Colman, A. M. (2003). Beyond rationality: Rigor without mortis in game theory. *Behavioral and Brain Sciences, 26,* 180-198.

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

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 domains. 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 section 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 certainties, 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 and 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 payoff 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. Empirical 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. 73-4); 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 paper 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 section 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 as regards 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 bring about 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. 358-359), 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 section 5 of the target article) is quite separate from the problem of cooperation in social dilemmas (discussed in section 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 section 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, in which 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 by 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 zeroth-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 zeroth-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 per cent had progressed beyond first-order reasoning.

Hedden and Zhang's (2002) experiments were shrewdly designed and well executed, although I have drawn attention 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 16th 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 by 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. Furthermore, 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 reply. 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 $1000 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 per cent 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 per cent 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

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

Camerer, C. F. & Lovallo, D. (1999) Overconfidence and excess entry: An experimental approach. *American Economic Review* 89: 306-318.

Cheng, K., Pena, J., Porter, M. A., & Irwin, J. D. (2002) Self-control in honeybees. *Psychonomic Bulletin Review* 9: 259-263.

Christiansen, M. H. & Chater, N. (1999) Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science* 23: 157-205.

Colman, A. M. (2003) Depth of strategic reasoning in games. *Trends in Cognitive Sciences* 7: 2-4.

Cooper, R. W., DeJong, D. V., Forsythe, R., & Ross, T. W. (1993) Forward induction in the Battle-of-the-Sexes game. *American Economic Review* 83: 1303-1313.

Edgeworth, F. Y. (1881/1967) *Mathematical psychics.* Augustus M. Kelley/Kegan Paul. (Original work published 1881)

Elster, J. (1989) *Solomonic judgements: Studies in the limitations of rationality.* Cambridge University Press.

Fehr, E. & Gächter, S. (2002) Altruistic punishment in humans. *Nature* 415: 137-140.

Føllesdal, D. (1982) The status of rationality assumptions in interpretation and in the explanation of action. *Dialectica* 36: 301-316.

Hedden, T. & Zhang, J. (2002) What do you think I think you think? Strategic reasoning in matrix games. *Cognition* 85: 1-36.

Heider, F. (1958) *The psychology of interpersonal relations.* Wiley.

Holland, J. H. (1996) The rationality of adaptive agents. In: *The rational foundations of economic behaviour: Proceedings of the IEA conference, Turin, Italy,* pp. 281-297, ed. K. J. Arrow, E. Colombatto, M. Perlman, & C. Schmidt. Macmillan.

Kahneman, D. & Tversky, A. (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47: 263-291.

Kahneman, D. & Tversky, A., ed. (2000) *Choices, values, and frames.* Cambridge University Press.

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).

Kreps, D. M. & Wilson, R. (1982b) Sequential equilibria. *Econometrica* 50: 863-894.

Kyburg, H. E., Jr. (1987) Bayesian and non-Bayesian evidential updating. *Artificial Intelligence* 31:271-293.

Levi, I. (1986) *Hard choices.* Cambridge University Press.

Lewis, D. K. (1979) Prisoner’s dilemma is a Newcomb problem. *Philosophy and Public Affairs* 8: 235-240.

McClennen, E. F. (1985) Prisoner's dilemma and resolute choice. In: *Paradoxes of rationality and cooperation: Prisoner's Dilemma and Newcomb's problem*, pp. 94-104, ed. R. Campbell & L. Sowden. University of British Columbia Press.

Miller, G. & Isard, S. (1964) Free recall of self-embedded English sentences. *Information and Control* 7: 293-303.

Nash, J. F. (2002) Non-cooperative games. In: *The essential John Nash* (pp. 51-84), ed. H. W. Kuhn & S. Nasar. Princeton University Press.

Quine, W. V. (1962) Paradox. *Scientific American* 206(4): 84-96.

Rabin, M. (1993) Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281-1302.

Rachlin, H. (2000) *The science of self-control*. Harvard University Press.

Rachlin, H. & Green, L. (1972) Commitment, choice and self-control. *Journal of the Experimental Analysis of Behavior* 17: 15-22.

Rescher, N. (1975) *Unselfishness: The role of the vicarious affects in moral philosophy and social theory*. University of Pittsburgh Press.

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. 531-537, ed. R. L. de Mantras & D. L. Poole. Morgan Kaufmann.

Sen, A. K. (1985) Rationality and uncertainty. *Theory and Decision* 18: 109-127.

Thomas, J. P. & McFadyen, R. G. (1995) The confidence heuristic: A game-theoretic analysis. *Journal of Economic Psychology* 16: 97-113.

Tversky, A. (1996) Rational theory and constructive choice. In: *The rational foundations of economic behaviour: Proceedings of the IEA conference, Turin, Italy*, pp. 185-197, ed. K. J. Arrow, E. Colombatto, M. Perlman, & C. Schmidt. Macmillan.

Tversky, A. & Kahneman, D. (1992) Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5: 297-323.

Walley, P. (1991) *Statistical reasoning with imprecise probabilities*. Chapman & Hall.