the incentive to cooperate (Schuster 2001; 2002). Finally, social influences can affect how outcomes are allocated following an act of cooperation, because of factors such as sex, age, aggression, or status, which influence priority of access (e.g., Boesch & Boesch 1989; Noë 1990). In humans, allocation can also follow from bargaining or prior arrangements. Instrumental rationality is also violated when cooperation persists even though individual behavior might be more profitable (e.g., Packer et al. 1990).

A second alternative is to incorporate free-ranging conditions into experimental models of cooperation. One example is a simple model in which pairs of laboratory rats (*Rattus norvegicus*) are rewarded with a saccharine solution for coordinating back-and-forth shuttling within a shared chamber in which social interaction is unrestricted (Schuster 2001; 2002). Two experiments are relevant to instrumental rationality. In one, the option of competing over outcome allocation was modeled by reinforcing cooperative shuttles with intermittent presentations of either one or two cups of saccharine solution. Although pairs varied in how the single outcomes were allocated, from near "sharing" to strong dominance by one partner, there was *no* relationship between outcome disparity and the level of performance (Schuster et al., in preparation). The second experiment measured choice between entering one chamber, where the rats shuttled alone, or another chamber in which they always cooperated with the same partner. This situation is analogous to the game theory choice between cooperation and noncooperation, with the important difference being that the cooperation option, as in real life, was also a social option. Even though there was no advantage from choosing either chamber, because outcomes and the proportions of reinforced shuttles were matched, cooperation was strongly preferred (Schuster 2001; 2002; Schuster et al., in preparation).

By incorporating the social dimensions of cooperation into experimental models, it becomes possible to clarify the differences between ultimate explanations based on game theory or proximate explanations based on social processes. This clarification would not be needed if predictions from the two kinds of explanations were congruent. But congruence seems more likely when behavior is *not* social, so that the abstract prediction of maximizing fitness or expected utility can be better matched by the influence of feedback from individual experience (e.g., Staddon 1983). The cooperation bias shows that the advantage from using game-theory models, eliminating the "noise" generated by uncontrollable social interactions, can no longer be defended. The same methods are also inadequate for analyzing why the preference for cooperation deviates from instrumental rationality.

So why not go all the way and explain cooperation under free-ranging conditions by jettisoning the game-theory approach, in both method and theory, in favor of alternative models that are more faithful to the reality of cooperation in the real world? Then we could begin to study what is impossible with Colman's approach, namely, the experimental analysis of deviations from predictions based on instrumental rationality, and the likely identification of, as yet undiscovered, sources of reinforcement that underlie cooperation as we know it.

Locally rational decision-making

Richard M. Shiffrin

Psychology Department, Indiana University, Bloomington, IN 47405.
shiffrin@indiana.edu
http://www.cogs.indiana.edu/people/homepages/shiffrin.html

Abstract: Colman shows that normative theories of rational decision-making fail to produce rational decisions in simple interactive games. I suggest that well-formed theories are possible in local settings, keeping in mind that a good part of each game is the generation of a rational approach appropriate for that game. The key is rationality defined in terms of the game, not individual decisions.

Colman gives an intriguing, interesting, and at times amusing account of the failures of normative theories of rational decision-making. He suggests moving toward a "psychological" game theory that would be "primarily descriptive or positive rather than normative," and adds "a collection of tentative and ad hoc suggestions" (target article, sect. 8). I suggest that a well-formed psychological theory of rational decision-making may well be possible in local contexts (of a scope and generality large enough to be interesting). The approach is rooted in the thought that rationality itself is a psychological rather than axiomatic concept, justifying the need to reinvent it (or at least restrict it) for different settings.

I propose that all the decision-makers in a social/interactive game are faced with a dual task: They must decide (quite possibly without any communication) what theory of rational decision-making applies in that situation, and given that, whether a jointly rational solution exists, and what it is. The first of these tasks is not typically made explicit, but is a necessary consequence of the current lack of a general (axiomatic) theory of rational decision-making.

It will suffice for this commentary to consider the Centipede game (Colman's Fig. 5). This is a good exemplar of a social/interaction game without communication (except through the choices made), and with the goal for each player to maximize individual utility (not beat the other player). I assume that both players know that both players are rational, and not subject to the sundry "irrational" forces that lead human decision-makers to their choices. I also assume that each player knows his or her own mapping of monetary payoffs onto subjective utility, but does not know the mapping for the other player, other than the shared knowledge that a larger payoff (in monetary amount, say) corresponds to a larger utility. Note that this assumption (in most cases) eliminates the possibility that a rational strategy will involve a probabilistic mixture. Assuming that player A's mixture of choices affects player B's mixture of outcomes, player A generally cannot know whether the utility to B of a given mixture exceeds that for some other fixed or mixed payoff.

Therefore, the players at the outset of a game will both consider the same finite set of strategies S_j, where a given strategy consists of the ordered set of decisions $\langle D(1_A), D(2_B), D(3_A), D(4_B), \ldots D(N)\rangle$, where $D(I)$ is one of the choices allowed that player by the sequence of previous choices in that strategy. A game utility U_j is associated with each strategy: $\langle U_{jA}, U_{jB}\rangle$. Each player's goal is to reach a strategy that will maximize his or her personal U_j, in the knowledge that both players are rational and both have this goal.

In a Centipede game with many trials (say, 20), backward induction seems to lead to the "irrational" decision to stop (defect) on trial 1, even though both players can gain lots of money by playing (cooperating) for many trials. Of course, backward induction is flawed when used here in the usual way: Player A defects on, say, trial 15 in the certainty that Player B will defect on trial 16. But trial 15 could not have been reached unless B had been cooperating on all previous choices, so certainty is not possible. Thus, by cooperating on the first trial, the player eliminates backward induction as a basis for reasoning, and allows cooperation to emerge as a rational strategy. Yet, the forces in favor of defecting grow over trials, until backward induction seems to regain its force on the penultimate choice (e.g., trial 19 of 20, or 3 of 4).

Consider, therefore, a two-trial version of Colman's Centipede game. Both players at the outset consider the three possible strategies: ⟨stop⟩, ⟨play, stop⟩, ⟨play, play⟩, with associated payoffs of ⟨0,0⟩, ⟨−1,10⟩, ⟨9, 9⟩. The players look for a rational solution, in the hope that one exists (they share the knowledge that some games may not have a rational solution). So each player reasons: Which of the three strategies could be rational? Player B might like ⟨play, stop⟩, but both players could not decide this strategy was rational. If it were, A would stop on trial 1 (forcing a better outcome). Therefore, both players know ⟨play, stop⟩ could not be a rational strategy. Of the two remaining strategies, both players have little trouble seeing ⟨play, play⟩ as the rational choice, given that ⟨9, 9⟩ is preferred to ⟨0,0⟩.

This solution is "selfish," not predicated on maximizing joint return. It derives from the shared knowledge of playing a two-trial social game: In a one-trial game even a rational, cooperative decision-maker would clearly defect. Rationality is defined in terms of the entire game and total payoffs, not the payoff on a given trial. This approach could perhaps be considered a kind of generalization of the "Stackelberg reasoning" discussed by Colman, but is even more closely related to "dependency equilibria" discussed by Spohn (2001). It can be generalized and formalized (though not in this commentary). I note only that it gives justification for cooperative choices in simultaneous-choice games, such as the Prisoner's Dilemma (and sequential-play extensions of those games).

Perhaps the chief objection to this approach involves the perception that accepted causal precepts are violated: What is to stop B from defecting once trial 2 is reached? This issue is reminiscent of that obtaining in Newcomb's paradox (Nozick 1969), or the "toxin" puzzle (Kavka 1983), but in those cases a defense of a seemingly irrational later choice depends on uncertainty concerning an earlier causal event (I say "seemingly" because I am quite certain a Newcomb's chooser should take "one" and the "toxin" should be imbibed). The present case is more troublesome, because the first choice is known when the last choice is made. I nonetheless defend cooperation with the primary argument that rationality ought to be, and in fact must be, defined in terms of the entire game and not an individual decision within that game.

"Was you ever bit by a dead bee?" – Evolutionary games and dominated strategies

Karl Sigmund

Institut für Mathematik, Universität Wien, 1090 Vienna, Austria.
karl.sigmund@univie.ac.at **http://mailbox.univie.ac.at/karl.sigmund/**

Abstract: On top of the puzzles mentioned by Colman comes the puzzle of why rationality has bewitched classical game theory for so long. Not the smallest merit of evolutionary game theory is that it views rationality as a limiting case, at best. But some problems only become more pressing.

Aficionados of Humphrey Bogart will recognize this title's question as being a running gag from the film "To Have and Have Not." Apparently, if you step barefoot on a dead bee, you are likely to get hurt. The assumption that human behavior is rational died a long time ago, for reasons Colman summarizes very well, but it has failed to be buried properly. And if you carelessly tread on it, you will learn about its sting.

The question is, of course, why one should tread on it in the first place. There seems no reason ever to come close. The hypothesis that humans act rationally has been empirically refuted not only in the context of interactive decisions, but also for individual decision-making, where, in a way, it is even more striking. Indeed,

some of the observed interactive behavior can be explained in rational terms, if the utility function is modified by a term depending on the payoff difference between the player and the coplayers (see Fehr & Schmidt 1999). But this device, a "fix" that resembles the modifications of epicycles in the Ptolemaic model of celestial mechanics, cannot explain deviations from rationality in individual decision-making as evidenced, for instance, by the paradoxes of Allais (see, e.g., Allais & Hagen 1979) or Ellsberg (1961).

The founding fathers of game theory had little knowledge of such experiments. But it seems difficult to understand why, to our day, after all the work by Tversky, Kahnemann (see e.g., Kahnemann & Tversky 1979), and many others, full rationality can still be termed "not such a bad assumption." Not every scientific idealization deserves as much respect as that of a perfect gas! Clinging to human rationality, in the face of evidence, must be a way of protecting faith in the existence of a "unique rational solution" for every game – a supernatural claim.

Game theory is the conceptual tool for analyzing social interactions in terms of methodological individualism. That it should be used in any normative sense smacks of dogmatism. Game theory is a branch of mathematics, and in this sense is not more "normative" or "descriptive" than algebra. It helps to analyze the logical consequences of certain assumptions. The assumption of fully rational agents is just one of many alternatives. Its prominent role is caused by force of habit alone. Almost two hundred years ago, mathematicians rejected the creed in a unique set of geometrical axioms. Why should there be a unique set of postulates for game theory?

The rationality axiom is obviously not needed in game-theoretical analyses dealing with the chemical warfare between bacterial mutants, the mating behavior of male lizards, or the economical solidarity between students (see Fehr & Gächter 2000; Kerr et al. 2002; Sinervo & Lively 1996). Even the term "bounded rationality" seems ill-advised in such contexts, implying to lay-persons that rationality is the norm that bacteria, lizards, and undergraduates fail to achieve.

In applications to real-life situations (as opposed to philosophical puzzles), game theory can do just as well without the postulate of rationality, and Occam's razor demands, therefore, to get rid of it. That it held out for so long is mostly due to historical contingency.

An illustration of historical contingency at work is the fact that John Nash, in his Ph.D. thesis, explicitly stated that his equilibrium notion could be motivated, not only by an appeal to rational players, but also by what he called the "mass action" approach. Oddly, this section was deleted in the published version from 1950 (see Weibull 1995). It seems that a reviewer had discarded it. Nash's mass action approach was resuscitated decades later in evolutionary game theory: Thinking in terms of populations came naturally to evolutionary biologists. No longer do the players have to be rational; all they need is some propensity for adopting successful strategies. This can be due to learning, to imitation, to infection, or to inheritance (see, e.g., Gintis 2000; Hofbauer & Sigmund 1998; Weibull 1995).

But, and here comes the sting, getting rid of the rationality axiom as a foundational postulate does not get rid of all problems. Evolutionary games lead, in many cases, back to the puzzles described by Colman. It only places them in the context of natural science. Whenever successful strategies spread, dominated strategies will get eliminated, defection will evolve in the Prisoner's Dilemma game, and selfishness will be just as self-defeating as it is between rational players bent on out-smarting their equally rational coplayers.

This is the dead bee's revenge. Far from explaining it away, evolutionary game theory emphasizes the urgency of the paradox. There are societies out there – not only in philosophical mind games – that display cooperation, although it is a dominated strategy. Opting for the evolutionary approach is beneficial, nevertheless, because it opens up so many testable solutions to the puzzles. Consider, for example, the Ultimatum game. Here an experimenter offers ten dollars to a pair of test persons, provided they keep to the following rules: A toss of the coin decides who of the two is the "Proposer" and must decide which part of the ten dollars to offer to the coplayer. If the "Responder" accepts the offer, this is how the money is split between the two players. If the "Responder" rejects the offer, the experimenter pockets the money. In each case, the game is over, and all go their separate ways – no haggling, and no further rounds.

In real experiments, small offers get rejected by most Responders, and most Proposers offer a substantial share. This blatantly contradicts the usual rationality assumptions, whereby Proposers ought to offer the minimal amount and Responders ought to accept it. Numerical simulations of evolving populations of players yield the same prediction. But, whereas the rationality axiom just leads to an impasse, the evolutionary approach suggests ways out. If one assumes, for example, (a) that players usually interact only within their neighborhood (rather than with a randomly chosen member of a large, well-mixed crowd); or (b) that there is always some small percentage of players who would never offer, as Proposers, less than they would accept as Responders; or (c) that players occasionally offer less if they learn, somehow, that their coplayer is likely to swallow it; then offers coming close to reality will evolve (see Nowak et al. 2000; Page & Nowak 2002; Page et al. 2002). None of these three hypotheses need be right; but all allow for testable predictions. Game theory is not only a tool for philosophical debates, but – rid of the straitjacket of rationality – it is an instrument for every social science.

Irrationality, suboptimality, and the evolutionary context

Mark Steer and Innes Cuthill

School of Biological Sciences, University of Bristol, Bristol, BS8 1UG, United Kingdom. **mark.steer@bristol.ac.uk** **i.cuthill@bristol.ac.uk**

Abstract: We propose that a direct analogy can be made between optimal behaviour in animals and rational behaviour in humans, and that lessons learned by the study of the former can be applied to the latter. Furthermore, we suggest that, to understand human decisions, rationality must be considered within an evolutionary framework.

We believe that Colman raises valuable and interesting points about the nature of rational choice in humans. Nonetheless, we would like to make the important point that behaviour considered to be irrational within the confines of an experimental situation may nonetheless be rational within a wider context. We believe there are illuminating parallels between the study of the adaptive value of behaviour (in terms of individual optimality or evolutionary stability) and that of rationality in decision-making. Just as a rational decision is one that maximizes some measure of utility, so to a behavioural ecologist, an optimal decision is one that maximizes Darwinian fitness given certain constraints. Thus, we believe that the appropriate research program to understand the rationality (or otherwise) of decision-making in humans should be analogous to that needed to understand the adaptive value of behaviour in the face of evolution by natural selection. These issues are of broad concern, not just confined to game-theoretic situations.

Imagine, for example, an investigation into the foraging behaviour of a bird in an aviary. It has a choice between foraging in two locations. At location A, situated deep in a bush, the bird experiences a low rate of food intake; at the more open location B, the rate of intake is much higher. Contrary to the predictions of a simple model of energetic intake maximization, the bird prefers to feed at A. Why?

Although the experimenters appreciate that the bird is in no danger of predation, it doesn't necessarily follow that the bird does. Foraging in the open may be deemed too risky, even though

the rate of gain is higher. It is only when we start to take these considerations into account, and include them in the model, that the bird's behaviour begins to make sense (see Houston & McNamara 1989; 1999; McNamara 1996 for further discussion). Of course, we must test our new assumptions independently before accepting our revised hypothesis.

So is the bird making optimal, or suboptimal, foraging decisions? This depends on what information we can expect the bird to have about its current situation. If it can perceive that it is in no danger of predation, then to carry on exhibiting antipredator behaviour may indeed be considered suboptimal. However, if the bird is incapable of perceiving the decreased threat, either through a lack of the cues it would use in nature or an inflexible decision rule (that nonetheless works well under natural conditions), then the behaviour may be optimal within the context for which the rule evolved. The information that the bird bases its decisions upon depends not only on its perception of the current environment, but also on its own developmental and evolutionary history. Indeed, to an evolutionary game theorist, it is appropriate to ask why a decision rule that may be susceptible to developmental experience (e.g., learning) has evolved as opposed to a relatively inflexible, innate, rule.

We suggest that a direct comparison can be made between our investigation of the apparent suboptimality of the bird foraging under cover and how one should investigate the apparent irrationality of, for example, an altruistic person in a one-shot game. Although the behaviour of the person in question might seem to be irrational within the confines of the experimental set-up, we have to consider not only whether the person perceives (or even *can* perceive) the game in the way the experimenters have conceived it, but also whether it makes evolutionary sense for that person to do so.

We would go further than simply drawing analogies between the study of optimality in animal decision-making and of rationality in human behaviour. We believe that a full understanding of human decision-making, even if the rules are the result of conscious reasoning rather than inflexible preprogrammed strategies, requires an evolutionary perspective. As behavioural ecologists, our belief is that natural selection should have equipped humans, just as other animals, with rules that maximize fitness in the appropriate environment. Therefore, from our point of view, the most rational decision for a human to make should be that which maximizes fitness. If indeed human decisions are constrained in some way by evolutionary influences, then considering our evolutionary past could be instrumental in understanding why seemingly irrational decisions are made in certain circumstances.

Not only might human behaviour be less flexible than we imagine, the subject may perceive that the game is being played with a wider range of people than the experimenter has planned: the experimenters, other players, and observers. For example, individuals behave differently if they are informed that they are playing another person, than if they are told they are playing against a computer (Baker & Rachlin 2002).

So we return to our main point. Whether our bird is behaving optimally or not depends to some degree on whether it is reasonable to expect that bird to know that predation risk is zero. Similarly, whether a person's choices in a game can be considered (ir)rational depends on whether we expect that person to have understood the precise experimental paradigm and then acted accordingly. The importance of looking at the wider context when considering responses to games is paramount to understanding how people might react. When individuals make seemingly irrational choices in a predictable fashion, there are three possibilities among which we need to distinguish: First, subjects don't possess the capabilities to fully assess the conditions of the situation, precipitating fully rational behaviour as far as the subject is concerned. Or, subjects might fully understand the conditions of the game, however, because of a developmental or evolutionary hangup, they perform the (evolutionarily) right behaviour in the wrong context. The behaviour is thus locally irrational, but rational within a wider framework. Or, third, all the conditions of the game are well understood by a subject, and the decisions truly are irrational (and thus maladaptive?).

Bridging psychology and game theory yields interdependence theory

Paul A. M. Van Lange and Marcello Gallucci

Department of Social Psychology, Free University, 1081 BT Amsterdam, The Netherlands. **pam.van.lange@psy.vu.nl** **m.gallucci@psy.vu.nl**

Abstract: This commentary focuses on the parts of psychological game theory dealing with preference, as illustrated by team reasoning, and supports the conclusion that these theoretical notions do not contribute above and beyond existing theory in understanding social interaction. In particular, psychology and games are already bridged by a comprehensive, formal, and inherently psychological theory, interdependence theory (Kelley & Thibaut 1978; Kelley et al. 2003), which has been demonstrated to account for a wide variety of social interaction phenomena.

Understanding social interaction phenomena is obviously a key issue in the social and behavioral sciences. It is, therefore, surprising that several important theories, such as rational choice theory, game theory, and complementary forms of decision theory, tend to focus on the individual level of behavior, as if there is no mutual dependence between individuals. In other words, these theories tend to neglect the conceptual importance of *interdependence* in understanding social interaction – the realization that important psychological processes, including deviations of rationality and self-interest, are ultimately rooted in the dimensions of interdependence underlying interaction situations.

In the target article, Colman proposes psychological game theory as an alternative to "orthodox theory" in accounting for deviations of rationality and self-interest, thereby extending orthodox theory in at least two important respects. First, it addresses *interpersonal psychology* by emphasizing the role of beliefs and expectations relevant to a situation and the actions and beliefs of the interaction partner. Second, it addresses a *collective level of analysis,* emphasizing, for example, the importance of collective preferences in settings of interdependence.

Although we are in agreement with both extensions, we do not think that these extensions make a very novel contribution to existing social psychological theory. Moreover, we share the critical view expressed by Colman that psychological game theory "amounts to nothing more than a collection of tentative and ad hoc suggestions for solving the heterogeneous problems that have been highlighted in earlier sections" (sect. 8). Notwithstanding their heuristic value, the suggestions would have been considerably less tentative and less ad hoc, if Colman had discussed psychological game theory in relation to existing social psychological theories. In particular, we argue that the parts of psychological game theory dealing with preference, illustrated mainly with team reasoning, are already well-understood in terms of interdependence theory (Kelley & Thibaut 1978; Kelley et al. 2003; Rusbult & Van Lange 2003), which in many ways can be conceptualized as an inherently psychological, yet formal, theory of social interaction. As we will see, after a brief discussion of interdependence theory, it is not clear whether the part of psychological game theory dealing with preference – a part that is very essential to game theory – contributes above and beyond interdependence theory in understanding social interaction.

Interdependence theory. Interdependence theory may be characterized by at least three qualities. First, using games and related conceptual tools, interdependence theory provides a taxonomy of interaction situations, which can be analyzed in terms of several dimensions, such as degree and mutuality of dependence, basis for dependence, corresponding versus conflicting interest, temporal structure, and information availability (Kelley et al. 2003;

Rusbult & Van Lange 2003). This taxonomy allows one to characterize interaction situations. For example, a social dilemma would be characterized as one involving relatively high levels of interdependence, based on unilateral actions of the partner, and also characterized by a fairly strong conflict of interest; and social dilemmas may differ in terms of temporal structure (e.g., single-trial vs. iterated) and information availability (e.g., complete or incomplete information regarding one another's preferences).

Second, interdependence theory assumes that the outcomes in any interaction situation ("the given preferences") may be psychologically *transformed* into a subjective situation representing effective preferences, which are assumed to guide behavior and ultimately social interaction. Examples of transformation rules are interaction goals such as enhancement of both one's own and other's outcomes (MaxJoint), equality in outcomes (MinDiff), other's outcomes (MaxOther), or relative advantage over other's outcomes (MaxRel). Although transformations may be a product of careful reasoning and thought, they may also occur in a fairly automatic manner, involving very little thought or deliberation. As such, transformations deviate not only from self-interest, but also from rationality, in that individuals are not assumed to obey criteria of strict rationality. More importantly, transformations are assumed to accompany cognitive and affective processes in guiding behavior and shaping interaction (see Kelley et al. 2003; Rusbult & Van Lange 2003).

Finally, interdependence theory focuses on both individual and collective levels of analyses, in that it explicitly seeks to understand *social interaction*, which is conceptualized as a product of two individuals (with their basic preferences and transformational tendencies) and the interaction situation. Social interactions are also assumed to shape relatively stable embodiments of transformations, at the intrapersonal level (i.e., dispositions such as social value orientation), at the relationship level (i.e., partner-specific orientations, such as commitment), and at the cultural level (i.e., broad rules for conduct, such as social norms; see Rusbult & Van Lange 2003; Van Lange et al. 1997).

Interdependence theory and psychological game theory. As noted earlier, the parts of psychological game theory dealing with preference seem to be well-captured by interdependence theory. Needless to say, the notion of transformation explicates deviations of rationality and self-interest, and it is a theory that focuses on both individual and collective levels of analysis. Moreover, although transformations are individual-level rules, they do have strong implications for the collective level of analysis. For example, transformations may be interrelated with group-based variables, such as group identification ("we-thinking"), group attachment, or feelings of group responsibility. A case in point is the demonstration that individuals with prosocial orientation define rationality at the collective level, not at the individual level, thus judging cooperation as more intelligent than noncooperation in social dilemmas (cf. goal-prescribes-rationality principle, Van Lange 2000). Also, interdependence theory emphasizes the conceptual importance of beliefs, expectations, and interpersonal trust. Following the seminal work of Kelley and Stahelski (1970), the transformations that people actually adopt are assumed to be strongly conditioned by trust, beliefs, and expectations regarding the transformations pursued by particular interaction partners.

Utility of a transformational analysis. We should also briefly comment on Colman's suggestion that a transformational analysis is not especially helpful in understanding the Hi-Lo Matching game. Let us analyze this particular interaction situation for five transformations. First, a transformational analysis indicates that the orientations toward enhancing one's own outcomes (individualism), joint outcomes (cooperation), and a partner's outcomes (altruism) prescribe matching, and more strongly so for Heads than for Tails. Second, mere enhancement of relative advantage (competition) and equality in outcomes (egalitarianism) are irrelevant in this particular situation, because all four cells present equal outcomes for self and other. Given that cooperation and individualism are prevalent orientations, and given that often cooperation is accompanied by egalitarianism (see Van Lange 1999), the transformational analysis indicates that most people will be oriented toward matching Heads (followed by matching Tails). Pure forms of competition or egalitarianism lead to indifference, which in fact may hinder effective coordination between two individuals.

Thus, a transformation analysis may very well account for the fact that people tend to be fairly good at coordinating in the Hi-Lo Matching game and related situations. At the same time, the transformational analysis suggests that one reason people may not be able to coordinate is that at least one individual is merely interested in outperforming the interaction partner (or, less likely, merely interested in obtaining equality in outcomes). More generally, a comprehensive transformation analysis (which includes not only cooperation, as discussed by Colman) helps us understand this specific situation, even though we agree with Colman's implicit assumption that a transformational analysis is typically more strongly relevant to motivational dilemmas, involving more pronounced conflicts among cooperation, equality, individualism, and competition.

Conclusion. Over the past 25 years, interdependence theory has inspired several programs of research in areas as diverse as relationships, norms and roles, interpersonal dispositions, social dilemmas, group decision-making, and negotiation. It is a comprehensive, logical, and psychological theory of social interaction, thereby, to some degree, using the language (and logic) of game theory. Our discussion indicates that the parts of psychological game theory dealing with preference (and illustrated by team reasoning) do not extend interdependence theory in terms of theoretical potential, logic (including parsimony), or psychological breadth. Perhaps the contribution of Colman's article is more strongly rooted in conceptualizing specific lines of reasoning, such as Stackelberg reasoning, and reasoning focusing on common beliefs and nonmonotonic reasoning, which tend to deviate from self-interest or rationality. We are most confident about one broad message that Colman's article shares with interdependence theory – that is, the conviction that "bridging" social psychology with game theory is essential to the further development of the science of social interaction. After all, one needs games ("the situational structure") and the psychology of two individuals ("the processes," i.e., transformations, along with cognition and affect) to understand social interaction.

Toward a cognitive game theory

Ivaylo Vlaev[a] and Nick Chater[b]

[a]*Department of Experimental Psychology, University of Oxford, Oxford, OX1 3UD, United Kingdom;* [b]*Institute for Applied Cognitive Science, Department of Psychology, University of Warwick, Coventry, CV4 7AL, United Kingdom.*
ivaylo.vlaev@psy.ox.ac.uk **nick.chater@warwick.ac.uk**

Abstract: We argue that solving the heterogeneous problems arising from the standard game theory requires looking both at reasoning heuristics, as in Colman's analysis, and at how people represent games and the quantities that define them.

Colman's elegant and persuasive article describes psychological game theory by introducing formal principles of reasoning, and focuses on several nonstandard reasoning processes (team reasoning, Stackelberg reasoning, and epistemic and nonmonotonic reasoning). The goal is to explain psychological phenomena in game-playing that orthodox game theory, and its conventional extensions, cannot explain. We argue that, in addition, a model is needed of how the economic agent perceives and mentally represents the game initially, before any (strategic) reasoning begins. For instance, the perceived utility of various outcomes might change depending on the previous games seen.

As an illustration of such a possible model, here we offer some initial results from a research program that aims to ground ac-

counts of rationality in general, and decision theory in particular, on the underlying cognitive mechanisms that produce the seemingly paradoxical behavior. Colman's sections 2 and 4 discuss the basic underlying assumptions of expected utility theory and game theory. Existing models of rational choice and interactive game-theoretic decision making typically assume that only the attributes of the game need be considered when reaching a decision; that is, these theories assume that the utility of a risky prospect or strategy is determined by the utility of the outcomes of the game, and transforms the probabilities of each outcome. Decisions are assumed to be based on these utilities.

Our results demonstrate, however, that the attributes of the previously seen prospects and games influence the decisions in the current prospect and game, which suggests that prospects and games are not considered independently of the previously played ones (Stewart et al., in press; Vlaev & Chater, submitted). In particular, Stewart et al. (in press) have revealed the phenomena of "prospect relativity": that the perceived value of a risky prospect (e.g., "p chance of x") is relative to other prospects with which it is presented. This is counter to utility theory, according to which the perceived value of each prospect should be dependent only on its own attributes. Stewart et al. suggest that this phenomenon arises in the representation of the magnitudes that define the prospects, and suggest that the phenomenon has a common origin with related effects in the perception of sensory magnitudes (Garner 1954; Laming 1997; Lockhead 1995).

We have found similar effects, providing a new type of anomaly for orthodox game theory. People play repeated one-shot Prisoner's Dilemma games (Vlaev & Chater, submitted). The degree to which people cooperate in these games is well-predicted by a function of the pay-offs in the game, the cooperation index as proposed in Rapoport & Chammah (1965). Participants were asked on each round of the game to predict the likelihood that their coplayer will cooperate, and then to make a decision as to whether to cooperate or defect. The results demonstrated that the average cooperation rate and the mean predicted cooperation of the coplayer in each game strongly depend on the cooperativeness of the preceding games, and specifically on how far the current game was from the end-points of the range of values of the cooperation index in each session. Thus, in games with identical cooperation indices, people cooperated more and expected more cooperation in the game with higher rank position relative to the other cooperation index values in the sequence. These findings present another challenge to the standard rational choice theory and game theory, as descriptive theories of decision-making under uncertainty, and also to other theories where games are independently considered.

Our proposed account for these results, and also for other problems related to the independence assumption, is that people have poor notions of absolute cooperativeness, risk, and utility, and instead make their judgments and decisions in relative terms, as is described in some existing psychophysical and cognitive theories of perception and judgment of information about magnitudes (intensities of stimulus attributes). Thus, this account departs fundamentally from previous work in this field, by modeling the highly flexible and contextually variable way in which people represent magnitudes, like sums of money, probabilities, time intervals, cooperativeness, and so forth, rather than by assuming that these can be represented on absolute internal psychological scales (i.e., even if these scales exist, they stretch or contract depending on the other stimuli in the environment). We conjecture that the results from the two studies presented here suggest that people use context as a sole determinant of the utility of a strategy, which is a form of a more ecologically adaptive rationality, and therefore any descriptive account of game-theoretic behavior, especially in sequential social dilemmas, should incorporate a model of agents' lower-level cognitive perceptual processes.

This discussion does not answer the paradoxes posed in the target article, but here we would like to make the stronger claim that there are many more phenomena that the standard approach cannot explain (and there will be more to be discovered), and that in order to develop a decent account of human decision behavior in games, a much more radical approach is needed. Our results imply that Colman's version of psychological game theory, as based only on nonstandard forms of reasoning, needs to be supplemented by a more general "cognitive game theory," which grounds decision-making in the underlying cognitive mechanisms that produce the decision behavior. Such a cognitive approach could also include collective rationality criteria (which, as Colman states, are lacking in the standard decision theory), because, for example, categorization of the coplayer as being very similar to me could strongly affect my common belief in each other's rationality, or at least in the similarity of the reasoning processes that we would employ. Also, the perception of the players and the game as being similar to a previous interactive situation, in which the coplayers acted in a certain way (e.g., chose a certain focal point), would enforce the belief that, in the current situation, the coplayers would act in a similar way.

From rationality to coordination

Paul Weirich

Philosophy Department, University of Missouri, Columbia, MO 65211.
weirichp@missouri.edu
http://web.missouri.edu/~philwww/people/weirich.html

Abstract: Game theory's paradoxes stimulate the study of rationality. Sometimes they motivate the revising of standard principles of rationality. Other times they call for revising applications of those principles or introducing supplementary principles of rationality. I maintain that rationality adjusts its demands to circumstances, and in ideal games of coordination it yields a payoff-dominant equilibrium.

Game theory raises many puzzles about rationality, which is why it is so fascinating to philosophers. Responding to the puzzles is a good way of learning about rationality. Colman insightfully reviews many of game theory's paradoxes and uses them to argue that, in games, people follow nonstandard principles of reasoning. He does not, however, claim that those nonstandard principles have normative force. Do principles of rationality need revision in light of the paradoxes of game theory? Rationality is a rich topic, and familiar principles are not likely to capture all its nuances. Nonetheless, standard principles of rationality are very versatile. In this commentary, I make a few general points about rationality and then show that extended applications of the standard principles resolve some paradoxes.

A standard principle of rationality is to maximize utility. The literature advances several interpretations of this principle. Nearly all acknowledge that impediments may provide good excuses for failing to meet it; the principle governs ideal cases. Game theory presents decision problems that are non-ideal in various ways. Perhaps in games some failures to maximize utility are excused. Rationality may lower standards in difficult cases.

In *Equilibrium and Rationality* (Weirich 1998), I generalize the principle of utility maximization to accommodate non-ideal cases in which no option maximizes utility. I assume that standards of rationality adjust to an agent's circumstances, so that in every decision problem some option is rational. The generalization of the decision principle leads to a generalization of Nash equilibrium that makes equilibrium exist more widely. A principle Colman calls "rational determinacy" supports the generalizations. The version I endorse asserts that rationality is attainable, but not that rationality is attainable in one way only. It allows for multiple solutions to a game. Associated with the principle of rational determinacy is the view that achieving an equilibrium is just one requirement of rationality, and meeting it is not sufficient for full rationality. Principles of rationality govern equilibrium selection also. I ascribe to the principle of equilibrium selection called pay-

off dominance. What principles of individual rationality support it?

Colman applies principles of team reasoning and Stackelberg reasoning to coordination games such as the Hi-Lo Matching game, where communication is impossible, and, as a result, neither agent can influence the other's strategy. These principles are replacements for utility maximization. Team reasoning replaces individual goals with team goals. Stackelberg reasoning replaces maximization of good results with maximization of good news. Even if these novel principles of reasoning have descriptive value, neither is promising as an account of rationality. Team reasoning conflicts with individualism, and Stackelberg reasoning conflicts with consequentialism.

Stackelberg reasoning uses evidence that a choice furnishes to help evaluate the choice. Its approach to equilibrium selection resembles the grounding of equilibrium in a strategy's self-ratification (see Harper 1991; 1999; Jeffrey 1983; McClennen 1992; Shin 1991; Weirich 1988; 1994). However, it does not discriminate between a strategy's causal and evidential consequences. As causal decision theory points out, rational decision-making attends to causal consequences exclusively (see Gibbard 1992; Joyce 1999, Ch. 5; and Weirich 2001, sect. 4.2).

Principles of rationality regulate the circumstances of decision problems and not just the choices made in the decision problems. Their influence on those circumstances affects the solutions to the decision problems. Principles of rationality govern belief, for example, and through belief's control of action, influence rational choices. In particular, precedence shapes rational beliefs in ways that sustain conventions of coordination. Similarly, principles of rationality regulate an agent's circumstances in Colman's ideal version of Hi-Lo so that utility maximization leads to the superior form of coordination.

In Hi-Lo, the players' evidence about each other influences their beliefs about the strategies they will adopt. As Colman describes the players, each chooses *H* only if the other does, and neither gets the ball rolling. However, a rational agent who foresees the possibility of playing Hi-Lo prepares for the game. He inculcates a disposition to choose *H* and lets others know about his disposition. Although he cannot influence his counterpart's strategy during the game, prior to the game he sets the stage for an optimal outcome. When he enters the game, he has already influenced his counterpart's beliefs about his choice. Knowing about the disposition he has acquired, his counterpart believes he will choose *H* and so maximizes utility by choosing *H* also.

Acquiring the disposition to choose *H* is rational. It leads to the superior form of coordination. Choosing *H* is rational also. It maximizes utility given the belief that one's counterpart will choose *H*. Both agents in Hi-Lo may rationally acquire the disposition and make the choice. Neither one's rational steps undercut the reasons for the other's.

When communication is possible, the goal of optimization requires an agent in Hi-Lo to plump for *H* and advertise his choice. Plumping for *H* is rational because it elicits *H* from his counterpart. Acquiring a disposition to pick *H* is rational for the same reason when communication is not possible. Having a disposition to pick *H* elicits *H* from one's counterpart. It does not matter that the disposition has bad consequences if one's counterpart were to choose *L*. That possibility is not realized. The disposition is rational because of its good consequences, even if it entails a disposition to make an irrational choice in a counterfactual situation.

When one has an opportunity to prepare for Hi-Lo, it is rational to put oneself into a decision situation such that a strategy of maximum utility has a utility at least as great as a strategy of maximum utility in any other decision situation into which one might put oneself. Rational decision preparation yields a decision problem in which rationality prospers.

A rational agent providently handles events prior to a game. She prepares for the game in ways that improve its likely outcome. Cultivating dispositions to choose certain ways is sensible preparation for coordination problems. Enriching an account of rationality to cover such preparation helps explain successful coordination without drastic revision of the principles for reasoning in games.

Author's Response

Beyond rationality: Rigor without mortis in game theory

Andrew M. Colman

School of Psychology, University of Leicester, Leicester LE1 7RH, United Kingdom. **amc@le.ac.uk** **www.le.ac.uk/home/amc**

Abstract: Psychological game theory encompasses formal theories designed to remedy game-theoretic indeterminacy and to predict strategic interaction more accurately. Its theoretical plurality entails second-order indeterminacy, but this seems unavoidable. Orthodox game theory cannot solve payoff-dominance problems, and remedies based on interval-valued beliefs or payoff transformations are inadequate. Evolutionary game theory applies only to repeated interactions, and behavioral ecology is powerless to explain cooperation between genetically unrelated strangers in isolated interactions. Punishment of defectors elucidates cooperation in social dilemmas but leaves punishing behavior unexplained. Team reasoning solves problems of coordination and cooperation, but aggregation of individual preferences is problematic.

R1. Introductory remarks

I am grateful to commentators for their thoughtful and often challenging contributions to this debate. The commentaries come from eight different countries and an unusually wide range of disciplines, including psychology, economics, philosophy, biology, psychiatry, anthropology, and mathematics. The interdisciplinary character of game theory and experimental games is illustrated in **Lazarus**'s tabulation of more than a dozen disciplines studying cooperation. The richness and fertility of game theory and experimental games owe much to the diversity of disciplines that have contributed to their development from their earliest days.

The primary goal of the target article is to argue that the standard interpretation of instrumental rationality as expected utility maximization generates problems and anomalies when applied to interactive decisions and fails to explain certain empirical evidence. A secondary goal is to outline some examples of *psychological game theory,* designed to solve these problems. **Camerer** suggests that *psychological* and *behavioral game theory* are virtually synonymous, and I agree that there is no pressing need to distinguish them. The examples of psychological game theory discussed in the target article use formal methods to model reasoning processes in order to explain powerful intuitions and empirical observations that orthodox theory fails to explain. The general aim is to broaden the scope and increase the explanatory power of game theory, retaining its rigor without being bound by its specific assumptions and constraints.

Rationality demands different standards in different do-

mains. For example, criteria for evaluating formal arguments and empirical evidence are different from standards of rational decision making (Manktelow & Over 1993; Nozick 1993). For rational decision making, expected utility maximization is an appealing principle but, even when it is combined with consistency requirements, it does not appear to provide complete and intuitively convincing prescriptions for rational conduct in all situations of strategic interdependence. This means that we must either accept that rationality is radically and permanently limited and riddled with holes, or try to plug the holes by discovering and testing novel principles.

In everyday life, and in experimental laboratories, when orthodox game theory offers no prescriptions for choice, people do not become transfixed like Buridan's ass. There are even circumstances in which people reliably solve problems of coordination and cooperation that are insoluble with the tools of orthodox game theory. From this we may infer that strategic interaction is governed by psychological game-theoretic principles that we can, in principle, discover and understand. These principles need to be made explicit and shown to meet minimal standards of coherence, both internally and in relation to other plausible standards of rational behavior. Wherever possible, we should test them experimentally.

In the paragraphs that follow, I focus chiefly on the most challenging and critical issues raised by commentators. I scrutinize the logic behind several attempts to show that the problems discussed in the target article are spurious or that they can be solved within the orthodox theoretical framework, and I accept criticisms that appear to be valid. The commentaries also contain many supportive and elaborative observations that speak for themselves and indicate broad agreement with many of the ideas expressed in the target article.

R2. Interval-valued rational beliefs

I am grateful to **Hausman** for introducing the important issue of rational beliefs into the debate. He argues that games can be satisfactorily understood without any new interpretation of rationality, and that the anomalies and problems that arise in interactive decisions can be eliminated by requiring players not only to choose rational strategies but also to hold rational beliefs. The only requirement is that subjective probabilities "must conform to the calculus of probabilities."

Rational beliefs play an important role in Bayesian decision theory. Kreps and Wilson (1982b) incorporated them into a refinement of Nash equilibrium that they called *perfect Bayesian equilibrium,* defining game-theoretic equilibrium for the first time in terms of strategies *and beliefs.* In perfect Bayesian equilibrium, strategies are best replies to one another, as in standard Nash equilibrium, and beliefs are *sequentially rational* in the sense of specifying actions that are optimal for the players, given those beliefs. Kreps and Wilson defined these notions precisely using the conceptual apparatus of Bayesian decision theory, including belief updating according to Bayes' rule. These ideas prepared the ground for theories of *rationalizability* (Bernheim 1984; Pearce 1984), discussed briefly in section 6.5 of the target article, and the psychological games of Geanakoplos et al. (1989), to which I shall return in section R7 below.

Hausman invokes rational beliefs in a plausible – though I believe ultimately unsuccessful – attempt to solve the payoff-dominance problem illustrated in the Hi-Lo Matching game (Fig. 2 in the target article). He acknowledges that a player cannot justify choosing *H* by assigning particular probabilities to the co-player's actions, because this leads to a contradiction (as explained in sect. 5.6 of the target article).[1] He therefore offers the following suggestion: "If one does not require that the players have point priors, then Player I can believe that the probability that Player II will play *H* is not less than one-half, and also believe that Player II believes the same of Player I. Player I can then reason that Player II will definitely play *H,* update his or her subjective probability accordingly, and play *H."*

This involves the use of interval-valued (or set-valued) probabilities, tending to undermine **Hausman**'s claim that it "does not need a new theory of rationality." Interval-valued probabilities have been axiomatized and studied (Kyburg 1987; Levi 1986; Snow 1994; Walley 1991), but they are problematic, partly because *stochastic independence,* on which the whole edifice of probability theory is built, cannot be satisfactorily defined for them, and partly because technical problems arise when Bayesian updating is applied to interval-valued priors. Leaving these problems aside, the proposed solution cleverly eliminates the contradiction that arises when a player starts by specifying a point probability, such as one-half, that the co-player will choose *H,* and ends up deducing that the probability is in fact unity. Because "not less than one-half" includes both one-half and unity, the initial belief is not contradicted but merely refined from a vague belief to a certainty.

This is not strictly Bayesian updating, because it is driven by deduction rather than empirical data, but it is unnecessary to pursue that problem. More importantly, what *reason* does a player have for believing that the probability is not less than one-half that the co-player will choose *H*? The *HH* equilibrium is highly salient by virtue of being payoff-dominant, but Gilbert (1989b) showed that this does *not* imply that we should expect our co-players to choose *H,* because *mere salience does not provide rational agents with a reason for action* (see sect. 5.5 of the target article). As far as I know, this important conclusion has never been challenged.

The proposed solution begins to look less persuasive when we realize that there are other interval-valued beliefs that do the trick equally well. If each player believes that the probability is *not less than three-quarters* that the co-player will play *H,* then once again these beliefs can be refined, without contradiction, into certainties. This suggests that *not less than one-half* is an arbitrary choice from an infinite set of interval-valued priors.

In fact, **Hausman** need not have handicapped himself with his controversial and decidedly nonstandard interval-valued probabilities. He could merely have required each player to believe from the start, *with certainty,* that the co-player will choose *H.* That, too, would have escaped the contradiction, but it would also have exposed a question-begging feature of the solution.

This leads me to the most serious objection, namely, that the proposed solution does not actually deliver the intuitively obvious payoff-dominant solution. It gives no obvious reason why we should not require each player to believe that the probability is not less than one-half that *the co-player will choose L.* If these beliefs are refined into cer-

tainties, then the players choose the Pareto-inefficient *LL* equilibrium. In other words, a belief that *the co-player will choose L* becomes self-confirming provided only that both players adopt it, in exactly the same way that a belief that *the co-player will choose H* does, although these two beliefs are mutually exclusive. This is a variation of a well-known problem in rational expectations theory (Elster 1989, pp. 13–15).

The point of section 5.6 of the target article is to argue that orthodox game theory fails to justify or explain the intuitively obvious payoff-dominant *HH* solution. **Hausman**'s suggestion falls short of being a complete solution because of technical problems with interval-valued beliefs, and because it seems, on examination, to have other shortcomings. Nevertheless, it is the most resourceful and challenging attempt among all the commentaries to solve a problem discussed in the target article without recourse to psychological game theory.

R2.1. Are social dilemmas paradoxical?

I feel impelled to comment on the following assertion by **Hausman** about the single-shot Prisoner's Dilemma game (PDG) shown in Figure 4 of the target article: "Although rationality is indeed *collectively* self-defeating in a PDG, there is no paradox or problem with the theory of rationality" (emphasis in original). It was Parfit (1979) who first described rationality as *self-defeating* in the PDG. It is true that he claimed it to be collectively and not individually self-defeating, but he did not mean to imply that it embodied no paradox or problem of rationality. If both players choose strategically dominant and hence rational *D* strategies, then the *collective* payoff to the pair (the sum of their individual payoffs) is less than if they both choose cooperative *C* strategies. If the dilemma amounted to nothing more than that, then I would agree that "there is no paradox or problem."

But the PDG places a player in a far deeper and more frustrating quandary. Each player receives a *better individual* payoff from choosing *D* than from choosing *C*, whatever the co-player chooses, yet if both players choose *D,* then each receives a *worse individual* payoff than if both choose *C.* That is what makes the PDG paradoxical and causes rationality to be self-defeating. I discuss this in section 6.5 of the target article and point out in sections 6.9 and 6.11 that the same paradox haunts players in multi-player social dilemmas.

I am not even sure that it is right to describe rationality in the PDG as *collectively* but not *individually* self-defeating. As **Krueger** reminds us, the logician Lewis claimed that the PDG and Newcomb's problem[2] are logically equivalent. Lewis (1979) was quite emphatic:

> Considered as puzzles about rationality, or disagreements between two conceptions thereof, they are one and the same problem. Prisoner's Dilemma *is* Newcomb's problem – or rather, two Newcomb's problems side by side, one per prisoner. (p. 235, emphasis in original)

This turned out to be controversial (Campbell & Sowden 1985, Part IV), and **Krueger**'s own comments show rather effectively that the correspondence is far from clear, but everyone agrees that the two problems are at least closely related. Nevertheless, in Newcomb's problem there is only one decision maker, and choosing the dominant (two-box) strategy must therefore be *individually* self-defeating in that case.

What is a *paradox*? The word comes from the Greek *paradoxos,* meaning beyond (*para*) belief (*doxa*). Quine (1962) defined it as an apparently valid argument yielding either a contradiction or a *prima facie* absurdity. He proposed a threefold classification into *veridical paradoxes,* whose conclusions are true; *falsidical paradoxes,* whose conclusions are false; and *antinomies,* whose conclusions are mutually contradictory. The PDG is obviously a veridical paradox, because what we can deduce about it is true but *prima facie* absurd. A classic example of a veridical paradox is Hempel's paradox,[3] and the PDG seems paradoxical in the same sense. Newcomb's problem, which is logically equivalent or at least closely related to the PDG, is indubitably paradoxical.

R3. Payoff-transformational approaches

Van Lange & Gallucci are clearly underwhelmed by the solutions outlined in the target article. They "do not think that these extensions make a very novel contribution to existing social psychological theory." Social psychology is notoriously faddish, but surely what is important is how well the extensions solve the problems in hand, not how novel they are. They should be judged against competing theories, and Van Lange & Gallucci helpfully spell out their preferred solution to one of the key problems. They tackle the payoff-dominance problem, arguing that *interdependence theory,* with its payoff transformations, provides a complete solution. If they are right, then this simple solution has been overlooked by generations of game theorists, and by the other commentators on the target article; but I believe that they misunderstand the problem.

Van Lange & Gallucci's discussion focuses on the Hi-Lo Matching game shown in Figure 2 of the target article. They assert that maximization of individual payoffs (*individualism*), joint payoffs (*cooperation*), and co-player's payoffs (*altruism*) all lead to successful coordination on the payoff-dominant *HH* equilibrium (I have substituted the usual term "payoffs" where they write "outcomes," because an outcome is merely a profile of strategies). They then claim: "Given that cooperation and individualism are prevalent orientations, . . . the transformational analysis indicates that most people will be oriented toward matching *H* (followed by matching *L*)" (here I have corrected a slip in the labeling of strategies, replacing Heads and Tails with *H* and *L*). They believe that this "may very well account for the fact that people tend to be fairly good at coordinating in the Hi-Lo Matching game."

The *individualism* transformation is no transformation at all: it is simply maximization of individual payoffs. With the specified payoffs, the players have *no reason* to choose *H* (see sect. 5.6 of the target article). The *cooperation* transformation fails for the same reason, merely producing the bloated Hi-Lo Matching game shown in Figure 6. Although I do not discuss the *altruism* transformation in the target article, it fares no better. A simple proof is given in an endnote.[4]

Van Lange & Gallucci labor to show that the players prefer the *HH* outcome in the Hi-Lo Matching game under certain payoff transformations. But we do not need pay-

off transformations to tell us that – it is obvious by inspection of Figure 2. The problem is that, *in spite of their obvious preference for HH*, the players have no *reason* to choose the strategy *H*. "Wishes can never fill a sack," according to an Italian proverb, and that is why Harsanyi and Selten (1988) had to introduce the payoff-dominance principle as an axiom in their equilibrium selection theory. We need to explain how human players solve such games with ease. The fact that "people will be oriented toward matching *H*" does not magically entail that this "may very well account for the fact that people tend to be fairly good at coordinating in the Hi-Lo Matching game."

Other commentators remark that individual preferences do not automatically guarantee coordination on payoff-dominant outcomes. For example, **Hurley** comments: "In Hi-Lo, individuals have the same goals, yet individual rationality fails to guarantee them the best available outcome." As **Haller** puts it: "principles other than individual rationality have to be invoked for equilibrium selection."

Barclay & Daly share the opinion of **Van Lange & Gallucci** that "tinkering with utility functions" is all that is needed to solve the game, but they do not attempt to show how this can be done, so there can be no reasoned reply. Payoff transformations are potentially useful for psychological game theory, notably in Rabin's (1993) "fairness equilibria," discussed by **Carpenter & Matthews**, **Camerer**, and **Haller** (in passing), but they cannot solve the payoff-dominance problem, although it would be pleasant indeed if such a simple solution were at hand.

Team reasoning and Stackelberg reasoning, the two suggestions in the target article, both solve the problem but require nonstandard auxiliary assumptions. **Alvard** reminds us that cultural mechanisms solve cooperative problems so transparently that many do not recognize them as solutions at all. This brings to mind Heider's (1958) comment: "The veil of obviousness that makes so many insights of intuitive psychology invisible to our scientific eye has to be pierced" (p. 322).

R4. Is rationality dead?

Writing from the standpoint of evolutionary game theory (see sect. R5 below), **Sigmund** puts forward the radically dismissive view that rationality is dead: "The assumption that human behavior is rational died a long time ago. . . . The hypothesis that humans act rationally has been empirically refuted. . . . Even the term 'bounded rationality' seems ill-advised." Hofbauer and Sigmund (1998), in true evolutionary spirit, explained the development of their view of rationality in their superb monograph on evolutionary games:

> The fictitious species of rational players reached a slippery slope when the so-called "trembling hand" doctrine became common practice among game theorists . . . and once the word of "bounded rationality" went the round, the mystique of rationality collapsed. (p. xiv)

This invites the following question. Does **Sigmund** expect his readers to be persuaded that rationality is dead? If he rejects rationality in all its forms, then he can hardly claim that his own opinions are rationally based, and there is consequently no obvious reason why we should be persuaded by them. By his own account, his comments must have arisen from a mindless evolutionary process unrelated to truth. This view cannot be taken seriously, and it is debatable – though I shall resist the temptation to debate it – whether it is even possible for **Sigmund** to believe it.

It seems clear to me that people are instrumentally rational in the broad sense explained in section 2 of the target article. **Rapoport** agrees, and so do other commentators, explicitly and implicitly. All that this means is that human behavior is generally purposive or goal-directed. To deny this would be to deny that entrepreneurs try to generate profits; that election candidates try to maximize votes; that professional tennis players try to win matches; and that Al-Qaeda terrorists try to further their ideological goals. To deny human instrumental rationality is to deny that such activities are purposive.

The assumption of instrumental rationality has a privileged status because of its neutrality toward the ends that people seek. Whether people are motivated by a desire for money, status, spiritual fulfillment, altruistic or competitive objectives, devotion to family, vocation, or country, they are rational to the extent that they choose appropriate actions to promote their desires. **Barclay & Daly** appear to overlook this when they argue in favor of rejecting rationality even as a default assumption. They suggest that people may be driven by motives such as "concern for the welfare of others," and that this leads to decisions that are "not in accordance with predictions of RCT [rational choice theory]." But in RCT and game theory such motives are assumed to be fully reflected in the players' utility functions. Rationality is interpreted as behavior that optimally fulfils an agent's desires, *whatever* these may be.

We have to treat other people as broadly rational, for if they were not, then their reactions would be haphazard and unpredictable. We assume by default that others are rational. The following *Gedankenexperiment* illustrates this nicely (cf. Elster 1989, p. 28). Imagine a person who claimed to prefer *A* to *B* but then deliberately chose *B* when *A* was also available. We would not, in the absence of special circumstances, infer that the choice was irrational. We would normally infer that the person did not really prefer *A*, or perhaps that the choice of *B* was a slip or an error. This shows rather effectively that rationality is our default assumption about other people's behavior.

There are certainly circumstances in which people behave irrationally. Introspection, anecdotal evidence, and empirical research all contribute clear examples. I find the following introspective example, originally formulated by Sen (1985) and mentioned in **Rapoport**'s commentary, especially persuasive. A family doctor in a remote village has two patients, *S* and *T*, both critically ill with the same disease. A certain drug gives excellent results against the disease, but only one dose is available. The probability of success is 90 per cent for Patient *S* and 95 per cent for Patient *T*. To maximize expected utility (EU), the doctor should administer it to *T*. But there are many doctors who would prefer to toss a coin, to give *S* and *T* equal chances of receiving the drug, although this mixed strategy yields a lower EU. It is difficult not to empathize with a doctor who is reluctant to "play God" in this situation, although tossing a coin obviously violates the axioms of instrumental rationality.

Anecdotal examples abound. Behavior that ignores future consequences, such as the actions of a person descending into drug addiction, are obviously irrational. Em-

pirical research has focused on anomalies such as the Allais and Ellsberg paradoxes (see, e.g., Dawes 1988, Ch. 8). Each of these involves a pair of intuitively compelling choices that can be shown to be jointly incompatible with the axioms of expected utility theory. In addition, a great deal of empirical research has been devoted to heuristics and biases that deviate from rationality (Bell et al. 1988; Kahneman et al. 1982; Kahneman & Tversky 2000). When violations are pointed out to decision makers, they tend to adjust their behavior into line with rational principles, suggesting that people's choices are sometimes in conflict with their own normative intuitions (Tversky 1996). But what attracts attention to all these phenomena is precisely that they *are* deviations from rational behavior.

People evidently take no pride in their occasional or frequent lapses from rationality (Føllesdal 1982; Tversky 1996). Anecdotal and experimental evidence of irrationality does not alter the fact that people are generally rational. The fact that birds and bats and jumbo jets fly does not refute Newton's universal law of gravitation. By the same token, the fact that human decision makers deviate from rationality in certain situations does not refute the fundamental assumption of instrumental rationality.

Less often discussed than bounded rationality and irrationality is the fact that people are sometimes even more rational than orthodox game theory allows. In section 5 of the target article, I show that players frequently succeed in coordinating, to their mutual advantage, where game theory fails. In section 6, I show that players frequently cooperate in social dilemmas, thereby earning higher payoffs than conventionally rational players. In section 7, I show that players frequently ignore the logic of backward induction in sequential games, thereby outscoring players who follow game theory. These examples suggest that human players are, on occasion, *super-rational* inasmuch as they are *even more* successful at maximizing their expected utilities than orthodox game theory allows.

R5. Evolutionary games

I discuss evolutionary game theory briefly in section 1.3 of the target article, but several commentators (**Alvard**, **Barclay & Daly**, **Butler**, **Casebeer & Parco**, **Sigmund**, and **Steer & Cuthill**) take me to task for assigning too little importance to evolutionary and adaptive mechanisms. Evolutionary approaches are certainly fashionable, and I believe that they have much to offer. I have contributed modestly to the literature on evolutionary games myself. However, because the target article was devoted to examining *rationality* in strategic interaction, evolutionary games are only obliquely relevant.

Sigmund traces the origin of the evolutionary approach to a passage in John Nash's Ph.D. thesis. The passage is missing from the article that emerged from the thesis (Nash 1951), but the thesis has now been published in facsimile (Nash 2002), and my reading of the key passage suggests that Nash interpreted his computational approach as a method of approximating rational solutions by simulation, analogous to the Newton-Raphson iterative method for solving equations. He imagined a game repeated many times by players who "accumulate empirical information on the relative advantage of the various pure strategies at their disposal" (Nash 2002, p. 78) and choose best replies to the co-players' strategies. He showed how this adaptive learning mechanism causes the strategies to converge toward an equilibrium point.

Contemporary evolutionary game models, whether they involve adaptive learning processes (*à la* Nash) or replicator dynamics, are designed to explore the behavior of goal-directed automata. Either the automata adjust their strategies in response to the payoffs they receive in simulated interactions, or their relative frequencies in the population change in response to payoffs. In either case they are programmed to maximize payoffs, and in that limited sense they are instrumentally rational, even though their behavior is generated without conscious thought or deliberate choice, as **Barclay & Daly** and **Steer & Cuthill** correctly point out. One of the pioneers of genetic algorithms has gone so far as to claim that evolutionary models can be used "to explore the extent to which we can capture human rationality, both its limitations and its inductive capacities, in computationally defined adaptive agents" (Holland 1996, p. 281).

Gintis makes the important point that evolutionary game theory cannot solve all the problems of orthodox game theory, because it is relevant only to large populations and repeated interactions. It cannot solve the problems that arise in isolated interactions.

Indeed, evolutionary or adaptive mechanisms are a far cry from rational choice. Human decision makers can and do anticipate the future consequences of their actions, whereas genetic and other evolutionary algorithms are backward-looking, their actions being determined exclusively by past payoffs (plus a little randomness in stochastic models). They function by unthinking evolution, learning, and adaptation. This is not intended as a criticism – backward-looking nostalgia may not be as limiting as it appears to be. It is worth recalling that the behaviorist school of psychology also explained human and animal behavior by a backward-looking and unthinking adaptive mechanism, namely, reinforcement. Behaviorism had a dominant influence on psychology throughout the 1940s and 1950s and remains influential even today.

R5.1. Learning effects

Casebeer & Parco claim that an experiment on three-player Centipede games by Parco et al. (2002) directly contradicts both psychological and traditional game theory. Parco et al. found that play converged toward equilibrium over 60 repetitions of the game, especially when very large monetary incentives were assigned to the payoffs. These interesting learning and incentive effects contradict neither traditional game theory nor the nonstandard approaches that I tentatively discuss in section 8.4 of the target article, namely, epistemic and non-monotonic reasoning. They suggest to me that players gradually learn to understand backward induction, through the course of repetitions of the rather complex game, especially when much is at stake. Convergence toward equilibrium is characteristic of iterated games in general.

I agree with **Kokinov** that strategic decisions are often made by analogy with previous experiences and, in particular, that there are circumstances in which people tend to repeat strategies that were successful in the past and to avoid strategies that were unsuccessful. This is most likely to occur in repeated games of *incomplete information,* in

which players do not have enough information to select strategies by reasoning about the game. The most common form of incomplete information is uncertainty about the co-players' payoff functions. The mechanism that Kokinov proposes is an analogical version of a strategy for repeated games variously called *win-stay, lose-change* (Kelley et al. 1962); *simpleton* (Rapoport & Chammah 1965, pp. 73–74); or *Pavlov* (Kraines & Kraines 1995), and it is remarkably effective in some circumstances.

R6. Behavioral ecology

Turning now to behavioral ecology, I agree with **Alvard**, **Butler**, and **Lazarus** that human beings and their cognitive apparatus are products of natural selection, and that evolution may help to explain some of the problems discussed in the target article, although it may be over-ambitious to suggest, as **Alvard** does, that "many of the ad hoc principles of psychological game theory introduced at the end of the target article might be deductively generated from the principles of evolutionary theory."

Steer & Cuthill advocate a radically evolutionary interpretation. They believe that our most rational decisions are those that maximize Darwinian fitness – that is, our lifetime reproductive success, or the number of offspring that we produce. That is how rationality is implicitly defined in evolutionary game theory (see sect. R5 above), and in that context the interpretation works well enough. But maximizing offspring cannot be taken as the ultimate underlying motive of all human behavior, because it simply does not fit the facts. Most purposive actions are driven by motives far removed from reproduction, and there are common forms of purposive behavior, such as contraception and elective sterilization, that clearly diminish Darwinian fitness.

Butler identifies the most prominent biological theories that help to explain cooperation in social dilemmas (see sect. 7 of the target article). (1) *Kin selection* (Hamilton 1964) involves cooperation with close relatives, sacrificing individual payoffs in order to maximize the total number of one's genes that are passed on. (2) *Reciprocal altruism* (Trivers 1971) may occur when selfish motives exist for cooperating in long-term relationships. (3) *Indirect reciprocity* (Alexander 1987) operates in established groups when an individual can benefit in the long run by establishing a reputation for cooperativeness.

These three theories certainly help to explain why cooperation occurs in certain circumstances – **Hancock & DeBruine** discuss some interesting and relevant evidence from research into facial resemblance in games – but it seems clear that they cannot provide a complete answer. None of them can explain why cooperation occurs among genetically unrelated strangers in isolated interactions lacking opportunities for reputation-building. Yet we know that it does, in many cases.

A further suggestion of **Butler**'s escapes this criticism. He quotes from Price et al. (2002): "punitive sentiments in collective action contexts have evolved to reverse the fitness advantages that accrue to free riders over producers." It is not unusual for people who take advantage of the cooperation or generosity of others to find themselves socially ostracized or worse. There is now powerful experimental evidence that this tends to promote and maintain cooperation. Fehr and Gächter (2002) studied *altruistic punishment* of defectors, costly to those who administer it, in public goods dilemmas. They found that cooperation flourishes when punishment is possible and tends to break down when it is not.

Gintis also mentions punishment as a possible explanatory mechanism, and **Barclay & Daly** agree with the suggestion of Price et al. (2002) that a propensity to punish defectors may have evolved. Can punishment of defectors explain cooperation in social dilemmas?

Punishment is invariably costly to those who administer it, and hence, is altruistic, because it takes time and energy and invites retaliation. Therefore, natural selection should tend to eliminate it. If the theory is to work, then we must assume that failure to punish defectors is treated as free-riding and hence as a form of second-degree defection that is itself subject to sanctions from other group members. But that raises the question of sanctions against third-degree defectors, who neglect to punish second-degree defectors, and so on, leading to an infinite regress that collapses under its own weight. Juvenal's *Quis custodiet ipsos custodes?* (Who is to guard the guards themselves?) was never more pertinent. Altruistic punishment seems to be a fact of life, but it does not *explain* cooperation. It replaces the problem of explaining cooperation with that of explaining punishment of defectors.

Lazarus makes several useful suggestions for interdisciplinary research on strategic interaction. I am less sure about the relevance of functional brain imaging, discussed by **Berns** and more briefly by **Butler**. This research is intriguing, and useful discoveries are being made, but it is hard to believe that brain imaging "will help resolve the apparent paradoxes." By way of analogy, consider the current debate in the field of artificial intelligence about the "strong AI" proposition that a computer capable of passing the Turing test – by responding to inputs in a manner indistinguishable from a human being – would necessarily have a mind and be capable of thought. No one believes that studying the electronic circuitry of computers will help to resolve this problem, and for analogous reasons I doubt that functional brain imaging can help resolve the conceptual problems associated with strategic interaction.

R6.1. Does unselfishness explain cooperation in social dilemmas?

I agree with **Fantino & Stolarz-Fantino** that people are taught from an early age to be unselfish and cooperative, that such behavior tends to be rewarded throughout life, and that unselfish and cooperative behavior is often reciprocated. However, it is important to point out that, in orthodox game theory, unselfish motives *cannot* explain cooperation in the Prisoner's Dilemma game (PDG) and other social dilemmas. At best, unselfish motives might explain cooperation in experimental games in which the payoffs presented to the players correspond to social dilemmas but the players' utility functions include cooperative or altruistic motives that transform them into other games in which cooperation is an unconditionally best strategy. In any experimental game in which this occurs, *the players are not playing a social dilemma:* extraneous sources of utility have transformed the game into something else.

Rescher (1975) mounted the most strenuous and sustained attempt to solve the paradox of the PDG along these lines, by appealing to unselfish motives and values. He

claimed that "the PDG presents a problem for the conventional view of rationality only when we have been dragooned into assuming the stance of the theory of games itself" (p. 34). Disdainfully placing "dilemma" in quotation marks, Rescher argued that

> the parties were entrapped in the "dilemma" because they did not internalize the welfare of their fellows sufficiently. If they do this, and do so in sufficient degree, they can escape the dilemmatic situation. (p. 48)

This argument collapses as soon as it is pointed out that the players' utilities represented in the payoff matrix are *not* based on a disregard of each other's interests. On the contrary, they are assumed to reflect the players' preferences, taking fully into account their motives, values, tastes, consciences, and moral principles, including any concerns they may have for the welfare of others. **Hancock & DeBruine**'s comment that "non-economic factors influence behavior" is obviously right, provided that economic factors are sufficiently narrowly defined. Further, the evidence that they cite for the effects of personal attractiveness on behavior in the Ultimatum game (see also sect. R7 below) is interesting and instructive, but it is important to remember that utility theory and game theory are entirely neutral with regard to the sources and nature of players' utilities.

Rescher (1975) treated the numbers in the payoff matrix as "'raw,' first-order utilities" and then transformed them into "'cooked,' other-considering, second-order ones" (p. 46) in order to demonstrate how to neutralize the dilemma of the PDG, overlooking the fact that the payoff matrix actually dishes up pre-cooked utilities in the first place. Furthermore, there is no guarantee that cooking raw utilities would invariably neutralize the dilemma. In some games, the payoffs may represent a social dilemma only *after* unselfish motives and values are factored in. As **Camerer** points out, in experimental games, the best we can do is to measure monetary payoffs, but the underlying theory applies to von Neumann-Morgenstern utilities, and these are certainly assumed to include non-monetary components.

Edgeworth's (1881) famous dictum that "the first principle of economics is that every agent is actuated only by self-interest" (p. 16) is trivially – in fact, tautologically – true in modern utility theory. Rational agents try to maximize their expected utilities whenever they are free to choose, and this must be so because their utility functions are defined by their choices. An agent's utility function may nevertheless include concern for the welfare of others, and I believe that, for most non-psychopaths, it does. That, at least, is the standard theory. Whether players' preferences can invariably be represented by static utilities is a moot point – see my comments on team reasoning in section 8.1 of the target article and my outline of the psychological games of Geanakoplos et al. (1989) in section R7 immediately below.

R7. Psychological games and sequential rationality

Carpenter & Matthews are right to point out that the earlier psychological games of Geanakoplos et al. (1989), and theories descended from their work, offer persuasive answers to some of the problems that I discuss. One of the referees of the target article drew my attention to this earlier work, and I was able to insert a brief mention of it in the final version. I agree that it is highly relevant, and that Geanakoplos et al. were the first to use the term *psychological games*, though apparently not *psychological game theory.*

In the theory of Geanakoplos et al. (1989), players' preferences depend not only on the outcomes of a game but also on their beliefs – the arguments of players' utility functions include both outcomes and expectations. The theory models intuitively plausible emotional aspects of strategic interactions, such as surprise, pride, anger, and revenge. Geanakoplos et al. argue persuasively that these factors cannot in general be adequately represented in conventional utility functions. This subverts the orthodox game-theoretic view, defended by **Barclay & Daly**, that relevant psychological factors can always be represented in the payoff functions.

To illustrate the basic idea, a simple psychological game can be constructed from the Ultimatum game, which was mentioned by several commentators. In the Ultimatum game, a monetary prize of $100 (for example) is divided between Player I and Player II as follows. Player I makes a single take-it-or-leave-it proposal for a division of the prize, Player II either accepts or rejects it, and neither player receives anything if the proposal is rejected. From a game-theoretic point of view, Player I should offer Player II one penny, and Player II should accept it, because a penny is better than nothing. Numerous experiments have shown that human players deviate sharply from game theory: Player I usually offers much more than one penny – often a 50–50 split – and Player II usually rejects any offer smaller than about one-quarter of the prize value.

Suppose Player I proposes the following split: $99 for Player I and $1 for Player II. A Player II who is resigned to Player I taking the lion's share of the prize may follow orthodox game theory and accept the offer, preferring $1 to nothing. But a Player I who expects a 50–50 offer may be sufficiently proud or angry to reject the proposal, leaving both players with nothing – emotions aroused by the inequity of the proposal may outweigh the $1. Intuitively, this outcome is a second credible equilibrium, and in the theory of Geanakoplos et al. (1989), it emerges as a *psychological Nash equilibrium.* The particular payoffs, and hence the equilibrium that is likely to be chosen, depend on Player II's expectations.

Carpenter & Matthews do a superb job of tracing the development of these intriguing ideas through the work of Rabin (1993) and others. These are among the most exciting recent developments in game theory, at least from a psychological viewpoint. They help to explain several puzzling phenomena, including cooperation in social dilemmas.

This leads **Carpenter & Matthews** to pose the following reasonable question: "What observed behavior will the 'new psychological game theory' explain that an old(er) . . . one cannot?" To this I reply that the theories discussed in the target article already explain focusing in pure coordination games and selection of payoff-dominant equilibria. They may ultimately help to explain cooperation in backward-induction games such as the Centipede game. The older theories have not, as far as I know, explained these phenomena. Many other strategic phenomena that also remain unexplained by the older theories may yield to new approaches in the future.[5]

R8. Unit of rational agency

I am grateful to **Hurley** for drawing attention to the relevance of Regan's (1980) book on utilitarianism and cooperation. Although Regan did not use the terminology of rational choice theory, he tackled problems closely linked to those addressed in sections 5 and 6 of the target article. In Chapters 2 and 7, he explained with painstaking thoroughness why individualistic payoff maximization, or what he calls *act utilitarianism,* cannot solve the payoff-dominance problem, and in later chapters he put forward and defended a theory of *cooperative utilitarianism* that clearly anticipated team reasoning.

Some commentators are skeptical about the claim in section 8.1 of the target article that team reasoning is inherently non-individualistic. In particular, **Barclay & Daly** "looked in vain for evidence or argument" to support this contention. They claim that team reasoning involves nothing more than "incorporating nonstandard preferences into the decision makers' utility functions." I thought I had shown in section 8.1 of the target article why this is not so, but for those who remain unconvinced, Regan's (1980) book should eradicate any lingering smidgen of doubt.

A standard assumption of decision theory and game theory is that the unit of rational agency is the individual. **Hurley** rejects this assumption and argues that the dogma of individualism is ultimately responsible for the problems of coordination and cooperation that I discuss. This may be so, but I need to point out a non-trivial problem associated with collective agency and related ideas, including (I regret) team reasoning.

Hurley points out that collective agency does not necessarily require collective preferences or collective utility: "As an individual I can recognize that a wholly distinct agent can produce results I prefer to any I could bring about, and that my own acts would interfere [with this process]." But a collective agent representing or implementing the preferences of several individuals needs a method of aggregating their preferences into a unique choice of action or strategy. The problem is that even if each individual has rational preferences in the sense defined in section 2.1 of the target article, a collective agent acting on their behalf cannot, in general, choose rationally or make a reasonable decision. **Hurley** understands that individual rationality can co-exist with collective irrationality but does not follow the implications of this to its awkward conclusion.

Rationality tends to break down at the collective level because of *Arrow's impossibility theorem* (Arrow 1963). This theorem establishes that there can be no rational collective agency implementing the preferences of a group. Even if the members of a group have rational individual preferences, there can in general be no non-dictatorial procedure for aggregating these preferences to reach a decision without violating minimal conditions of fairness and workableness. Arrow proved that if a procedure meets three mild and uncontroversial conditions, then it must be dictatorial. A simple account is given in Colman (1995a, Ch. 10).

Arrow's original proof relies on the profile of individual preferences leading to *Condorcet's paradox of voting.* The simplest example is a group of three individuals judging three options labeled x, y, and z. Suppose that one individual prefers $x > y > z$ (strictly prefers x to y and y to z); a second prefers $y > z > x$; and a third prefers $z > x > y$. Then the group prefers x to y by a majority (because the first and third voters prefer x to y), prefers y to z by a majority (because the first and second voters prefer y to z), and prefers z to x by a majority (because the second and third voters prefer z to x). These collective preferences violate the axiom of transitivity mentioned in section 2.1 of the target article and are therefore irrational.

This means that if the unit of agency is the group, or even an individual agent acting on behalf of the group, in the manner of a trade union negotiator, then there is in general no satisfactory procedure whereby the agent can choose rationally from the set of available options. This poses an intractable problem for the notion of rational collective agency whenever there are more than two individuals and more than two options. Binary decisions escape this particular problem; Arrow's theorem kicks in only when there are three or more individuals and options.

In practice, of course, families, firms, organizations, and other groups *do* sometimes act collectively, but such actions cannot in general be instrumentally rational. Many organizations are managed dictatorially. Arrow's theorem shows that those that are not are liable to encounter situations in which they are doomed to act inconsistently or to find themselves unable to act at all.

Hurley cites the slime mold as a biological example of collective agency. There are other life forms that challenge our usual conception of individuality. Earthworms can be subdivided into two or more independently acting individuals. Sea urchins do not have fully centralized nervous systems and cannot therefore act as individuals. Sponges have no nervous systems at all and hence no individuality in the sense that ordinary unicellular and multicellular organisms are individuals.

In human beings, **Hurley** argues that the unit of agency may sometimes be *below* the level of the individual. **Monterosso & Ainslie** discuss how this might arise in intertemporal choices, in which a person functions as two or more agents with different preferences, as when a short-term preference for eating an ice-cream conflicts with a longer-term preference for slimming. I tend to agree that there is nothing sacrosanct about the individual as the unit of agency, but such subhuman agents (if they will forgive me for calling them that) raise similar problems of consistency and rationality to those outlined above.

People have non-rational ways of coping with problems of self-control, including *resolute choice* (Machina 1991; McClennen 1985, 1990) and various pre-commitment strategies. A frequently quoted pre-commitment strategy from Greek mythology is that of Ulysses, who had himself bound to the mast of his ship in order to prevent himself from yielding to the temptations of the Sirens when the time came to sail near their island. Surprisingly, other animals are also apparently capable of commitment and resolute choice.[6]

R8.1. Is the payoff-dominance principle individually rational?

Weirich provides a thoughtful and subtle analysis of the payoff-dominance principle discussed in section 5.6 of the target article. According to this principle, if one equilibrium point payoff-dominates all others in a game, in the sense of yielding every player a strictly higher payoff than any other

equilibrium point, then rational players will play their parts in it.

I argue in the target article that the payoff-dominance principle cannot be derived from standard assumptions of individual rationality alone. I discuss team reasoning and Stackelberg reasoning as possible ways forward. **Weirich** rejects these approaches on the grounds that "team reasoning conflicts with individualism, and Stackelberg reasoning conflicts with consequentialism." He outlines how the payoff-dominance principle might be based on assumptions of individual rationality, suitably extended.

The payoff-dominance principle was originally introduced by Harsanyi and Selten (1988), hence it is worth pointing out that they agree with me that the principle *cannot* be based on individual rationality. This does not prove me right, but it will make me feel better if I turn out to be wrong. After discussing their subsidiary risk-dominance principle, which *is* based in individual rationality, they write:

> In contrast, payoff dominance is based on *collective* rationality: it is based on the assumption that in the absence of special reasons to the contrary, rational players will choose an equilibrium point yielding all of them higher payoffs, rather than one yielding them lower payoffs. That is to say, it is based on the assumption that rational individuals will cooperate in pursuing their common interests if the conditions permit them to do so. (Harsanyi & Selten 1988, p. 356, emphasis in original)

The point is that, other things being equal, a player who simply maximizes individual payoffs has *no reason* to prefer a payoff-dominant equilibrium point. Thus, there seems to be a hidden inconsistency between **Weirich**'s rejection of team reasoning on the ground that it conflicts with individualism, and his reliance on the payoff-dominance principle.

The extensions to individual rationality that **Weirich** puts forward to ground payoff dominance involve pre-play communication, or what is often called *cheap talk.* For example, he suggests that a rational player preparing to play the Hi-Lo Matching game (shown in Fig. 2 of the target article) "inculcates a disposition to choose *H* and lets others know about his disposition."

The nature and function of the pre-play communication is not specified sufficiently formally to be analyzed rigorously, but this turns out be immaterial, because even if it does indeed lead players to choose the payoff-dominant equilibrium point, I believe that the solution can be shown to be illusory. In particular, pre-play communication cannot secure the foundations of the payoff-dominance principle. A counterexample is Aumann's version of the Stag Hunt game, shown in Figure R1.

		II	
		C	***D***
I	***C***	**9, 9**	**0, 8**
	D	**8, 0**	**7, 7**

Figure R1. Stag Hunt game

Note that this game is really a Hi-Lo Matching game with extra bits and pieces in the cells off the main (top-left to bottom-right) diagonal. As in the Hi-Lo Matching game, there are two pure-strategy equilibrium points at *CC* and *DD,* and the first payoff-dominates the second by a small margin. But for both players *C* is a much riskier choice than *D,* because it entails the possibility of a zero payoff, whereas the worst possible payoff from a *D* choice is 7. In other words, the *maximin* strategy is *D* and the *DD* equilibrium is *risk-dominant,* but both players strictly prefer *CC.*

According to Harsanyi and Selten (1988, pp. 358–59), pre-play communication is useless in this game, because it is in the individual interests of each player to encourage the co-player to choose *C,* by pretending to have a "disposition" to choose *C* and letting the co-player know this (to use **Weirich**'s terminology), but then to play safe by choosing *D.* It was this Stag Hunt game that persuaded Harsanyi and Selten reluctantly to insert the payoff-dominance principle into their theory as an axiom.

For these reasons, I believe that **Weirich**'s suggestion, though apparently innocuous and sensible, is unsatisfactory. Although it gives the desired result in the simple Hi-Lo Matching game, it cannot provide a *general* solution to the payoff-dominance problem.

Janssen agrees with **Weirich** that the "principle of coordination," as he calls it, can be "rationalized on individualistic grounds." But he bases his rationalization on a "principle of optimality" that obviously requires collective reasoning – it is really just Harsanyi and Selten's payoff-dominance principle in disguise – apparently undermining his claim that he does not need "we thinking" to rationalize payoff dominance. I have shown in the preceding paragraphs why I believe that individualistic reasoning cannot supply a firm foundation for the payoff-dominance principle. **Janssen**'s analysis of his version of the Hi-Lo Matching game (his Table 1) seems to rely on self-confirming expectations. My comments toward the end of section R2 above apply equally here.

On the other hand, I agree entirely with **Janssen** that the problem of coordination (discussed in sect. 5 of the target article) is quite separate from the problem of cooperation in social dilemmas (discussed in sect. 6). These problems should not be confused. Furthermore, I welcome his useful suggestion that psychological game theory should take account of framing effects. Too little attention has been paid to framing effects in the literature on game theory and experimental games, though Bacharach (1993) is a striking exception, as **Janssen** points out, and so is Geanakoplos et al. (1989) (see sect. R7 above). The commentaries of **Jones & Zhang**, discussed in section R10 below, and **Vlaev & Chater**, discussed in section R11, are also relevant to this suggestion.

R8.2. Does insufficient reason explain payoff dominance?

Gintis rejects the solutions that I propose for the payoff-dominance problem in isolated interactions, namely, team reasoning and Stackelberg reasoning. He rejects team reasoning on the ground that the *principle of insufficient reason* provides a satisfactory solution. However, in section 5.6 of the target article, I argue that any attempt to solve the problem on the basis of the principle of insufficient reason is logically flawed, and **Gintis** makes no attempt to reply to

that argument. I do not believe that a solution based on the principle of insufficient reason can be defended.

In addition, **Gintis** finds Stackelberg reasoning "implausible" because it allegedly fails to work in the Battle of the Sexes game: "Stackelberg reasoning in this game would lead the players never to coordinate, but always to choose their preferred strategies." This would be a valid objection – in fact, a devastating one – if true. But the Battle of the Sexes is not a Stackelberg-soluble game, because its Stackelberg strategies are out of equilibrium; therefore, the theory makes *no prediction whatsoever* about the choices of the players. Section 8.2 of the target article explains all this and includes the sentence: "Stackelberg reasoning mandates the choice of Stackelberg strategies only in games that are Stackelberg soluble." **Gintis** is perfectly entitled to consider Stackelberg reasoning as implausible, of course, but not for the reason that he gives.

Team reasoning and Stackelberg reasoning may not be appealing as first philosophy, but they do at least plug an explanatory hole. There may be better solutions to the payoff-dominance problem, but until someone formulates them, we are stuck with the theories that we have.

R9. Indeterminacy of psychological game theory

Perugini makes the valid point that psychological game theory, consisting as it does of a plurality of ad hoc theoretical approaches to particular classes of games, generates a kind of second-order indeterminacy. Various nonstandard forms of psychological game-theoretic reasoning may produce determinate local solutions, but they do not add up to a comprehensive theory because they "offer no tools to select among these different reasoning concepts" in specific cases. Perugini illustrates this problem vividly by pointing out that team reasoning is no better than orthodox game theory at explaining human behavior in the Ultimatum game.

Along similar lines, **Kokinov** points out that different forms of reasoning involve different optimization criteria and common beliefs, and that there is nothing to specify "how and when these additional criteria are triggered and where the common beliefs come from." **Haller** comments that "novel solution concepts may be compelling in some contexts and unconvincing under different but similar circumstances," as when Stackelberg reasoning yields unsatisfactory solutions if applied to certain classes of Stackelberg-solvable games that he identifies.

This is all true. There does not exist a psychological game theory that is both comprehensive and free of the drawbacks of orthodox game theory. In the absence of such a theory, we need particular remedies for particular problems. This is not so very different from the current state of theoretical development in any branch of psychology – there is no comprehensive grand theory, just a collection of more modest theories that explain certain classes of behavior but are apt to generate absurd or empirically incorrect predictions when applied in the wrong contexts. I do not feel any need to apologize for the heterogeneity of psychological game theory, though of course a comprehensive and rigorous grand theory would be much better.

From the standpoint of cognitive psychology, **Shiffrin** grasps the nettle of theoretical plurality with both hands. He suggests that rationality should be interpreted not as an axiomatic system of general applicability, but as a psychological concept defined in relation to particular games. According to this view, a decision maker must first decide what theory of rational decision making applies to the current game, then whether a jointly rational solution exists, and, if so, what it is. Shiffrin illustrates this by applying Spohn's (2001) theory of dependency equilibria to the Centipede game. Although the general approach seems quite radical, it looks promising.

I tend to agree with **Shiffrin** that there must be something wrong with the backward induction argument as it is usually applied to the Centipede game (summarized in sect. 7.4 of the target article). The argument is persuasive, and that may be because it is valid, but it is possible for an argument to be valid – necessarily true if its premises are true – but unsound if one or more of its premises is false. The premises of the backward induction argument are the common knowledge and rationality (CKR) assumptions set out in section 4 of the target article, and they are certainly inadequate if **Shiffrin** is right in thinking that rationality must be defined in terms of how the entire game is played, rather than how each decision is made. This seems closely related to the notion of resolute choice (see the end of sect. R8 above).

R10. Depth of strategic reasoning

According to **Jones & Zhang**, although the CKR axioms are designed to make normative decision theory applicable to games (see sect. 3 of the target article), they are far too limiting. These commentators argue that rational choice theory can be salvaged if players are assumed to be instrumentally rational and to anchor their rationality not on a priori assumptions of their co-players' rationality, but on theory-of-mind models of their co-players "based on general experience with human behavior."

This is an interesting and plausible approach, but it has one worrying anomaly. It assumes that players are instrumentally rational but that they do not necessarily model their co-players as instrumentally rational. It seems unreasonable for rational players not to credit their co-players with rationality equal to their own. Apart from everything else, the asymmetry implies that players' models of one another could never be common knowledge in a game. This may not be a knock-down argument, but it does seem potentially problematic.

In support of their approach, **Jones & Zhang** discuss a fascinating pair of experiments by Hedden and Zhang (2002) on depth of strategic reasoning. The CKR assumptions imply indefinitely iterated recursive reasoning ("I think that you think that I think . . ."), but Hedden and Zhang found that players tend to operate at shallow levels only. Some zero-order reasoning was observed, with players choosing strategies myopically, without considering their co-players' viewpoints; but most players began with first-order reasoning, defined as behavior that maximizes payoffs against co-players who use zero-order reasoning. When pitted against first-order co-players, some of the experimental players began to use second-order reasoning, but even after 30 repetitions of the game, fewer than 40 percent had progressed beyond first-order reasoning.

Hedden and Zhang's (2002) experiments were shrewdly designed and well executed, although I have drawn atten-

tion elsewhere to some significant methodological problems with them (Colman 2003). The findings broadly corroborate those of earlier experiments on depth of reasoning in so-called *beauty contest games* and other games that are solvable by iterated deletion of strongly dominant strategies (notably Stahl & Wilson 1995). Findings from disparate experimental games converge on the conclusion that human players generally manage only first-order or at most second-order depth of strategic reasoning.

It is worth commenting that research into cognitive processing of recursively embedded sentences has also shown that people can handle only one or two levels of recursion (Christiansen & Chater 1999; Miller & Isard 1964). The following four-level embedded sentence is virtually incomprehensible: *The article that the commentary that the student that the professor that the university hired taught read criticized was written by me.* One level of embedding causes no problems: *The article that the commentary criticized was written by me.* Two levels of embedding can be handled with effort and concentration: *The article that the commentary that the student read criticized was written by me.* Three or more levels are impossible to process and look ungrammatical. There are evidently severe limitations to human cognitive capacities for multi-level recursive thinking in language as in games.

I agree with **Jones & Zhang** that facts like these need to be taken into account in any descriptively accurate game theory. But they seem to show that human players are themselves imperfectly rational, not merely that they model their co-players as irrational. In any event, a theory according to which players are instrumentally rational but do not credit their co-players with the same sophistication as themselves seems internally unsatisfactory.

In the Centipede game, singled out for discussion by these commentators, their assumptions do indeed appear to allow Player II to respond to a cooperative opening move by Player I. This may enable Player II to model Player I as a tit-for-tat player and therefore to respond cooperatively. However, it seems that the backward induction argument may nevertheless be retained, in which case, unless I am mistaken, Player I may be left with a reason to cooperate and a reason to defect – a contradiction. The Centipede game is a notoriously hard nut to crack.

R11. Prospect relativity in games

When people first think about repeated games, they often fall into the trap of assuming that any theoretical conclusions about a one-shot game can be applied to each repetition of it by the same players. The *supergame* that results when a *stage game* is repeated a number of times is, in fact, a new game with its own equilibrium points, and conclusions about the stage game cannot be applied straightforwardly to the supergame. Psychologically, however, players frequently think about each repetition as a separate game.

A grandiose question arising from this is whether we should model *all* the games in a player's life as a single supergame. We would probably want to call it the Game of Life, had John Conway not already taken the name for his cellular automata. It seems highly unlikely that different games have absolutely no bearing on one another but equally unlikely that people analyze them all together. This is an empirical question, and **Vlaev & Chater** take a step in the direction of answering it. They cite evidence that prospects in risky individual decisions cannot be considered independently of previous risky decisions, and that such *prospect relativity* also occurs in games. They are probably right in suggesting that psychological game theory needs to be supplemented by a *cognitive game theory.*

The findings on game relativity that **Vlaev & Chater** cite relate to the iterated Prisoner's Dilemma game, though the findings may turn out to apply across different games. They found that cooperation and expectations of cooperation in each stage game were strongly dependent on cooperativeness in preceding games. Their explanation for these findings is that players have poor notions of absolute cooperativeness, risk, and utility, and that they therefore make relative judgments. This suggestion fits in with evidence from cognitive psychology, and (if I understand the findings correctly) *prospect theory* (Kahneman & Tversky 1979; Tversky & Kahneman 1992). It is also closely related to the evidence cited by **Fantino & Stolarz-Fantino** of a pronounced effect of past history on decision making. This work provides another answer to **Janssen**'s plea for more research into "how people describe the game situation to themselves."

R12. Research methodology

I think that **Schuster**'s analogy between psychological game theory and scientific doctrines that are "amended again and again in a vain attempt to forge an accommodation with a new reality" is a little unfair. He may have in mind Ptolemaic epicycles, postulated to explain the observed deviations of the orbits of some celestial bodies from perfect circles before Copernicus introduced a heliocentric astronomy in the sixteenth century, and mentioned by **Sigmund**. Sigmund referred to epicycles in connection with a relatively innocent amendment designed to bring game theory more closely in line with intuition and empirical observations.

The purpose of any theory is to explain, and there are three ways in which it can prove inadequate: through being *indeterminate, misleading,* or *unfalsifiable.* A theory is *indeterminate* if it fails to generate clear predictions; it is *misleading* if it generates predictions that are refuted by empirical observations; and it is *unfalsifiable* if there are no empirical observations that could refute it and therefore no possibility of testing it. Some aspects of game theory are certainly misleading inasmuch as they generate predictions that are refuted by empirical observations, especially in social dilemmas, backward induction games, and Ultimatum games; but its most serious and obvious failing is its systematic indeterminacy. Ptolemaic epicycles and similar theoretical amendments are objectionable because they render theories unfalsifiable. Neither orthodox game theory nor psychological game theory can be accused of that.

Science advances by replacing old theories with new ones that make better predictions. Newton's theory explained the motions of the planets, moons, and comets in the solar system without epicycles, and it survived empirical tests that could have falsified it. For centuries it appeared to yield no misleading predictions, until, in 1859, astronomers discovered that the planet Mercury drifts from the predicted orbit by what turned out to be 43 seconds of arc, or roughly one hundredth of a degree, per century. Further-

more, it failed to predict bending of light and black holes. In 1916, Einstein put forward a general theory of relativity that removed these inadequacies and also withstood empirical tests that could have falsified it. It now appears that Einstein's theory does not predict cosmic radiation satisfactorily, and no doubt it too will be replaced by something better in due course. That is how science advances in ideal cases.

I believe that **Schuster**'s characterization of experimental games is misleading. He asserts that "the basic design of laboratory experiments" involves a total absence of social interaction between participants: "anonymous players are physically isolated in separate cubicles." This may be a fair description of many experiments, but it is far from being universal. Communication is integral to experiments based on Ultimatum games and bargaining games in general, and it often plays an important part in experiments on coalition-formation in cooperative games. Even in research into behavior in dyadic and multi-player social dilemmas, numerous experiments, dating back to 1960, have focused on the effects of verbal and nonverbal communication between players (see Colman 1995a).

The bleak picture that **Schuster** paints of experimental games, with players isolated in solitary confinement and a total "absence of real-life social interaction," contains a grain of truth, but it is an exaggeration. One of his suggested alternatives, "to study examples [of real cooperation] . . . under free-ranging conditions, where cooperation is intrinsically social," is fraught with problems. Ethological investigations are certainly useful, especially in research with animals, but the lack of experimental manipulation of independent variables and problems of controlling extraneous variables limit the conclusions that can be drawn from them. His other suggestion, "to incorporate free-ranging conditions into experimental models of cooperation that allow social and non-social variables to be manipulated," seems more promising, and he cites some interesting animal research along those lines.

R13. Concluding remarks

I approached the commentaries with an open mind, and many of the criticisms seemed cogent and damaging when I first read them. After thinking about them carefully, I came to the conclusion that some are indeed valid, and I acknowledge them in this response. In particular, I accept that the theoretical plurality of psychological game theory generates an unwelcome second-order indeterminacy, and that there are earlier theoretical contributions that provide solutions to some – though not all – of the problems discussed in the target article. However, the various attempts to show that these problems are not really problematic if viewed in the correct light, or to show how they can be solved without recourse to psychological game theory or nonstandard assumptions, turn out on careful examination to be based on misunderstandings or misleading arguments. When an argument is expressed informally, it sometimes appears far more compelling than it really is.

After studying and replying to the commentaries, my interpretations of the fundamental issues raised in the target article remain substantially unchanged, although I have learned a great deal. On the central questions, my opinions have actually been reinforced by being exposed to criticisms that appeared convincing at first but less persuasive on closer inspection.

I am more confident than before that the standard interpretation of instrumental rationality as expected utility maximization does not and cannot explain important features of interactive decision making. This central thesis has been endorsed by several of the commentators and subjected to critical examination from many different angles by others, and I believe that it has survived intact and has even been fortified. If the central thesis is correct, then psychological game theory, in some form or another, is needed to provide a more complete and accurate understanding of strategic interaction. This is an exciting challenge.

Replying to the commentaries has sharpened and clarified many of the issues and helped me to view them from fresh angles. Seriously interested readers will also gain a broader perspective and clearer insight into the fundamental problems and solutions by reading the target article along with the commentaries and my response, rather than by reading the target article alone.

ACKNOWLEDGMENTS

I am grateful to Ken Hughes and Caroline Salinger for helpful comments on an earlier draft of this article.

NOTES

1. Monterosso & Ainslie claim that this justification for choosing H is "descriptively accurate" and "prescriptively rational," but they do not explain how this can be so, given that it leads to a contradiction.

2. In front of you is a transparent box containing $1,000 and an opaque box containing either $1 million or nothing. You have the choice of taking either the opaque box only, or both boxes. You have been told, and believe, that a predictor of human behavior, such as a sophisticated computer programmed with psychological information, has already put $1 million in the opaque box if and only if it has predicted that you will take only that box, and not the transparent box as well, and you know that the predictor is correct in most cases (95 percent of cases, say, although the exact figure is not critical). Both strategies can apparently be justified by simple and apparently irrefutable arguments. The expected utility of taking only the opaque box is greater than that of taking both boxes, but the strategy of taking both boxes is strongly dominant in the sense that it yields a better payoff irrespective of what is already in the boxes. For a thorough examination of this problem, see Campbell and Sowden (1985).

3. A researcher wishes to test the hypothesis that all ravens are black. According to the logic of empirical induction, every black raven that is observed is a confirming instance that renders the hypothesis more probable. However, the propositions "All ravens are black" and "All non-black objects are not ravens" are logically equivalent, having the same truth value and differing merely in wording. It follows that, on a rainy day, instead of examining ravens, the researcher could stay indoors and examine non-black objects, such as a green book, a blue curtain, a white lampshade, and so on, checking that they are not ravens, because each of these is also a confirming instance of the hypothesis. Most logicians agree that this conclusion is true, and that its *prima facie* absurdity arises from a psychological illusion rooted in misguided intuition. (On the other hand, perhaps it is a refutation of induction.)

4. In Van Lange's (1999) model, all social value orientations are interpreted as maximizations of simple linear functions of the variables W_1 (own payoff), W_2 (co-player's payoff), and W_3 ("equality in outcomes"). Although W_3 is not formally defined, from Van Lange's examples it is obviously equal to $-|W_1 - W_2|$. *Altruism* is simply maximization of W_2, and because in the Hi-Lo Matching game $W_1 = W_2$, this is equivalent to maximizing W_1. It is not hard

to see that no linear combination of these three variables can solve the payoff-dominance problem. Note first that, because $W_3 = -|W_1 - W_2|$, any linear function of W_1, W_2, and W_3 can be expressed as $aW_1 + bW_2$, where a and b are suitably chosen real numbers. Furthermore, because $W_1 = W_2$ in the Hi-Lo Matching game, maximizing $aW_1 + bW_2$ amounts to maximizing W_1 for any values of a and b, and this is simply individualistic payoff maximization, which leaves neither player with any reason for choosing H, as shown in section 5.6 of the target article.

5. Among those that spring readily to mind are behavior in market entry games (Camerer & Lovallo 1999); coordination through the confidence heuristic (Thomas & McFadyen 1995); timing effects in games with asymmetric equilibria (Cooper et al. 1993); and depth-of-reasoning effects in normal-form games (Colman 2003; Hedden & Zhang 2002).

6. In the first experimental demonstration of commitment and self-control in animals, Rachlin and Green (1972) presented five hungry pigeons with a repeated choice between an immediate small reward (two seconds eating grain) and a delayed larger reward (four seconds delay followed by four seconds eating grain). All of the pigeons chose the immediate small reward on virtually every trial. The same pigeons were then presented with a repeated choice between (a) 16 seconds delay followed by the choice described above between an immediate small reward and a delayed larger reward; and (b) 20 seconds delay followed by the larger reward with no choice. Four of the five pigeons chose (b) on most trials – three of them on more than 80 percent of trials. This looks to me very much like resolute choice (Machina 1991; McClennen 1985; 1990). A similar phenomenon has more recently been observed in honeybees (Cheng et al. 2002). For a review of research into self-control, see Rachlin (2000).

References

Letters "a" and "r" appearing before authors' initials refer to target article and response, respectively.

Abell, P., ed. (1991) *Rational choice theory.* Edward Elgar. [aAMC]

Ainslie, G. (2001) *Breakdown of will.* Cambridge University Press. [JM]

Ainslie, G. & Monterosso, J. (2003) Building blocks of self-control: Increased tolerance for delay with bundled rewards. *Journal of the Experimental Analysis of Behavior* 79:83–94. [JM]

Alexander, R. D. (1987) *The biology of moral systems.* Aldine de Gruyter. [PB, DJB, rAMC]

Allais, M. & Hagen, O. (1979) *Expected utility theory and the Allais paradox.* Kluwer. [KS]

Alvard, M. (in press) The adaptive nature of culture. *Evolutionary Anthropology.* [MA]

Anand, P. (1990) Two types of utility: An experimental investigation into the prevalence of causal and evidential utility maximization. *Greek Economic Review* 12:58–74. [aAMC]

Anderson, J. R. & Schooler, L. J. (1991) Reflections of the environment in memory. *Psychological Science* 2(6):396–408. [MJ]

Anderson, R. B., Tweney, R. D., Rivardo, M. & Duncan, S. (1997) Need probability affects retention: A direct demonstration. *Memory and Cognition* 25(6):867–72. [MJ]

Andras, P., Roberts, G. & Lazarus, J. (2003) Environmental risk, cooperation and communication complexity. In: *Adaptive agents and multi-agent systems,* ed. E. Alonso, D. Kudenko & D. Kazakov. Springer-Verlag. [JL]

Andreoni, J. & Miller, J. H. (1993) Rational cooperation in the finitely repeated Prisoner's Dilemma: Experimental evidence. *The Economic Journal* 103:570–85. [aAMC]

Antonelli, G. A. & Bicchieri, C. (1994) Backwards forward induction. In: *Theoretical aspects of reasoning about knowledge: Proceedings of the Fifth Conference (TARK 1994),* pp. 24–43, ed. R. Fagin. Morgan Kaufmann. [aAMC]

Arrington, R. L. (1998) *Western ethics: An historical introduction.* Blackwell. [JL]

Arrow, K. J. (1963) *Social choice and individual values, 2nd edition.* Wiley. [arAMC, RS]

Arrow, K. J., Colombatto, E., Perlman, M. & Schmidt, C., eds. (1996) *The rational foundations of economic behavior: Proceedings of the IEA Conference, Turin, Italy.* Macmillan. [aAMC]

Aumann, R. J. (1976) Agreeing to disagree. *Annals of Statistics* 4:1236–39. [aAMC]

(1995) Backward induction and common knowledge of rationality. *Games and Economic Behavior* 8:6–19. [aAMC]

(1996) A note on backward induction: Reply to Binmore. *Games and Economic Behavior* 17: 138–46. [aAMC]

(1998) On the Centipede game. *Games and Economic Behavior* 23:97–105. [aAMC]

(2000) Economic theory and mathematical method: An interview. In: R. J. Aumann, *Collected papers, vol. 1.* MIT Press. [aAMC]

Aumann, R. J. & Brandenburger, A. (1995) Epistemic conditions for Nash equilibrium. *Econometrica* 63:1161–80. [aAMC]

Axelrod, R. (1984) *The evolution of cooperation.* Basic Books. [aAMC, JL]

(1997) *The complexity of cooperation: Agent-based models of competition and collaboration.* Princeton University Press. [aAMC, JL]

Axelrod, R. & Hamilton, W. D. (1981) The evolution of cooperation. *Science* 211:1390–96. [JL]

Bacharach, M. (1987) A theory of rational decision in games. *Erkenntnis* 27:17–55. [aAMC]

(1993) Variable universe games. In: *Frontiers of game theory,* ed. K. Binmore, A. Kirman & P. Tani. MIT Press. [arAMC, MCJ]

(1999) Interactive team reasoning: A contribution to the theory of co-operation. *Research in Economics* 53:117–47. [aAMC, SH]

Bacharach, M. & Hurley, S. (1991) Issues and advances in the foundations of decision theory. In: *Foundations of decision theory,* ed. M. Bacharach & S. Hurley. Blackwell. [aAMC]

Baker, F. & Rachlin, H. (2002) Teaching and learning in a probabilistic prisoner's dilemma. *Behavioural Processes* 57:211–26. [MS]

Bargh, J. A. & Ferguson, M. J. (2000) Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin* 126:925–45. [JIK]

Baron-Cohen, S. (1995) *Mindblindness: An essay on autism and theory of mind.* MIT Press. [MA]

Barton, R. & Dunbar, R. (1997) Evolution of the social brain. In: *Machiavellian intelligence II: Extensions and evaluations,* ed. A. Whiten & R. Byrne. Cambridge University Press. [DJB]

Basu, K. (1990) On the non-existence of a rationality definition for extensive games. *International Journal of Game Theory* 19:33–44. [aAMC]

Batson, C. D. (1987) Prosocial motivation: Is it ever truly altruistic? *Advances in Experimental Social Psychology* 20:65–122. [JL]

Bell, D. E., Raiffa, H. & Tversky, A. (1988) Descriptive, normative, and prescriptive interactions in decision making. In: *Decision making: Descriptive, normative, and prescriptive interactions,* ed. D. E. Bell, H. Raiffa & A. Tversky. Cambridge University Press. [arAMC]

Berkeley, D. & Humphreys, P. (1982) Structuring decision problems and the "bias heuristic." *Acta Psychologica* 50:201–52. [aAMC]

Bernheim, B. D. (1984) Rationalizable strategic behavior. *Econometrica* 52:1007–28. [arAMC]

Bernoulli, D. (1738/1954) Specimen theoriae novae de mensura sortis. *Comentarii Acdemii Scientarum Imperialis Petropolitanae* 5:175–92. (English trans.: Sommer, L. (1954) Exposition of a new theory on the measurement of risk. *Econometrica* 22:23–36.) [AR]

Bicchieri, C. (1989) Self-refuting theories of strategic interaction: A paradox of common knowledge. *Erkenntnis* 30:69–85. [aAMC]

(1993) *Rationality and coordination.* Cambridge University Press. [aAMC]

Bicchieri, C. & Antonelli, G. A. (1995) Game-theoretic axioms for local rationality and bounded knowledge. *Journal of Logic, Language, and Information* 4:145–67. [aAMC]

Binmore, K. (1987) Modeling rational players: Part 1. *Economics and Philosophy* 3:179–214. [aAMC]

(1992) *Fun and games: A text on game theory.* Heath. [aAMC]

(1994a) *Playing fair: Game theory and the social contract, vol. I.* MIT Press. [arAMC]

(1994b) Rationality in the Centipede. In: *Theoretical aspects of reasoning about knowledge: Proceedings of the Fifth Conference (TARK 1994),* pp. 150–59, ed. R. Fagin. Morgan Kaufmann. [aAMC]

Binmore, K., Gale, J. & Samuelson, L. (1995) Learning to be imperfect: The ultimatum game. *Games and Economic Behavior* 8:56–90. [JPC]

Björnerstedt, J. & Weibull, J. (1996) Nash equilibrium and evolution by imitation. In: *The rational foundations of economic behavior,* ed. K. Arrow, E. Colombatto, M. Perlman & E. Schmidt. Macmillan. [JPC]

Boesch, C. & Boesch, H. (1989) Hunting behavior of wild chimpanzees in the Tai National Park. *American Journal of Physical Anthropology* 78:547–73. [RS]

Bonanno, G. (1991) The logic of rational play in games of perfect information. *Economics and Philosophy* 7:37–65. [aAMC]

Bowley, A. L. (1924) *The mathematical groundwork of economics.* Oxford University Press. [HH]

Boyd, R. & Richerson, P. J. (1991) Culture and cooperation. In: *Cooperation and prosocial behaviour,* ed. R. A. Hinde & J. Groebel. Cambridge University Press. [JL]
(1992) Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13:171–95. [PB]
(1995) Why does culture increase human adaptability? *Ethology and Sociobiology* 16:125–43. [MA]
(1996) Why culture is common, but cultural evolution is rare. *Proceedings of the British Academy* 88:77–93. [MA]
Breiter, H. C., Aharon, I., Kahneman, D., Dale, A. & Shizgal, P. (2001) Functional imaging of neural responses to expectancy and experience of monetary gains and losses. *Neuron* 30(2):619–39. [GSB]
Brewer, M. B. & Kramer, R. M. (1986) Choice behavior in social dilemmas: Effects of social identity, group size, and decision framing. *Journal of Personality and Social Psychology* 50:543–49. [aAMC]
Broome, J. (1991) Rationality and the sure-thing principle. In: *Thoughtful economic man,* ed. G. Meeks. Cambridge University Press. [aAMC]
Buunk, B. P. & Schaufeli, W. B. (1999) Reciprocity in interpersonal relationships: An evolutionary perspective on its importance for health and well-being. *European Review of Social Psychology* 10:259–91. [JL]
Call, J. & Tomasello, M. (1999) A nonverbal theory of mind test: The performance of children and apes. *Child Development* 70:381–95. [SH]
Camerer, C. F. (1995) Individual decision making. In: *Handbook of experimental economics,* ed. J. H. Kagel & A. E. Roth. Princeton University Press. [aAMC]
(1997) Progress in behavioral game theory. *Journal of Economic Perspectives* 11:167–88. [aAMC]
(1999) Behavioral economics: Reunifying psychology and economics. *Proceedings of the National Academy of Science, USA* 96(10):575–77. [GSB]
(2003) *Behavioral game theory: Experiments on strategic interaction.* Princeton University Press. [CFC]
Camerer, C. F. & Ho, T.-H. (1999) Experience-weighted attraction learning in games: A unifying approach. *Econometrica* 67:827–74. [WDC]
Camerer, C. F., Ho, T.-H. & Chong, J. K. (2000) A cognitive hierarchy theory of one-shot games. (unpublished manuscript). http://www.hss.caltech.edu/~camerer/camerer.html [CFC]
Camerer, C. F. & Lovallo, D. (1999) Overconfidence and excess entry: An experimental approach. *American Economic Review* 89:306–18. [rAMC]
Campbell, R. & Snowden L., eds. (1985) *Paradoxes of rationality and cooperation: Prisoner's dilemma and Newcomb's problem.* University of British Columbia Press. [arAMC]
Cartwright, J. (2000) *Evolution and human behavior.* Macmillan. [DJB]
Casajus, A. (2001) *Focal points in framed games.* Springer-Verlag. [aAMC]
Case, D., Fantino, E. & Goodie, A. S. (1999) Base-rate training without case cues reduces base-rate neglect. *Psychonomic Bulletin and Review* 6(2):319–27. [EF]
Casebeer, W. D. (2003) *Natural ethical facts: Evolution, connectionism, and moral cognition.* MIT Press. [WDC]
Casebeer, W. D. & Churchland, P. S. (2003) The neural mechanisms of moral cognition: A multiple-aspect approach to moral judgment and decision-making. *Biology and Philosophy* 18:169–94. [WDC]
Charlton, B. G. (1997) The inequity of inequality: Egalitarian instincts and evolutionary psychology. *Journal of Health Psychology* 2(3):413–25. [JL]
Chase, I. D. (1980) Cooperative and noncooperative behaviour. *American Naturalist* 115:827–57. [RS]
Cheney, D. & Seyfarth, R. (1990) *How monkeys see the world: Inside the mind of another species.* University of Chicago Press. [MA]
Cheng, K., Pena, J., Porter, M. A. & Irwin, J. D. (2002) Self-control in honeybees. *Psychonomic Bulletin Review* 9:259–63. [rAMC]
Christiansen, M. H. & Chater, N. (1999) Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* 23:157–205. [rAMC]
Cialdini, R. B., Schaller, M., Houlihan, D., Arps, K., Fultz, J. & Beaman, A. L. (1987) Empathy-based helping: Is it selflessly or selfishly motivated? *Journal of Personality and Social Psychology* 52(4):749–58. [JL]
Cohen, L. J. (1981) Can human irrationality be experimentally demonstrated? *Behavioral and Brain Sciences* 4:317–70. [aAMC]
Coleman, A. A., Colman, A. M. & Thomas, R. M. (1990) Cooperation without awareness: A multiperson generalization of the minimal social situation. *Behavioral Science* 35:115–121. [aAMC]
Coleman, J. S. & Fararo, T. J. (1992) *Rational choice theory: Advocacy and critique.* Sage. [aAMC]
Colman, A. M. (1995a) *Game theory and its applications in the social and biological sciences, 2nd edition.* Butterworth-Heinemann. [arAMC]
(1995b) Prisoner's Dilemma, chicken, and mixed-strategy evolutionary equilibria. *Behavioral and Brain Sciences* 18:550–51. [aAMC]
(1997) Salience and focusing in pure coordination games. *Journal of Economic Methodology* 4:61–81. [aAMC]
(1998) Rationality assumptions of game theory and the backward induction paradox. In: *Rational models of cognition,* ed. M. Oaksford & N. Chater. Oxford University Press. [aAMC]
(2003) Depth of strategic reasoning in games. *Trends in Cognitive Sciences* 7:2–4. [rAMC]
Colman, A. M. & Bacharach, M. (1997) Payoff dominance and the Stackelberg heuristic. *Theory and Decision* 43:1–19. [aAMC]
Colman, A. M. & Stirk, J. A. (1998) Stackelberg reasoning in mixed-motive games: An experimental investigation. *Journal of Economic Psychology* 19:279–93. [aAMC, HG, JIK]
Colman, A. M., Stirk, J. & Park, J. (2001) Non-standard reasoning in games: Theory and experiments on Stackelberg reasoning. Paper presented at the V SAET Conference, Ischia, July 2–8, 2001. [aAMC]
Colman, A. M. & Wilson, J. C. (1997) Antisocial personality disorder: An evolutionary game theory analysis. *Legal and Criminological Psychology* 2:23–34. [aAMC]
Cooper, R. W., DeJong, D. V. & Forsythe, R. (1996) Cooperation without reputation: Experimental evidence from Prisoner's Dilemma games. *Games and Economic Behavior* 12:187–218. [aAMC]
Cooper, R. W., DeJong, D. V., Forsythe, R. & Ross, T. W. (1990) Selection criteria in coordination games: Some experimental results. *American Economic Review* 80:218–33. [aAMC]
(1993) Forward induction in the Battle-of-the-Sexes game. *American Economic Review* 83:1303–13. [rAMC]
Cosmides, L. & Tooby, J. (1992) Cognitive adaptations for social exchange. In: *The adapted mind: Evolutionary psychology and the generation of culture,* ed. J. H. Barkow, L. Cosmides & J. Tooby. Oxford University Press. [JL]
Crawford, V. P. & Haller, H. (1990) Learning how to cooperate: Optimal play in repeated coordination games. *Econometrica* 58:571–95. [aAMC]
Cubitt, R. P. & Sugden, R. (1994) Rationally justifiable play and the theory of non-cooperative games. *Economic Journal* 104:798–803. [aAMC]
(1995) Games and decisions. *Greek Economic Review* 17:39–60. [aAMC]
Damasio, A. (1994) *Descartes' error: Emotion, reason and the human brain.* Putnam. [DJB]
Dawes, R. M. (1973) The commons dilemma game: An n-person mixed-motive game with a dominating strategy for defection. *Oregon Research Institute Research Bulletin* 13:2. [aAMC]
(1980) Social dilemmas. *Annual Review of Psychology* 31:169–93. [aAMC]
(1988) *Rational choice in an uncertain world.* Harcourt, Brace, Jovanovich. [arAMC]
(2000) A theory of irrationality as a "reasonable" response to an incomplete identification. *Synthese* 122:133–63. [aAMC]
Dawes, R. M., van de Kragt, J. C. & Orbell, J. M. (1988) Not me or thee but we: The importance of group identity in eliciting cooperation in dilemma situations; Experimental manipulations. *Acta Psychologica* 68:83–97. [aAMC]
(1990) Cooperation for the benefit of us: Not me, or my conscience. In: *Beyond self-interest,* ed. J. Mansbridge. University of Chicago Press. [aAMC]
Dawkins, R. (1989) *The selfish gene, 2nd edition.* Oxford University Press. [aAMC]
DeBruine, L. M. (2002) Facial resemblance enhances trust. *Proceedings of the Royal Society of London B* 269:1307–12. [PJH]
(unpublished manuscript) Facial resemblance as a factor influencing outcomes in economic games. [PJH]
Dennett, D. C. (1995) *Darwin's dangerous idea.* Simon and Schuster. [RS]
Dixit, A. K. & Nalebuff, B. J. (1991) *Thinking strategically: The competitive edge in business, politics, and everyday life.* Norton. [aAMC]
Doyle, J. R., O'Connor, D. J., Reynolds, G. M. & Bottomley, P. A. (1999) The robustness of the asymmetrically dominated effect: Buying frames, phantom alternatives, and in-store purchases. *Psychology and Marketing* 16:225–43. [aAMC]
Dufwenberg, M. & Kirchsteiger, G. (1998) A theory of sequential reciprocity. Tilburg Center for Economic Research, Discussion Paper No. 9837. [JPC]
Dugatkin, L. A. (1995) Partner choice, game theory and social behavior. *Journal of Quantitative Anthropology* 5:3–14. [RS]
(1997) *Cooperation among animals: An evolutionary perspective.* Oxford University Press. [JL, RS]
Eckel, C. C. & Wilson, R. K. (1998a) Reciprocal fairness and social signalling: Experiments with limited reputations. Paper presented at the American Economic Association Meeting, New York, NY, 1998. [PB, PJBH]
(1998b) Reputation formation in simple bargaining games. Paper presented at the Midwest Political Science Association Meeting, Chicago, IL. [PJBH]
Edgeworth, F. Y. (1881/1967) *Mathematical psychics.* Augustus M. Kelley/Kegan Paul. (Original work published 1881.) [arAMC]
Eells, E. (1985) Causality, decision, and Newcomb's paradox. In: *Paradoxes of rationality and cooperation: Prisoner's Dilemma and Newcomb's problem,* ed. R. Campbell & L. Sowden. University of British Columbia Press. [aAMC]

El-Gamal, M. A., McKelvey, R. D. & Palfrey, T. R. (1993) A Bayesian sequential experimental study of learning in games. *Journal of the American Statistical Association* 88(422):428–35. [aAMC]

Ellsberg, D. (1961) Risk, ambiguity and the Savage axiom. *Quarterly Journal of Economics* 75:643–69. [KS]

Elster, J. (1986) Introduction. In: *Rational choice,* ed. J. Elster. Basil Blackwell. [aAMC]

(1989) *Solomonic judgements: Studies in the limitations of rationality.* Cambridge University Press. [rAMC]

Erev, I. & Roth, A. E. (1998) Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* 88:848–81. [WDC]

Fagin, R., Halpern, J. Y., Moses, Y. & Vardi, M. Y. (1995) *Reasoning about knowledge.* MIT Press. [aAMC]

Falk, A., Fehr, E. & Fischbacher, U. (2000) Testing theories of fairness – Intentions matter. Institute for Empirical Research in Economics, Working Paper No. 63. [JPC]

Falk, A. & Fischbacher, U. (2000) A theory of reciprocity. Institute for Empirical Research in Economics, Working Paper No. 6. [JPC]

Fantino, E. (1998a) Behavior analysis and decision making. *Journal of the Experimental Analysis of Behavior* 69:355–64. [EF]

(1998b) Judgment and decision making: Behavioral approaches. *The Behavior Analyst* 21:203–18. [EF]

Fantino, E. & Stolarz-Fantino, S. (2002a) From patterns to prosperity: A review of Rachlin's *The Science of Self-Control. Journal of the Experimental Analysis of Behavior* 78:117–25. [EF]

(2002b) The role of negative reinforcement; or: Is there an altruist in the house? *Behavioral and Brain Sciences* 25(2):257–58. [EF]

Farrell, J. (1987) Cheap talk, coordination, and entry. *Rand Journal of Economics* 18:34–39. [aAMC]

(1988) Communication, coordination and Nash equilibrium. *Economics Letters* 27:209–14. [aAMC]

Feger, H. (1991) Cooperation between groups. In: *Cooperation and prosocial behaviour,* ed. R. A. Hinde & J. Groebel. Cambridge University Press. [JL]

Fehr, E. & Gächter, S. (2000) Cooperation and punishment in public goods experiments. *American Economic Review* 90:980–94. [PB, KS]

(2002) Altruistic punishment in humans. *Nature* 415: 137–40. [rAMC]

Fehr, E. & Schmidt, K. (1999) A theory of fairness, competition and cooperation. *Quarterly Journal of Economics* 114:817–68. [KS]

Fey, M., McKelvey, R. D. & Palfrey, T. R. (1996) An experimental study of constant-sum Centipede games. *International Journal of Game Theory* 25:269–87. [aAMC, WDC]

Fishburn, P. C. (1988) Normative theories of decision making under risk and under uncertainty. In: *Decision making: Descriptive, normative, and prescriptive interactions,* ed. D. E. Bell, H. Raiffa, & A. Tversky. Cambridge University Press. [aAMC]

Flood, M. M. (1954) Environmental non-stationarity in a sequential decision-making experiment. In: *Decision processes,* ed. R. M. Thrall, C. H. Coombs & R. L. Davis. Wiley. [MJ]

Foddy, M., Smithson, M., Schneider, S. & Hogg, M. eds. (1999) *Resolving social dilemmas: Dynamic, structural, and intergroup aspects.* Psychology Press. [aAMC]

Føllesdal, D. (1982) The status of rationality assumptions in interpretation and in the explanation of action. *Dialectica* 36:301–16. [rAMC]

Forsythe, R., Horowitz, J., Savin, N. & Sefton, M. (1994) Replicability, fairness and play in experiments with simple bargaining games. *Games and Economic Behavior* 6:347–69. [RS]

Frank, R. H. (1988) *Passions within reason: The strategic role of the emotions.* W. W. Norton. [WDC, JL, RS]

Frank, R. H., Gilovitch, T. & Regan, D. (1993) The evolution of one-shot cooperation: An experiment. *Ethology and Sociobiology* 14:247–56. [PB]

Freud, S. (1911) Formulations on the two principles of mental functioning. In: *The standard edition of the complete psychological works of Sigmund Freud, vol. 12,* ed. and trans. J. Strachey. Hogarth Press. [aAMC]

Friedman, J. W. (1991) *Game theory with applications to economics, 2nd edition.* Oxford University Press. [aAMC]

Friedman, J. W., ed. (1996) *The rational choice controversy: Economic models of politics reconsidered.* Yale University Press. [aAMC]

Frisch, D. & Clemen, R. T. (1994) Beyond expected utility: Rethinking behavioral decision research. *Psychological Bulletin* 116:46–54. [aAMC]

Garner, W. R. (1954) Context effects and the validity of loudness scales. *Journal of Experimental Psychology* 48:218–24. [IV]

Gauthier, D. (1975) Coordination. *Dialogue* 14:195–221. [aAMC, MCJ]

Geanakoplos, J., Pearce, D. & Stacchetti, E. (1989) Psychological games and sequential rationality. *Games and Economic Behavior* 1:60–79. [arAMC, JPC]

Gentner, D., Holyoak, K. & Kokinov, B., eds. (2001) *The analogical mind.* MIT Press. [BK]

Gibbard, A. (1992) Weakly self-ratifying strategies: Comments on McClennen. *Philosophical Studies* 65:217–25. [PW]

Gigerenzer, G. & Goldstein, D. G. (1996) Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* 103:650–69. [aAMC]

Gigerenzer, G., Todd, P. M. & the ABC Research Group (1999) *Simple heuristics that make us smart.* Oxford University Press. [aAMC]

Gilbert, M. (1987) Modelling collective belief. *Synthese* 73:185–204. [aAMC]

(1989a) Folk psychology takes sociality seriously. *Behavioral and Brain Sciences* 12:707–708. [aAMC]

(1989b) Rationality and salience. *Philosophical Studies* 57:61–77. [arAMC]

(1990) Rationality, coordination and convention. *Synthese* 84:1–21. [aAMC]

(2000) Collective preferences, obligations, and rational choice. *Economics and Philosophy* 17:109–19. [aAMC]

Gillies, D. B. (1953) Some theorems on n-person games. Princeton University. (Unpublished doctoral dissertation.) [aAMC]

Gintis, H. (2000) *Game theory evolving: A problem-centered introduction to modeling strategic behavior.* Princeton University Press. [PB, HG, KS]

(2003) Solving the puzzle of human prosociality. *Rationality and Society* 15(2):155–87. [HG]

Glimcher, P. W. (2002) Decisions, decisions, decisions: Choosing a biological science of choice. *Neuron* 36:323–32. [GSB]

Gold, J. I. & Shadlen, M. N. (2001) Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences* 5(1):10–16. [GSB]

Goltz, S. M. (1993) Examining the joint roles of responsibility and reinforcement history in recommitment. *Decision Sciences* 24:977–94. [EF]

(1999) Can't stop on a dime: The roles of matching and momentum in persistence of commitment. *Journal of Organizational Behavior Management* 19:37–63. [EF]

Good, D. A. (1991) Cooperation in a microcosm: Lessons from laboratory games. In: *Cooperation and prosocial behavior,* ed. R. A. Hinde & J. Groebel. Cambridge University Press. [aAMC]

Goodie, A. S. & Fantino, E. (1995) An experientially derived base-rate error in humans. *Psychological Science* 6:101–106. [EF]

(1996) Learning to commit or avoid the base-rate error. *Nature* 380:247–49. [EF]

(1999) What does and does not alleviate base-rate neglect under direct experience. *Journal of Behavioral Decision Making* 12:307–35. [EF]

Goyal, S. & Janssen, M. (1996) Can we rationally learn to coordinate? *Theory and Decision* 40:29–49. [HH]

Green, D. P. & Shapiro, I. (1994) *Pathologies of rational choice theory: A critique of applications in political science.* Yale University Press. [aAMC]

Grzelak, J. (1988) Conflict and cooperation. In: *Introduction to social psychology: A European perspective,* ed. M. Hewstone, W. Stroebe, J.-P. Codol & G. M. Stephenson. Basil Blackwell. [aAMC]

Güth, W. (1995) On ultimatum bargaining experiments – A personal review. *Journal of Economic Behavior and Organization* 27:329–44. [MP]

Güth, W., Ockenfels, P. & Wendel, M. (1997) Cooperation based on trust: An experimental investigation. *Journal of Economic Psychology* 18:15–43. [aAMC]

Halford, G. S. (1993) *Children's understanding: The development of mental models.* Erlbaum. [BK]

Haller, H. (2000) Non-additive beliefs in solvable games. *Theory and Decision* 49:313–38. [HH]

Hamburger, H. (1973) *N*-person Prisoner's Dilemma. *Journal of Mathematical Sociology* 3:27–48. [aAMC]

Hamilton, W. D. (1964) The genetical evolution of social behavior. *Journal of Theoretical Biology* 7:1–52. [DJB, rAMC, PJBH]

Hancock, P. J. B. & Ross, K. (2002) What's a pretty face worth? (II): Factors affecting offer levels in the dictator game. Paper presented at the Human Behavior and Evolution Society Meeting, New Brunswick, NJ. [PJBH]

Hardin, R. (2002) *Trust and trustworthiness.* Russel Sage. [JL]

Harless, D. W. & Camerer, C. F. (1994) The predictive utility of generalized expected utility theories. *Econometrica* 62:1251–89. [aAMC]

Harper, W. (1991) Ratifiability and refinements. In: *Foundations of decision theory,* ed. M. Bacharach & S. Hurley. Blackwell. [PW]

(1999) Solutions based on ratifiability and sure thing reasoning. In: *The logic of strategy,* ed. C. Bicchieri, R. Jeffrey & B. Skyrms. Oxford University Press. [PW]

Harsanyi, J. C. (1962) Rational postulates for bargaining solutions in cooperative and non-cooperative games. *Management Science* 9:197–219. [aAMC]

(1966) A general theory of rational behavior in game situations. *Econometrica* 34:613–34. [aAMC]

(1967–1968) Games with incomplete information played by "Bayesian" players, Parts I-III. *Management Science.* 14:159–82, 320–34, 486–502. [aAMC]

Harsanyi, J. C. & Selten, R. (1988) *A general theory of equilibrium selection in games.* MIT Press. [arAMC, HH]

Hartl, J. A. & Fantino, E. (1996) Choice as a function of reinforcement ratios in delayed matching to sample. *Journal of the Experimental Analysis of Behavior* 66:11–27. [EF]

Hausman, D. (1992) *The inexact and separate science of economics.* Cambridge University Press. [DMH]
Hedden, T. & Zhang, J. (2002) What do you think I think you think? Strategic reasoning in matrix games. *Cognition* 85:1–36. [rAMC, MJ]
Heider, F. (1958) *The psychology of interpersonal relations.* Wiley. [rAMC]
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H. & McElreath, R. (2001) In search of Homo economicus: Behavioral experiments from 15 small-scale societies. *American Economic Review* 91:73–78. [PB]
Herrnstein, R. J. (1990) Rational choice theory. *American Psychologist* 45:356–67. [aAMC]
Hey, J. D. & Orme, C. (1994) Investigating generalizations of expected utility theory using experimental data. *Econometrica* 62:1291–1326. [aAMC]
Heyes, C. & Dickinson, A. (1993) The intentionality of animal action. In: *Consciousness,* ed. M. Davies & G. Humphreys. Blackwell. [SH]
Hill, K. (2002) Altruistic cooperation during foraging by the Ache, and the evolved predisposition to cooperate. *Human Nature* 13:105–28. [MA]
Hinde, R. A. (1987) *Individuals, relationships and culture: Links between ethology and the social sciences.* Cambridge University Press. [JL]
Hofbauer, J. & Sigmund, K. (1998) *Evolutionary games and population dynamics.* Cambridge University Press. [arAMC, KS]
Hofstadter, D. R. (1983) Metamagical themas: Computer tournaments of the Prisoner's Dilemma suggest how cooperation evolves. *Scientific American* 248(5):14–20. [aAMC]
Holland, J. H. (1996) The rationality of adaptive agents. In: *The rational foundations of economic behaviour: Proceedings of the IEA Conference, Turin, Italy,* pp. 281–97, ed. K. J. Arrow, E. Colombatto, M. Perlman, & C. Schmidt. Macmillan. [rAMC]
Hollis, M. (1987) *The cunning reason.* Cambridge University Press. [aAMC]
Hollis, M. & Sugden, R. (1993) Rationality in action. *Mind* 102:1–35. [DJB, aAMC]
Horgan, T. (1981) Counterfactuals and Newcomb's problem. *Journal of Philosophy* 78:331–56. [aAMC]
Houston, A. I. & McNamara, J. M. (1989) The value of food – Effects of open and closed economies. *Animal Behaviour* 37:546–62. [MS]
(1999) *Models of adaptive behaviour: An approach based on state.* Cambridge University Press. [MS]
Howard, J. (1988) Cooperation in the prisoner's dilemma. *Theory and Decision* 24:203–13. [SH]
Huber, J., Payne, J. W. & Puto, C. (1982) Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research* 9:90–98. [aAMC]
Hume, D. (1739–40/1978) *A treatise of human nature, 2nd edition,* ed. L. A. Selby-Bigge. Oxford University Press. (Original work published 1739–1740.) [aAMC, AR]
Hurley, S. (1989) *Natural reasons.* Oxford University Press. [SH]
(1991) Newcomb's problem, prisoners' dilemma, and collective action. *Synthese* 86:173–96. [SH]
(1994) A new take from Nozick on Newcomb's problem and prisoners' dilemma. *Analysis* 54:65–72. [SH]
(2003) Animal action in the space of reasons. *Mind and Language* 18(3):231–56). [SH]
Hutchins, E. (1995) *Cognition in the wild.* MIT Press. [SH]
Janssen, M. C. W. (2001a) On the principle of coordination. *Economics and Philosophy* 17:221–34. [MCWJ]
(2001b) Rationalizing focal points. *Theory and Decision* 50:119–48. [aAMC, HH, MCWJ]
Jeffrey, R. (1983) *The logic of decision, 2nd edition.* Chicago University Press. [aAMC, PW]
Joergensen, P. R. & Hancock, P. J. B. (2001) What's a pretty face worth?: Factors affecting offer levels in the ultimatum game. Paper presented at the Human Behavior and Evolution Society Meeting, London. [PJBH]
Jones, M. & Seick, W. R. (in press) Learning myopia: An adaptive recency effect in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* [MJ]
Joyce, J. M. (1999) *The foundations of causal decision theory.* Cambridge University Press. [AR, PW]
Kagel, J. H. & Roth, A. E. eds. (1995) *Handbook of experimental economics.* Princeton University Press. [aAMC]
Kahneman, D., Slovic, P. & Tversky, A., eds. (1982) *Judgment under uncertainty: Heuristics and biases.* Cambridge University Press. [arAMC]
Kahneman, D. & Tversky, A. (1973) On the psychology of prediction. *Psychological Review* 80:237–51. [EF]
(1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47:263–91. [rAMC, KS]
Kahneman, D. & Tversky, A., ed. (2000) *Choices, values, and frames.* Cambridge University Press. [rAMC]
Kaplan, H. & Hill, K. (1985) Food sharing among Ache foragers. Tests of explanatory hypotheses. *Current Anthropology* 26:223–45. [JL]
Kavka, G. (1983) The toxin puzzle. *Analysis* 43:33–36. [RMS]
Kelley, H. H., Holmes, J. W., Kerr, N. L., Reis, H. T., Rusbult, C. E. & Van Lange, P. A. M. (2003) *An atlas of interpersonal situations.* Cambridge University Press. [PAMVL]
Kelley, H. H. & Stahelski, A. J. (1970) Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology* 16:66–91. [PAMVL]
Kelley, H. H. & Thibaut, J. W. (1978) *Interpersonal relations: A theory of interdependence.* Wiley. [PAMVL]
Kelley, H. H., Thibaut, J. W., Radloff, R. & Mundy, D. (1962) The development of cooperation in the "minimal social situation." *Psychological Monographs* 76 (Whole No. 19). [rAMC]
Kelly, R. L. (1995) *The foraging spectrum: Diversity in hunter-gatherer lifeways.* Smithsonian Institution Press. [JL]
Kerr, B., Riley, M. A., Feldman, M. W. & Bohannan, B. J. M. (2002) Local dispersion promotes biodiversity in a real game of rock-paper-scissors. *Nature* 418:171–74. [KS]
Koehler, J. J. (1996) The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences* 19(1):1–53. [EF]
Kohlberg, L., ed. (1984) *The psychology of moral development: The nature and validity of moral stages.* Harper & Row. [JL]
Kokinov, B. (1992) Inference evaluation in deductive, inductive and analogical reasoning. In: *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society,* pp. 903–908. Erlbaum. [BK]
Kokinov, B. & Petrov, A. (2001) Integration of memory and reasoning in analogy-making: The AMBR model. In: *The analogical mind,* ed. D. Gentner, K. Holyoak & B. Kokinov. MIT Press. [BK]
Kraines, D. & Kraines, V. (1995) Evolution of learning among Pavlov strategies in a competitive environment with noise. *Journal of Conflict Resolution* 39:439–66. [arAMC]
Kramarz, F. (1996) Dynamic focal points in N-person coordination games. *Theory and Decision* 40:277–313. [HH]
Kramer, R. M. & Tyler, T. R. (1996) *Trust in organisations: Frontiers of theory and research.* Sage. [JL]
Krebs, J. R. & Davies, N. B. (1987) *An introduction to behavioural ecology, 2nd edition.* Basil Blackwell. [aAMC]
Kreps, D. (1988) *Notes on the theory of choice.* Westview. [aAMC]
(1990) *A course in microeconomic theory.* Princeton University Press. [aAMC]
Kreps, D. M. & Wilson, R. (1982a) Reputation and imperfect information. *Journal of Economic Theory* 27:253–79. [aAMC]
(1982b) Sequential equilibria. *Econometrica* 50:863–94. [rAMC]
Krueger, J. (1998) On the perception of social consensus. *Advances in Experimental Social Psychology* 30:163–240. [JIK]
Kyburg, H. E., Jr. (1987) Bayesian and non-Bayesian evidential updating. *Artificial Intelligence* 31:271–93. [rAMC]
Laming, D. R. J. (1997) *The measurement of sensation.* Oxford University Press. [IV]
Lazarus, J. (1995) Behavioural ecology and evolution. In: *Biological aspects of behavior,* ed. D. Kimble & A. M. Colman. Longman. [aAMC]
Lazarus, J. & Inglis, I. R. (1986) Shared and unshared parental investment, parent-offspring conflict and brood size. *Animal Behaviour* 34:1791–1804. [JL]
Lea, S. E. G., Tarpy, R. M. & Webley, P. (1987) *The individual in the economy.* Cambridge University Press. [aAMC]
Ledyard, J. O. (1995) Public goods: A survey of experimental research. In: *Handbook of experimental economics,* ed. J. H. Kagel & A. E. Roth. Princeton University Press. [aAMC]
Levi, I. (1986) *Hard choices.* Cambridge University Press. [rAMC]
Lewis, D. K. (1969) *Convention: A philosophical study.* Harvard University Press. [aAMC]
(1979) Prisoner's dilemma is a Newcomb problem. *Philosophy and Public Affairs* 8:235–40. [rAMC, JIK]
Lockhead, G. (1995) Psychophysical scaling methods reveal and measure context effects. *Behavioral and Brain Sciences.* 18(3):607–12. [IV]
Loewenstein, G., Weber, E., Hsee, C. & Welch, N. (2001) Risk as feelings. *Psychological Bulletin* 127:267–86. [DJB]
Logue, A. W. (1988) Research on self-control: An integrating framework. *Behavioral and Brain Sciences* 11:665–709. [EF]
Luce, R. D. & Raiffa, H. (1957) *Games and decisions: Introduction and critical survey.* Wiley/Dover. [aAMC, WDC, AR, RS]
Luce, R. D. & von Winterfeldt, D. (1994) What common ground exists for descriptive, prescriptive, and normative utility theories? *Management Science* 40:263–79. [AR]
Lumsden, M. (1973) The Cyprus conflict as a Prisoner's Dilemma Game. *Journal of Conflict Resolution* 17:7–32. [aAMC]
Machina, M. (1987) Choice under uncertainty: Problems solved and unsolved. *Economic Perspectives* 1:121–54. [aAMC]
(1991) Dynamic consistency and non-expected utility. In: *Foundations of decision theory,* ed. M. Bacharach & S. Hurley. Blackwell. [arAMC]

Manktelow, K. I. & Over, D. E. (1993) Introduction: The study of rationality. In: *Rationality: Psychological and philosophical perspectives,* ed. K. I. Manktelow & D. E. Over. Routledge. [arAMC]
Margolis, H. (1982) *Selfishness, altruism, and rationality: A theory of social choice.* Cambridge University Press. [JL]
Markman, A. & Moreau, C. (2001) Analogy and analogical comparison in choice. In: *The analogical mind,* ed. D. Gentner, K. Holyoak & B. Kokinov. MIT Press. [BK]
Maynard Smith, J. (1976) Evolution and the theory of games. *American Scientist* 64(1):41–45. [aAMC]
(1984) Game theory and the evolution of behavior. *Behavioral and Brain Sciences* 7:95–101. [aAMC]
Maynard Smith, J. & Price, G. R. (1973) The logic of animal conflict. *Nature* 246:15–18. [aAMC]
McCabe, K., Houser, D., Ryan, L., Smith, V. & Trouard, T. (2001) A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences, USA* 98(20):11832–35. [GSB]
McClennen, E. F. (1983) Sure-thing doubts. In: *Foundations of utility and risk theory with applications,* ed. B. F. Stigum & F. Wenstøp. Reidel. [aAMC]
(1985) Prisoner's dilemma and resolute choice. In: *Paradoxes of rationality and cooperation: Prisoner's Dilemma and Newcomb's problem,* ed. R. Campbell & L. Sowden. University of British Columbia Press. [rAMC]
(1990) *Rationality and dynamic choice.* Cambridge University Press. [arAMC]
(1992) The theory of rationality for ideal games. *Philosophical Studies* 65:193–215. [aAMC, PW]
McClintock, C. G. & Liebrand, W. B. G. (1988) The role of interdependent structure, individual value orientation, and another's strategy in social decision making: A transformational analysis. *Journal of Personality and Social Psychology* 55:396–409. [aAMC]
McElreath, R., Boyd, R. & Richerson, P. (in press) Shared norms can lead to the evolution of ethnic markers. *Current Anthropology* 44:122–29. [MA]
McKelvey, R. D. & Palfrey, T. R. (1992) An experimental study of the Centipede game. *Econometrica* 60:803–36. [WDC, aAMC]
McNamara, J. M. (1996) Risk-prone behaviour under rules which have evolved in a changing environment. *American Zoologist* 36:484–95. [MS]
Mehta, J., Starmer, C., & Sugden, R. (1994a) Focal points in pure coordination games: An experimental investigation. *Theory and Decision* 36:163–85. [aAMC, MCJ]
(1994b) The nature of salience: An experimental investigation in pure coordination games. *American Economic Review* 84:658–73. [aAMC, MCWJ]
Mesterton-Gibbons, M. & Dugatkin, L. A. (1992) Cooperation among unrelated individuals: Evolutionary factors. *The Quarterly Review of Biology* 67:267–81. [RS]
Milgrom, P. (1981) An axiomatic characterization of common knowledge. *Econometrica* 49: 219–22. [aAMC]
Milgrom, P. & Roberts, J. (1982) Predation, reputation, and entry deterrence. *Journal of Economic Theory* 27:280–312. [aAMC]
Milinski, M., Semmann, D. & Krambeck, H.-J. (2002) Reputation helps solve the "tragedy of the commons." *Nature* 415:424–26. [JL]
Miller, G. & Isard, S. (1964) Free recall of self-embedded English sentences. *Information and Control* 7:293–303. [rAMC]
Monderer, D. & Samet, D. (1989) Approximating common knowledge with common beliefs. *Games and Economic Behavior* 1:170–90. [aAMC]
Montague, P. R. & Berns, G. S. (2002) Neural economics and the biological substrates of valuation. *Neuron* 36:265–84. [GSB]
Montague, P. R., Berns, G. S., Cohen, J. D., McClure, S. M., Pagnoni, G., Dhamala, M., Wiest, M. C., Karpov, I., King, R. D., Apple, N. & Fisher, R. E. (2002) Hyperscanning: Simultaneous fMRI during linked social interactions. *NeuroImage* 16:1159–64. [GSB]
Monterosso, J., Ainslie, G., Mullen, P. & Gault, B. (2002) The fragility of cooperation: An empirical study employing false-feedback in a sequential iterated prisoner's dilemma. *Journal of Economic Psychology* 23:437–48. [JM]
Moser, P. K., ed. (1990) *Rationality in action: Contemporary approaches.* Cambridge University Press. [aAMC]
Nash, J. F. (1950a) Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences, USA* 36:48–49. [aAMC]
(1950b) The bargaining problem. *Econometrica* 18:155–62. [aAMC]
(1951) Non-cooperative games. *Annals of Mathematics* 54:286–95. [arAMC]
(1953) Two-person cooperative games. *Econometrica* 21:128–40. [aAMC]
(2002) Non-cooperative games. In: *The essential John Nash,* ed. H. W. Kuhn & S. Nasar. Princeton University Press. [rAMC]
Noë, R. (1990) A veto game played by baboons: A challenge to the use of the Prisoner's Dilemma as a paradigm for reciprocity and cooperation. *Animal Behaviour* 39:78–90. [RS]
Nowak, M. A., May, R. M. & Sigmund, K. (1995) The arithmetics of mutual help. *Scientific American* 1995(6):76–81. [aAMC]
Nowak, M. A., Page, K. M. & Sigmund, K. (2000) Fairness versus reason in the ultimatum game. *Science* 289:1773–75. [PB, KS]
Nowak, M. A. & Sigmund, K. (1993) A strategy of win-stay, lose-shift that outperforms tit-for-tat in the Prisoner's Dilemma game. *Nature* 364:56–58. [aAMC]
Nozick, R. (1969) Newcomb's problem and two principles of choice. In:*Essays in honor of Carl G. Hempel: A tribute on his sixty-fifth birthday,* ed. N. Rescher. Reidel. [aAMC, RMS]
(1993) *The nature of rationality.* Princeton University Press. [arAMC]
Oaksford, M. & Chater, N. (1996) Rational explanation of the selection task. *Psychological Review* 103:381–91. [MJ]
Ordeshook, P. C. (1986) *Game theory and political theory: An introduction.* Cambridge University Press. [aAMC]
Ostrom, E. (1990) *Governing the commons: The evolution of institutions for collective action.* Cambridge University Press. [JL]
Packer, C., Scheel, D. & Pusey, A. E. (1990) Why lions form groups: Food is not enough. *American Naturalist* 136:1–19. [RS]
Page, K. M. & Nowak, M. A. (2002) Empathy leads to fairness. *Bulletin of Mathematical Biology* 64:1101–16. [KS]
Page, K. M., Nowak, M. & Sigmund, K. (2002) The spatial ultimatum game. *Proceedings of the Royal Society (London) B* 267:2177–82. [KS]
Palameta, B. & Brown, W. M. (1999) Human cooperation is more than by-product mutualism. *Animal Behaviour* 57:F1-F3. [RS]
Parco, J. E., Rapoport, A. & Stein, W. E. (2002) The effects of financial incentives on the breakdown of mutual trust. *Psychological Science* 13:292–97. [WDC, rAMC]
Parfit, D. (1979) Is common-sense morality self-defeating? *Journal of Philosophy* 76:533–45. [arAMC]
(1984) *Reasons and persons.* Clarendon Press. [aAMC]
Park, J. R. & Colman, A. M. (2001) Team reasoning: An experimental investigation. Paper presented at the Fifth Conference of the Society for the Advancement of Economic Theory, Ischia, 2–8 July, 2001. [aAMC]
Pearce, D. G. (1984) Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52:1029–50. [arAMC]
Penton-Voak, I. S., Perrett, I. D. & Pierce, J. W. (1999) Computer graphic studies of the role of facial similarity in judgements of attractiveness. *Current Psychology* 18:104–17. [PJH]
Pettit, P. & Sugden, R. (1989) The backward induction paradox. *Journal of Philosophy* 86:169–82. [aAMC]
Povinelli, D. (1996) Chimpanzee theory of mind? In: *Theories of theories of mind,* ed. P. Carruthers & P. K. Smith. Cambridge University Press. [SH]
Price, M., Cosmides, L. & Tooby, J. (2002) Punitive sentiment as an anti-free rider psychological device. *Evolution and Human Behavior* 23:203–31. [DJB, rAMC]
Pruitt, D. G. & Kimmel, M. J. (1977) Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology* 28:363–92. [aAMC]
Poundstone, W. (1992) *Prisoner's Dilemma.* Oxford University Press. [aAMC, AR]
Quattrone, G. A. & Tversky, A. (1984) Causal versus diagnostic contingencies: On self-deception and the voter's illusion. *Journal of Personality and Social Psychology* 46:237–248. [aAMC]
Quine, W. V. (1962) Paradox. *Scientific American* 206(4):84–96. [rAMC]
Rabbie, J. M. (1991) Determinants of instrumental intra-group cooperation. In: *Cooperation and prosocial behaviour,* ed. R. A. Hinde & J. Groebel. Cambridge University Press. [JL]
Rabin, M. (1993) Incorporating fairness into game theory and economics. *American Economic Review* 83:1281–1302. [arAMC, JPC]
Rachlin, H. (2000) *The science of self-control.* Harvard University Press. [rAMC]
(2002) Altruism and selfishness. *Behavioral and Brain Sciences* 25(2):239–50. [EF, JIK]
Rachlin, H. & Green, L. (1972) Commitment, choice and self-control. *Journal of the Experimental Analysis of Behavior* 17:15–22. [rAMC]
Raiffa, H. (1992) Game theory at the University of Michigan, 1948–1952. In: *Toward a history of game theory,* ed. E. R. Weintraub. Duke University Press. [aAMC]
Ramsey, F. P. (1931) Truth and probability. In: *The foundations of mathematics and other logical essays,* ed. R. B. Braithwaite. Routledge and Kegan Paul. [aAMC]
Rapoport, A. (1962) The use and misuse of game theory. *Scientific American* 207(6):108–18. [aAMC]
(1988) Experiments with *n*-person social traps I. Prisoner's dilemma, weak prisoner's dilemma, volunteer's dilemma and largest number. *Journal of Conflict Resolution* 32:457–72. [AR]
(1989) *Decision theory and decision behavior: Normative and descriptive approaches.* Kluwer. [aAMC]
Rapoport, A. & Chammah, A. M. (1965) *Prisoner's Dilemma: A study in conflict and cooperation.* University of Michigan Press. [arAMC, IV]

Rapoport, A., Seale, D. A. & Parco, J. E. (2002) Coordination in the aggregate without common knowledge or outcome information. In: *Experimental business research,* ed. R. Zwick & A. Rapoport. Kluwer. [WDC]
Rawls, J. (1999) *A theory of justice,* revised edition. Oxford University Press. [JL]
Raz, J. (2000) *Engaging reason: On the theory of value and action.* Oxford University Press. [aAMC]
Regan, D. (1980) *Utilitarianism and co-operation.* Clarendon. [rAMC, SH]
Rescher, N. (1975) *Unselfishness: The role of the vicarious affects in moral philosophy and social theory.* University of Pittsburgh Press. [rAMC]
Richerson, P. & Boyd, R. (2001) Built for speed, not for comfort: Darwinian theory and human culture. *History and Philosophy of the Life Sciences* 23:423–63. [MA]
Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S. & Kilts, C. D. (2002) A neural basis for social cooperation. *Neuron* 35:1–20. [GSB]
Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G. & Kilts, C. D. (2002) A neural basis for cooperation. *Neuron* 35:395–405. [DJB]
Robbins, T. W. & Everitt, B. J. (1992) Functions of dopamine in the dorsal and ventral striatum. *Seminars in Neuroscience* 4:119–27. [GSB]
Roberts, G. (1997) Testing mutualism: A commentary on Clements & Stephens. *Animal Behaviour* 53:1361–62.
Roberts, G. & Sherratt, T. N. (1998) Development of cooperative relationships through increasing investment. *Nature* 394:175–79. [JL]
Rosenthal, R. W. (1981) Games of perfect information, predatory pricing and the chain-store paradox. *Journal of Economic Theory* 25:92–100. [aAMC]
Roth, A. E. (1995) Bargaining experiments. In: *Handbook of experimental economics,* ed. J. H. Kagel & A. E. Roth. Princeton University Press. [PB]
Rusbult, C. E. & Van Lange, P. A. M. (2003) Interdependence, interaction, and relationships. *Annual Review of Psychology* 54:351–75. [PAMVL]
Russell, B. (1954) *Human society in ethics and politics.* George Allen & Unwin. [aAMC]
Samet, D. (1996) Hypothetical knowledge and games with perfect information. *Games and Economic Behavior* 17:230–51. [aAMC]
Samuelson, L. (1992) Dominated strategies and common knowledge. *Games and Economic Behavior* 4:284–313. [aAMC]
(1997) *Evolutionary games and equilibrium selection.* MIT Press. [aAMC]
Savage, L. J. (1951) The theory of statistical decision. *Journal of the American statistics Association* 46:55–67. [aAMC]
(1954) *The foundations of statistics.* Wiley. (2nd edition, 1972.) [aAMC]
Schacter, D. L. (1999) The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist* 54(3):182–203. [MJ]
Scharlemann, J. P., Eckel, C. C., Kacelnik, A. & Wilson, R. K. (2001) The value of a smile: Game theory with a human face. *Journal of Economic Psychology* 22:617–40. [PJBH]
Schelling, T. C. (1960) *The strategy of conflict.* Harvard University Press. [aAMC, MCWJ]
(1973) Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *Journal of Conflict Resolution* 17:381–428. [aAMC]
Schillo, M., Funk, P. & Rovatsos, M. (2000) Using trust for detecting deceitful agents in artificial societies. *Applied Artificial Intelligence* 14:825–48. [JL]
Schroeder, D. A. (1995) *Social dilemmas: Perspectives on individuals and groups.* Praeger. [aAMC]
Schultz, W., Dayan, P. & Montague, P. R. (1997) A neural substrate of prediction and reward. *Science* 275:1593–99. [GSB]
Schuster, R. (2001) An animal model of cooperating dyads: Methodological and theoretical issues. *Mexican Journal of Behavior Analysis* 27:165–200. [RS]
(2002) Cooperative coordination as a social behavior: Experiments with an animal model. *Human Nature* 13:47–83. [RS]
Schuster, R., Berger, B. D. & Swanson, H. H. (1993) Cooperative social coordination and aggression. II. Effects of sex and housing among three strains of intact laboratory rats differing in aggressiveness. *Quarterly Journal of Experimental Psychology* 46B:367–90. [RS]
Selten, R. (1965) Spieltheoretische behandlung eines Oligopolmodells mit Nachfrageträgheit. *Zeitschrift für die gesamte Staatswissenschaft* 121:301–24, 667–89. [aAMC]
(1975) Re-examination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory* 4:25–55. [aAMC]
(1978) The chain store paradox. *Theory and Decision* 9:127–59. [aAMC, AR]
Selten, R. & Stoecker, R. (1986) End behavior in sequences of finite Prisoner's Dilemma supergames: A learning theory approach. *Journal of Economic Behavior and Organization* 7:47–70. [aAMC]
Sen, A. K. (1978) Rational fools: A critique of the behavioral foundations of economic theory. In: *Scientific models and men,* ed. H. Harris. Oxford University Press. [aAMC]
(1985) Rationality and uncertainty. *Theory and Decision* 18:109–27. [rAMC]
Sethi, R. & Somanathan, E. (2003) Understanding reciprocity. *Journal of Economic Behavior and Organization* 50:27. [DJB]
Shafir, E. (1993) Intuitions about rationality and cognition. In: *Rationality: Psychological and philosophical perspectives,* ed. K. I. Manktelow & D. E. Over. Routledge. [aAMC]
Shin, H. (1991) Two notions of ratifiability and equilibrium in games. In: *Foundations of decision theory,* ed. M. Bacharach & S. Hurley. Blackwell. [PW]
Simon, H. A. (1957) *Models of man: Social and rational.* Wiley. [aAMC]
Sinervo, B. & Lively, C. M. (1996) The rock-scissors-paper game and the evolution of alternative male strategies. *Nature* 340:240–43. [KS]
Slovic, P. & Lichtenstein, S. (1983) Preference reversals: A broader perspective. *American Economic Review* 73:596–605. [aAMC]
Smith, A. (1910) *The wealth of nations.* Dutton. (Original work published 1776). [aAMC]
Snow, P. (1994) Ignorance and the expressiveness of single- and set-valued probability models of belief. In: *Uncertainty in artificial intelligence: Proceedings of the Tenth Conference (UAI-1994),* pp. 531–37, ed. R. L. de Mantras & D. L. Poole. Morgan Kaufmann. [rAMC]
Sober, E. & Wilson, D. S. (1998) *Unto others: The evolution and psychology of unselfish behavior.* Harvard University Press. [RS]
Solnick, S. J. & Schweitzer, M. E. (1999) The influence of physical attractiveness and gender on ultimatum game decisions. *Organizational Behavior and Human Decision Processes* 793:199–215. [PJH]
Sosa, E. & Galloway, D. (2000) Man the rational animal? *Synthese* 122:165–78. [aAMC]
Spohn, W. (2001) Dependency equilibria and the causal structure of decision and game situations. Forschungsberichte der DFG-Forschergruppe; *Logik in der Philosophie* 56:645. Also to appear in: *Homo Oeconomicus* (2003). [rAMC, RMS]
Squires, D. (1998) Impossibility theorems for normal form games. *Theory and Decision* 44:67–81. [aAMC]
Staddon, J. E. R. (1983) *Adaptive behavior and learning.* Cambridge University Press. [RS]
Stahl, D. O. & Wilson, P. W. (1995) On players' models of other players: Theory and experimental evidence. *Games and Economic Behavior* 10:218–54. [arAMC]
Stander, P. E. (1992) Cooperative hunting in lions: The role of the individual. *Behavioral Ecology and Sociobiology* 29:445–54. [RS]
Stanovich, K. E. & West, R. F. (2000) Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences* 23:645–726. [aAMC]
Starmer, C. (2000) Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature* 38:332–82. [DJB, aAMC]
Stein E. (1996) *Without good reason: The rationality debate in philosophy and cognitive science.* Oxford University Press. [aAMC]
Stewart et al. (in press) Prospect relativity: How choice options influence decision under risk. *Journal of Experimental Psychology.* [IV]
Stolarz-Fantino, S. & Fantino, E. (1990) Cognition and behavior analysis: A review of Rachlin's *Judgment, decision, and choice. Journal of the Experimental Analysis of Behavior* 54:317–22. [EF]
(1995) The experimental analysis of reasoning: A review of Gilovich's *How we know what isn't so. Journal of the Experimental Analysis of Behavior* 64:111–16. [EF]
Stolarz-Fantino, S., Fantino, E. & Kulik, J. (1996) The conjunction fallacy: Differential incidence as a function of descriptive frames and educational context. *Contemporary Educational Psychology* 21:208–18. [EF]
Stolarz-Fantino, S., Fantino, E., Zizzo, D. & Wen, J. (2003) The conjunction effect: New evidence for robustness. *American Journal of Psychology* 116(1):15–34. [EF]
Sugden, R. (1991a) Rational bargaining. In: *Foundations of decision theory,* ed. M. Bacharach & S. Hurley. Blackwell. [aAMC]
(1991b) Rational choice: A survey of contributions from economics and philosophy. *Economic Journal* 101:751–85. [aAMC]
(1992) Inductive reasoning in games. In: *Rational interaction: Essays in honor of John C. Harsanyi,* ed. R. Selten. Springer-Verlag. [aAMC]
(1993) Thinking as a team: Towards an explanation of nonselfish behavior. *Social Philosophy and Policy* 10:69–89. [aAMC]
(1995) Towards a theory of focal points. *Economic Journal* 105:533–50. [aAMC, MCWJ]
(2000) Team preferences. *Economics and Philosophy* 16:175–204. [aAMC, MCJ]
Taylor, M. (1987) *The possibility of cooperation.* Cambridge University Press. [JL]
(1996) When rationality fails. In: *The rational choice controversy: Economic models of politics reconsidered,* ed. J. Friedman. Yale University Press. [aAMC]
Thaler, R. H. (1988) The Ultimatum Game. *The Journal of Economic Perspectives* 24:195–206. [PJH, MP]
(1992) *The winner's curse: Paradoxes and anomalies of economic life.* Princeton University Press. [aAMC]

Thomas, J. P. & McFadyen, R. G. (1995) The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology* 16:97–113. [rAMC]
Tomasello, M. (1999) *The cultural origins of human cognition.* Harvard University Press. [MA]
Tooby, J. & Cosmides, L. (1992) The psychological foundations of culture. In: *The adapted mind: Evolutionary psychology and the generation of culture,* ed. J. H. Barkow, L. Cosmides & J. Tooby. Oxford University Press. [JL]
Trivers, R. (1971) The evolution of reciprocal altruism. *Quarterly Review of Biology* 46:35–57. [DJB, rAMC]
(1985) *Social evolution.* Benjamin/Cummings. [DJB]
Tversky, A. (1969) Intransitivity of preferences. *Psychological Review* 76:31–48. [aAMC]
(1996) Rational theory and constructive choice. In: *The rational foundations of economic behaviour: Proceedings of the IEA Conference, Turin, Italy,* pp. 185–97, ed. K. J. Arrow, E. Colombatto, M. Perlman & C. Schmidt. Macmillan. [rAMC]
Tversky, A. & Kahneman, D. (1992) Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5:297–323. [rAMC]
van den Doel, H. & van Velthoven, B. (1993) *Democracy and welfare economics, 2nd edition.* Cambridge University Press. [JL]
van Huyck, J., Battalio, R., & Beil, R. (1990) Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review* 80:234–48. [MA, aAMC]
Van Lange, P. A. M. (1999) The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology* 77:337–49. [arAMC, PAV-L]
(2000) Beyond self-interest: A set of propositions relevant to interpersonal orientations. In: *European review of social psychology, vol. 11,* ed. M. Hewstone & W. Stroebe. Wiley. [PAV-L]
Van Lange, P. A. M. & De Dreu, C. K. W. (2001) Social interaction: Cooperation and competition. In: *Introduction to social psychology, 3rd edition,* ed. M. Hewstone & W. Stroebe. Blackwell. [aAMC]
Van Lange, P. A. M., Liebrand, W. B. G., Messick, D. M. & Wilke, H. A. M. (1992) Social dilemmas: The state of the art. In: *Social dilemmas: Theoretical issues and research findings,* ed. W. B. G. Liebrand, D. M. Messick & H. A. M. Wilke. Pergamon. [aAMC]
Van Lange, P. A. M., Otten, W., De Bruin, E. M. N. & Joireman, J. A. (1997) Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology* 73:733–46. [PAMVL]
Van Vugt, M. (1998) The conflicts in modern society. *The Psychologist* 11:289–92. [aAMC]
Vlaev, I. & Chater, N (submitted) Game relativity: How context influences decisions under uncertainty. [IV]
von Neumann, J. (1928) Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen* 100:295–320. [aAMC, AR]
von Neumann, J. & Morgenstern, O. (1944) *Theory of games and economic behavior,* first edition. Wiley. [aAMC, WDC, AR]
(1947) *Theory of games and economic behavior, second edition.* Princeton University Press (2nd edition, 1947; 3rd edition, 1953). [aAMC]
Wagstaff, G. (2001) Integrated psychological and philosophical approach to justice: Equity and desert. In: *Problems in contemporary philosophy, vol. 50.* Edwin Mellen. [JL]
Walley, P. (1991) *Statistical reasoning with imprecise probabilities.* Chapman & Hall. [rAMC]
Weber, M. (1922/1968) *Wirtschaft und Gesellschaft.* English edition, 1968: *Economy and society: An outline of interpretive sociology,* trans. G. Roth & G. Wittich. Bedminster Press. [aAMC]
Wegner, D. L. (2002) *The illusion of conscious will.* MIT Press. [JIK]
Weibull, J. (1995) *Evolutionary game theory.* MIT Press. [KS]
Weirich, P. (1988) Hierarchical maximization of two kinds of expected utility. *Philosophy of Science* 55:560–82. [PW]
(1994) The hypothesis of Nash equilibrium and its Bayesian justification. In: *Logic and philosophy of science in Uppsala,* ed. D. Prawitz & D. Westerståhl. Kluwer. [PW]
(1998) *Equilibrium and rationality: Game theory revised by decision rules.* Cambridge University Press. [aAMC, PW]
(2001) *Decision space: Multidimensional utility analysis.* Cambridge University Press. [PW]
Whiten, A. (1996) When does smart behaviour-reading become mindreading? In: *Theories of theories of mind,* ed. P. Carruthers & P. K. Smith. Cambridge University Press. [SH]
Wicksteed, D. R. (1933) *The common sense of political economy.* Routledge. [AR]
Wilson, E. O. (1998) *Consilience. The unity of knowledge.* Knopf. [GSB]
Yeager, L. B. (2001) *Ethics as a social science: The moral philosophy of social cooperation.* Edward Elgar. [JL]
Young, H. P. (1993) The evolution of conventions. *Econometrica* 61(1):57–58. [HG]
Zermelo, E. (1912) Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. *Proceedings of the Fifth International Congress of Mathematicians, Cambridge* 2:501–10. [aAMC]