Problems and Pseudo-Problems in Understanding Cooperation in Social Dilemmas

Andrew M. Colman and Briony D. Pulford

School of Psychology, University of Leicester, Leicester, United Kingdom

Abstract

Social dilemmas do not need to be solved by social projection or any other special mechanism or theory, because they are fully and unambiguously solved by the standard game-theoretic concept of strategic dominance. The real problems lie in explaining why human decision makers do not invariably choose the dominant-strategy solutions that are mandated by game theory, but those problems require explanation of the decision making of human agents with bounded rationality rather than any new solution of social dilemmas. The most influential explanations for the empirical findings are reciprocity, inclusive fitness, other-regarding preferences, team reasoning (properly understood in a way that escapes the criticism of it made in the target article), and similarity, with or without the social projection hypothesis.

Keywords: other-regarding preferences, prisoner's dilemma, reciprocity, similarity discrimination, social dilemma, team reasoning

The principal aim of this article is to clarify what is and what is not problematic in understanding cooperation in the dyadic Prisoner's Dilemma game and in multiplayer social dilemmas. A secondary aim is to provide a brief overview of the most important proposals for solving the genuine problems. Our commentary on the very interesting target article (Krueger, DiDonato, & Freestone, this issue) is embedded in a more general critical review of explanations for cooperation in social dilemmas. We hope to clarify some of the issues that are commonly misunderstood in the wide-ranging literature on cooperation in social dilemmas.

Figure 1 shows the payoff matrix of the Prisoner's Dilemma game in one of its most familiar forms (e.g., Axelrod, 1984). In the generalized payoff matrix for symmetric 2×2 games on the right, any game is a Prisoner's Dilemma game if the rank order of the payoffs obeys the inequalities $T > R > P > S$.

		II				II	
		C	D			C	D
I	C	3, 3	0, 5	I	C	R, R	S, T
	D	5, 0	1, 1		D	T, S	P, P

Figure 1. A typical Prisoner's Dilemma game on the left. *Note.* In the generalized payoff matrix for symmetric 2×2 games on the right, the Prisoner's Dilemma game is defined by the inequalities $T > R > P > S$.

To play the game, Player I chooses a strategy represented by a row, either *C* (cooperate) or *D* (defect), and independently (without knowledge of Player I's choice) Player II chooses a strategy represented by a column, either *C* or *D*. The outcome of the game is one of the four cells where the strategy choices intersect, and the pair of numbers in that cell are the payoffs to the two players, Player I receiving the first payoff and Player II the second. For convenience, payoffs are frequently interpreted as small amounts of money, and for most purposes this is harmless

enough, but in formal game theory payoffs are actually *utilities* representing the players' preferences among the possible outcomes, taking all relevant factors into account, not only the monetary or other objective values of the payoffs, but also feelings of generosity or spite, considerations of fairness, and all other extraneous factors that might affect a player's preferences.

This apparently simple game has attracted more attention from psychologists, economists, political scientists, sociologists, and philosophers than any other, for two main reasons. First, the game provides a simple model of a form of cooperation that arises commonly in social, economic, and political life. To give just one example, consider two restaurants competing for business in a small village and currently achieving equal market shares. Each restaurant manager is deciding whether to introduce a buy-one-get-one-free special offer. If either manager introduces the special offer unilaterally, then it will attract most of the customers from the rival restaurant, and increased revenues will more than offset the costs of the special offer, hence profits will increase; but if both managers introduce the special offer, then each will merely bear the costs without any revenue gains, and hence each will suffer reduced profits. It should be obvious that this is a Prisoner's Dilemma game in which introducing the special offer corresponds to choosing the defecting *D* strategy. Any dyadic interaction is a Prisoner's Dilemma game provided that both agents are better off if they both act in one way (cooperate) rather than another (defect), but each is tempted to defect unilaterally, because that yields the best payoff of all (and the worst of all to the other agent). Situations with this strategic structure are ubiquitous in everyday life.

The second reason why the Prisoner's Dilemma game has attracted so much attention is that it embodies a genuine paradox. Two perfectly rational players, who according to the standard assumptions of game theory know everything that can be known about the game, including the fact that both players are perfectly rational in the sense of always striving to maximize their individual payoffs relative to their knowledge and beliefs at the time of acting, will defect, although they would have received higher individual payoffs if they had both cooperated, and they know that too. The reason why they will defect is that defection yields a better payoff than cooperation whatever the co-player does: in the example in Figure 1 (left), a player receives 5 rather than 3 from defecting rather than cooperating if the co-player cooperates, and 1 rather than 0 from defecting rather than cooperating if the co-player defects. Hence, defection pays better than cooperation irrespective of the co-player's choice. The technical term for this is *strategic dominance*: *D* is a dominant strategy, and there should really be no argument about whether or not it is rational to defect, because it must always be rational to choose a dominant strategy and irrational to choose a dominated strategy. The strategic dominance argument is a straightforward and uncontroversial application to game theory of the *sure-thing principle* that Savage (1954/1972) included as one of the seven postulates of his axiomatic theory of rational decision making. Strategic dominance solves the game, in a game-theoretic sense, but the solution is undeniably paradoxical, because instrumental rationality is defined as choosing the best means to achieve the ends that you desire, yet two irrational players who cooperate in the Prisoner's Dilemma game end up with better payoffs than two rational players who defect.

There is no debate among game theorists about whether or not it is rational to defect in the Prisoner's Dilemma game. Whenever you find yourself in a situation in which you have to choose from two or more actions, one of which is better for you than any other in all circumstances that might arise, it must be rational to choose that action even if you do not know the circumstances. That is why it is rational to defect in the Prisoner's Dilemma game even if

you do not know what the co-player will decide, and there is nothing of substance to debate on that score. Furthermore, the Prisoner's Dilemma game is merely a special (two-player) case of the more general multiplayer social dilemma, defined by the following three properties: (a) there are n players ($n \geq 2$), each with two strategies (C and D); (b) an individual player receives a better payoff by choosing D than C irrespective of the number of other players choosing C or D ; and (c) every player receives a better individual payoff if all choose C than if all choose D . Here again, D is a dominant (sure-thing) strategy by definition (b above), hence there is no question but that a rational player will choose it with certainty.

Given that $T > R > P > S$, the dyadic Prisoner's Dilemma game shown in Figure 1 meets this tripartite definition of a social dilemma, hence it is a special case. Because both $T > R$ and $P > S$, each player receives a better payoff by choosing D than C irrespective of the co-player's choice, and because $R > P$, each player receives a better payoff if all (i.e., both) choose C than if all choose D . A further requirement, $2R > S + T$, is useful for studying repeated Prisoner's Dilemma games, because it ensures that joint cooperation pays better than turn-taking defection (C, D), (D, C), (C, D), . . . , but it should be clear from this discussion that it is not a necessary part of the definition of the game, although some researchers appear to believe that it is.

In game theory, a solution is a specification of what strategies rational players will choose and what the outcome will therefore be, and this class of games is fully solved: in social dilemmas, rational players will defect, and the outcome will therefore be joint defection. This outcome is the unique *Nash equilibrium* of the game, because it is the only outcome in which every player has chosen a *best reply*, defined as a strategy that yields the best payoff to the player choosing it, given the strategy chosen by the co-player(s). In the two-player case, given that the co-player has chosen D , a player can do no better than choose D , and joint defection is therefore a Nash equilibrium, whereas in every other outcome, at least one player could have done better by choosing differently, given the co-player's choice. For the same reason, joint defection is also the unique Nash equilibrium of a multiplayer social dilemma. Krueger et al. (this issue) claim that social projection can solve social dilemmas, but there is nothing to solve, because strategic dominance, resulting in Nash equilibrium, provides a complete and irrefutable solution.

The solution to the Prisoner's Dilemma game, and to social dilemmas in general, is paradoxical, but the problem lies in our misguided intuition, not in any error in the solution. It is an instance of what Quine (1962) called a *veridical* paradox, in which a sound argument leads to a conclusion that is surprising but true. A standard example of a veridical paradox is the following paradox of the other child. Mrs Smith has two children, at least one of whom is a girl. What is the probability that her other child is a girl? Our intuition tells us that the probability must be $1/2$, but the correct answer is actually $1/3$. Using G for a girl and B for a boy, and writing the first-born child on the left and the second-born on the right, there are just three equally likely possibilities in which at least one child is a girl, namely GB, BG, and GG, and only in the third case is the other child also a girl, so the probability is $1/3$. To check this empirically, toss a pair of coins repeatedly. Count only the outcomes in which at least one coin falls heads, and you will find that the proportion in which the other coin also falls heads approaches $1/3$ as the number of tosses increases. The paradox is veridical and the problem purely psychological, arising as it does from our misguided intuition—we do not need to develop new theories of sex determination in children—and the solution of the Prisoner's Dilemma game is similarly a veridical paradox.

Although no new solutions are required for social dilemmas, we do need to explain why people cooperate in these games. Experimental research confirms high levels of cooperation in

both one-shot and repeated social dilemmas of all sizes (see reviews by Balliet, Mulder, & Van Lange, 2011; Sally, 1995). What accounts for the tendency for many decision makers to cooperate? A number of interesting proposals have been put forward, including the social projection theory of Krueger et al. (this issue). We review some of the theories discussed in the target article together with some others that are missing from the target article. We focus on the theories that seem to us most important and interesting.

Reciprocity

Direct Reciprocity

Properly understood, the Prisoner's Dilemma is a game that is played once only, and the reasoning outlined in the paragraphs above applies to this *one-shot* version of the game. If the game is repeated a fixed number of times, and the number of repetitions is known in advance by both players, then it is easy to prove by *backward induction* that rational players will defect on every round (Colman, 1998, 2003; Luce and Raiffa, 1957, pp. 97-102); but if it is repeated an indefinite number of times, then the optimal strategy for a rational player is far from clear, and there are valid reasons to cooperate, the most obvious of these being reciprocity. Suppose you are expecting to play the Prisoner's Dilemma game in Figure 1 (left) an unknown number of times with the same co-player. Suppose that you judge the probability to be zero that your co-player will cooperate on the first round, and if you also defect on the first round, then you judge the probability to be zero that your co-player will cooperate on the second round; but if you cooperate on the first round, then you judge the probability to be $2/3$ that your co-player will reciprocate by cooperating on the second round. Perhaps personal experience has led you to these beliefs: their source does not matter to the argument that follows. Taking into consideration only these two rounds, which should be sufficient to get the point across, if you cooperate on both rounds, then your expected payoff is 0 on the first round and $(2/3 \times 3) + (1/3 \times 0)$ on the second, making a grand total of 2. On the other hand, if you defect on both rounds, then your expected payoff is 1 on each round, making the same grand total of 2. Obviously, if you judge the probability to be greater than $2/3$ that your co-player will reciprocate your cooperation on the second round, then your expected payoff is better from cooperating on both rounds than from defecting on both rounds. The dominance argument has obviously collapsed in this indefinitely repeated version of the game, because you have a valid reason to cooperate if you believe that the probability of reciprocal cooperation is any greater than $2/3$.

Trivers (1971) introduced the theory of *reciprocal altruism*, as he originally called it, and his own example was of a swimmer with cramp who has $1/2$ chance of drowning unless a passer-by jumps in and attempts a rescue, in which case there is a $1/20$ chance that both the swimmer and the passer-by will drown. A passer-by who believes that the tables will be turned at some later date, and that the swimmer will return the favor only if the passer-by offers help on this occasion, trades a $1/2$ chance of drowning at the later date (if nobody helps) for a $1/10$ chance (a $1/20$ chance of drowning now, while trying to save the swimmer, plus a $1/20$ chance later, when the swimmer returns the favor), and it is therefore rational for the passer-by to help the swimmer, despite the associated danger.

Direct reciprocity of this type is acknowledged to be one of the most powerful explanations for cooperation, and for the evolution of cooperation by natural selection (Imhof & Nowak, 2010; Trivers, 2005). Natural selection always favors payoff-maximizing strategies, and in evolutionary game theory these payoffs are measured in units of *Darwinian fitness*—expected lifetime reproductive success. In this sense, natural selection mimics rational choice.

Indirect Reciprocity

Even more powerful, especially in human communities, is the theory of *indirect reciprocity* (Alexander, 1987), according to which an agent's cooperative act can pay off in the long run, even if there is no possibility of the recipient reciprocating directly, if the act is observed by others who may be encouraged to cooperate in future with the agent, or who may spread the agent's good reputation among friends and associates by useful gossip (Feinberg, Willer, Stellar, & Keltner, 2012). Cooperation may thus enable people to cultivate reputations for cooperativeness that elicit indirect reciprocity from many others in the future. This mechanism provides a persuasive explanation for cooperation, and there is evidence that it has played a major role in the evolution of cooperation (Nowak & Sigmund, 1998, 2005).

However, both direct and indirect reciprocity are powerless to explain cooperation in one-shot, unrepeatable strategic interactions that are unobserved or that offer no opportunities for reputation enhancement. Some researchers have argued heroically along the following lines (e.g., Binmore, 1998, p. 343). The human brain has evolved by natural selection to solve problems posed in the environmental conditions of the Stone Age, the period from about 2.5 million to 6,000 BC, covering well over 99% of human evolutionary history, during which unobserved, one-shot strategic interactions were comparatively uncommon. Perhaps we have evolved to behave in one-shot, unobserved strategic interactions *as though* they are likely to be repeated or observed. Evolutionary biologists tend to be unimpressed by this type of argument, because it implies that natural selection, which normally exploits the minutest of differences with exquisite ruthlessness and efficiency, has somehow been unable to distinguish between strategic interactions that are repeated or observed and those that are not. This cognitive shortcoming would have exerted constant selective pressure and should have led to the spread of genes enabling individuals to exploit these important differences.

Strong Reciprocity

The theory of *strong reciprocity* or *altruistic punishment* (Fehr & Gächter, 2000, 2002; Gintis, Bowles, Boyd, & Fehr, 2005) was explicitly designed to explain cooperation in unrepeatable encounters lacking any opportunities for reputation management. The suggestion, bolstered by impressive evidence from experimental social dilemmas (presented as public goods games), is that many people have a tendency to cooperate voluntarily, if treated fairly, and to punish noncooperators. The propensity to punish noncooperators, whether physically, verbally, or by social ostracism, is presumed to have evolved as a mechanism to control cooperation. The theory is persuasive, as far as it goes, especially in light of the experimental evidence, and some people certainly do punish noncooperators in everyday life, but it replaces the original problem of explaining cooperation with a secondary one of explaining punishment, because punishment is costly in time, effort, and in terms of the real possibility of retaliation, and it is not clear why anyone should do it. Gintis (2000) put forward a theory designed to explain how punishment could have evolved, but it relies on "multilevel selection theory," a form of group selection that is firmly rejected by most evolutionary biologists nowadays.

Inclusive Fitness

Arguably the most important contribution to the theory of evolution since Darwin was the proof by Hamilton (1964, 1970) that a gene is favored by natural selection not only if it increases the fitness of its carrier (direct fitness), but also if it increases the fitness of other carriers of the

same gene (indirect fitness). *Inclusive fitness*, comprising direct and indirect fitness, turns out to provide a convincing explanation why a propensity to cooperate may be hardwired into us by natural selection. Hamilton focused attention on *altruism* rather than cooperation—we explain below why the argument applies equally to cooperation. Consider an altruistic act that (by definition) provides a benefit b to a recipient at a cost c to the altruist, with $c < b$, both costs and benefits being measured in units of Darwinian fitness. According to *Hamilton's rule* of inclusive fitness, a gene that causes organisms to perform such altruistic acts will be favored by natural selection if $rb > c$, where r is the coefficient of genetic relatedness between the cooperator and the recipient—roughly speaking, $1/2$ for a sibling, parent, or child, $1/4$ for a grandparent or grandchild, $1/8$ for a first cousin, and so on. Hamilton's rule means simply that altruism is favored by natural selection whenever the benefit to the recipient, discounted by the coefficient of genetic relatedness, is greater than the cost to the altruist.

II

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

Figure 2. A Prisoner's Dilemma constructed from mutually altruistic actions (C) providing a benefit $b = 5$ to the recipient at cost $c = 2$ to the altruist.

To see the relevance of this to cooperation, it is sufficient to understand that mutual cooperation in social dilemmas can be interpreted as mutual altruism. If two people are in a reciprocal interaction in which each has the opportunity to perform an altruistic act toward the other, then both are better off if both act altruistically than if both decline to do so, because then each receives a payoff of $b - c$, which is better than nothing, because $b > c$. Furthermore, the worst payoff of all goes to a player who performs the altruistic act while the other declines to perform it, because the altruist then pays the cost c without receiving any benefit b , and a player who declines to perform the altruistic act while the other player does perform it receives the best payoff of all, receiving the benefit b without paying any cost c . This payoff structure is evidently a simple (decomposable) Prisoner's Dilemma game, in which mutual altruism corresponds precisely to mutual cooperation. For example, if we arbitrarily set $b = 5$ and $c = 2$ (always subtracted), then we get the payoff matrix shown in Figure 2, and it is clearly a Prisoner's Dilemma game, because $T > R > P > S$ (see Figure 1, right). This suggests that mutual cooperation between relatives, at least, may be an instinctive pattern of behavior in certain circumstances, and one that is likely to have evolved by natural selection.

Other-Regarding Preferences

A standard game-theoretic assumption is that players are motivated exclusively to maximize their own individual payoffs. Recall that, strictly speaking, the payoffs are utilities, taking into account any concerns about the co-player's payoffs, hence the assumption that players are motivated to maximize their own payoffs is really tautological. However, we can distinguish between utilities and *objective* payoffs (monetary value, for example), and we can assume that players are motivated partly by their own objective payoffs and partly by the

objective payoffs of their co-player(s). It is then easy to explain cooperation in the Prisoner's Dilemma game and multiplayer social dilemmas in terms of other-regarding preferences.

In *interdependence theory* (see reviews by Rusbult & Van Lange, 2003; Van Lange, 2000), players' motivations are described in terms of *social value orientations* (SVOs) defined by *payoff transformations*. Players' other-regarding utilities are defined as functions of their own and their co-players' objective payoffs. Formally, if t_i and t_j are the objective payoffs to Players i and j in a two-player game, and Player i 's utility U_i is a function of these objective payoffs, then Player i is motivated to maximize $U_i = t_i$ under the *individualistic* SVO, $U_i = t_j$ under the *altruistic* SVO, $U_i = t_i + t_j$ under the *cooperative* SVO, and $U_i = t_i - t_j$ under the *competitive* SVO. These ideas were introduced by Messick and McClintock (1968) and McClintock (1972), early research concentrating on individualistic, cooperative, and competitive social value orientations. The assumption is that players are invariably motivated to maximize their own utilities, represented here by U , but that these utilities are not necessarily individualistic—in terms of objective payoffs, they may be cooperative, or competitive, or even altruistic, depending on individual differences and the particular circumstances of the interaction.

		II				II				II	
		C		D		C		D		C	
I	C	3, 3	0, 5	I	C	6, 6	5, 5	I	C	3, 3	5, 0
	D	5, 0	1, 1		D	5, 5	2, 2		D	0, 5	1, 1
Given matrix				Cooperative SVO				Altruistic SVO			

Figure 3. Payoff transformation of the given matrix of the Prisoner's Dilemma, using the cooperative and altruistic social value orientations.

Figure 3 shows the basic Prisoner's Dilemma game on the left and payoff transformations of it using the cooperative and altruistic SVOs. The individualistic SVO is simply the standard game-theoretic motivation in which players are motivated to maximize their individual payoffs without regard to the payoffs of co-players, and in the terminology of interdependence theory, it is represented by the *given matrix* on the left in Figure 3. In the cooperative SVO, players are motivated by the sum of their own and their co-players' payoffs, hence each of Player I's payoffs in the given matrix is replaced by the sum of both players' payoffs in the corresponding cell of the cooperatively transformed matrix, and the same is done for Player II's payoffs. In the cooperatively transformed payoff matrix, it turns out that cooperation is a dominant strategy for both players, because $6 > 5$ and $5 > 2$, and this implies that players who are cooperatively motivated, in the sense of wishing to do what is best for the pair rather than what is best for themselves as individuals, will cooperate. On the right of Figure 3 is the payoff transformation according to the altruistic SVO, representing the utilities of a player who is motivated exclusively to maximize the co-player's payoffs (each payoff in the given matrix being replaced by the co-player's payoff), and this also leads to a game in which cooperation is a dominant strategy for both players, because $3 > 0$ and $5 > 1$. It is clear that the cooperative and the altruistic SVOs can both explain cooperation in the Prisoner's Dilemma game.

The altruistic SVO may seem implausible to individualistically minded readers, but experimental research shows that it can be elicited reliably in certain circumstances (e.g., Batson

& Ahmad, 2001; see Batson & Shaw, 1991, for a review). Colman, Körner, Musy, & Tazdaït (2011) suggested the following homely example of how it can arise without special circumstances of the type studied by Batson and his colleagues. A jazz-loving man is married to a classical music lover, and each wishes to choose a musical recording as a wedding anniversary present for the other, knowing that they will spend hours listening to the music together. Intuitively, we should expect each spouse to maximize the other's objective payoff by choosing the other's preferred type of music. Both players would thereby maximize their own utilities, but those utilities would be altruistic rather than selfish. This is a dyadic example, and interdependence theory has been confined to two-player games, but Colman et al. worked out some mathematical implications of the altruistic SVO in multiplayer games and showed how it can provide a compelling interpretation of cooperation in social dilemmas of all sizes.

Team Reasoning

Theories of team reasoning (Bacharach, 1999, 2006; Sugden, 1993, 2005) offer an intuitively persuasive explanation of cooperation in social dilemmas. The key assumptions are that players are sometimes motivated to maximize collective payoffs, as in the cooperative social value orientation, and (crucially) that they adopt a distinctive mode of reasoning from preferences to decisions. In orthodox game-theoretic reasoning, players ask, *What do I want, and what should I do to achieve it?* In team reasoning, they ask, *What do we want, and what should I do to play my part in achieving it?* In social dilemmas, the answer is obviously, *We want joint cooperation, and I should play my part in achieving it by choosing C.*

Team-reasoning players begin by searching for a profile of strategies that maximizes the collective payoff of the group of players. If the maximizing profile is unique, then team-reasoning players select and play their component strategies of it; otherwise, the theory is indeterminate in the sense that it offers no team-reasoning solution in such a case. The target article refers to “murky notions of ‘collective rationality,’ ” but in everyday life, people often claim to be acting in the best interests of their family, company, department, sports team, or some other group to which they belong, rather than in their individual self-interest, and the sneer quotes around “collective rationality” seem out of place. Furthermore, there is strong experimental evidence that human decision makers are indeed influenced by collective rationality and team reasoning, even when the team-reasoning solution is not a Nash equilibrium, as in social dilemmas and many other games (Colman, Pulford, & Rose, 2008).

Krueger et al. (this issue) summarize team reasoning as follows:

With team reasoning, players do not focus on their individual payoffs, but on the sum of the payoffs within each cell of the matrix. They then see that what “we” want is the mutual cooperation payoff $2R$ as long as $2R > T + S$ But is it enough for a team player to say that she had done what was necessary without wondering if the other player will complete the deal? This question cannot be answered without making additional assumptions about players' expectations regarding the probability of cooperation, p_c . Alas, the theory does not say what role these estimates might play in the decision process. This is just as well because if estimates of p_c were admitted as part of a player's rationale, the team-reasoning hypothesis would devolve into the expected-reciprocity hypothesis.

This is a travesty of the theory as it is presented in the technical literature. A player has no reason to adopt the team-reasoning mode of playing in the absence of an expectation that the other player(s) will do likewise, and the theory assumes that this is a prerequisite of team reasoning. In Bacharach's (1999) formal version, the probability that a player will become a team-reasoner is represented by a parameter symbolized by the Greek letter omega, and players choose team-

reasoning strategies only when this probability is sufficiently high. The probability that a player will adopt a team-reasoning approach is thus represented by ω ($0 \leq \omega \leq 1$), and the value of ω is assumed to be known to every team member. According to the theory, a player will adopt the team-reasoning approach if and only if the value of ω exceeds a threshold for which the collective expected payoff of the group of players is maximized by so doing; otherwise, a player will lapse into the individual payoff maximization characteristic of standard game-theoretic reasoning.

		II	
		C	D
I	C	2, 2	-3, 3
	D	3, -3	1, 1

Figure 4. Prisoner's Dilemma game with nonobvious team-reasoning strategy.

Consider the Prisoner's Dilemma game displayed in Figure 4, an example analyzed by Bacharach (2006, p. 132). The collective payoff is maximized in the (C, C) outcome, when both players choose team-reasoning strategies C, because in that outcome the collective payoff is 4, assuming a natural interpretation of the collective payoff as the sum of the individual payoffs (in this outcome, $2 + 2$), and in every other outcome it is less than 4. A team-reasoning player will obviously choose C if $\omega = 1$, in which case both players are certain to choose C. If $\omega < 1$, and a player chooses C while the co-player lapses from team reasoning and chooses D, then the collective payoff will be 0 (the sum of 3 and -3). With these two facts in mind, the team-reasoning solution can be calculated as follows. The probability of both players team reasoning and choosing (C, C), with collective payoff 4, is ω^2 ; the probability of (C, D) or (D, C), with collective payoff 0, is $2\omega(1 - \omega)$; and the probability of (D, D), with collective payoff 2, is $(1 - \omega)^2$. It follows that the expected collective payoff is $4\omega^2 + 2(1 - \omega)^2 = 6\omega^2 - 4\omega + 2$. If team reasoning does not occur, then the players will lapse into individual reasoning and choose (D, D), with collective payoff 2 ($1 + 1$). Players will therefore choose team-reasoning C strategies if $6\omega^2 - 4\omega + 2 > 2$, that is, if $\omega > 2/3$, and they will choose individual-reasoning D strategies otherwise. It is certainly not the case that a player will choose a team-reasoning strategy whenever $2R > T + S$ (see Figure 1, right) "without wondering if the other player will complete the deal," as Krueger et al. (this issue) claim in their target article. Taking into account what the co-player is likely to do is the cornerstone of game-theoretic reasoning.

Similarity

Following an original idea suggested by Gauthier (1986), J. V. Howard (1988) and Danielson (1992) independently suggested a persuasive explanation, based on similarity, for cooperation in social dilemmas. They each provided a rigorous proof that rational decision makers have a reason to cooperate if they recognize their co-players to be identical to themselves. Both researchers formalized the problem in terms of game-playing automata that can compare their programs and recognize identical co-players when they meet them. The gist of the argument is that each player, knowing that the co-player is identical, can reason validly that any strategy choice is bound to be mirrored by the co-player's choice, because an identical co-player

must make the same choice in the same situation; therefore, because mutual cooperation yields a better individual payoff than mutual defection, it is rational to cooperate. Furthermore, Howard proved that a population of such players must be evolutionarily stable against invasion by mutants or migrants who invariably defect.

Howard (1988) implemented this *mirror strategy* in Basic programming language, and Danielson (1992) in Prolog. The (far from straightforward) computational implementation of the argument amounts to a certificate of its soundness, but the relevance of the argument to cooperation in human and animal populations is limited by the requirement that the cooperating players have to be literally identical. However, if the cost of cooperation is c and the benefit to the other player(s) is b , with $c < b$ as before, it is easy to prove that the payoff from cooperation exceeds that from defection whenever $p > (b + c)/2b$, where p is the probability that both players will choose the same strategy, and there is strong evidence from computational simulations with social dilemmas and a variety of other games that *similarity discrimination* evolves spontaneously and powerfully in the process of natural selection, with players cooperating selectively with co-players who are similar but not necessarily identical to themselves (Colman, Browning, & Pulford, 2012). It seems reasonable to conclude that similarity discrimination helps to explain cooperation and its evolution. Furthermore, observation of everyday life confirms that people are generally more inclined to cooperate with others who are similar to themselves in nationality, race, age, religion, and other characteristics, although that is more likely a result of social evolution than gene selection, because gene selection would be expected to lead to evolution of cheats who accept cooperation from similar others without cooperating themselves.

The social projection theory promoted by Krueger et al. (this issue) makes a much stronger claim. They argue that, in the absence of any other evidence, it is rational for players to apply Bayes' theorem and to assume that their co-players are likely to make the same strategy choices as themselves. They claim that it is rational for players to assume that if they decide to cooperate and have no other basis for judging the probable strategy of their co-players, then the Bayesian probability is 2/3 that their co-players will cooperate, and if they decide to defect, then the Bayesian probability is 2/3 that their co-players will defect; furthermore, if joint cooperation pays better than joint defection, that is a sound basis for choosing to cooperate. This is *evidential* decision making, and it is not a valid application of Bayes' theorem. Although it may be reasonable to use one's own action as evidence of the likely action of a co-player, it is not rational to apply the same reasoning to all available strategies as a basis for choosing the one that yields the best payoff, assuming that *any* choice will be matched by the co-player. Consider the following analogous argument. I am starting up a new business, and I intend to delay paying an associate in order to ease my cash flow. I have no idea what the probability is that my associate will delay paying me for the same reason but, using Bayes' theorem, I judge the probability to be 2/3 on the basis of my intention to delay paying. So far so good. But then, in a "last-minute intrigue," I notice that I would be better off if we both settled our bills on time. Using Bayes' theorem again, I hastily calculate the probability that my associate would reciprocate my timely payments to be 2/3, and on that basis I change my mind and decide to settle my bill on time after all. If it is obvious that this argument is fallacious, as we believe it is, for reasons explained in the following paragraph, then it should be clear that the social projection theory of cooperation in social dilemmas is equally fallacious.¹

If two events are correlated, then the occurrence of one is evidence that the other is likely to occur, but correlation does not imply causation. In the Prisoner's Dilemma game, although a player's decision to cooperate may provide evidence that the co-player is likely to cooperate, it

cannot cause the co-player to cooperate, because the players choose their strategies independently. A player who cooperates normally hopes that the co-player will cooperate also. But rational players choose strategies that are likely to produce desirable outcomes, not merely evidence that their co-players will play desirable strategies. In the Prisoner's Dilemma game, a rational player will therefore defect, because defection yields a better payoff than cooperation whether or not the co-player cooperates. Cooperation would provide desirable evidence—that the co-player is likely to cooperate—but not a desirable outcome, because it would yield a worse payoff against either of the co-player's strategies, and rational players try to optimize outcomes, not evidence.

It is possible, of course, that human decision makers use the form of evidential decision making that Krueger et al. (this issue) attempt to defend, and that evidential decision making may help to explain cooperation in social dilemmas. There is experimental evidence that people are seduced by evidential decision making in certain circumstances (Quattrone & Tversky, 1984) including decisions closely related to Prisoner's Dilemma games (Anand, 1990). But the theory does not have the "normative appeal" that is claimed for it, because the reasoning is unsound. Weber (1904-1905) argued that, during the Reformation, both entrepreneurs and workers began working harder, and saving money rather than spending it. According to the Calvinist doctrine of predestination that gained ascendancy at that time, those who were to be saved had been chosen by God at the beginning of time, and nothing that they could do could increase their chances of salvation if they were not already among the chosen. What they attempted to do was to produce evidence of having been chosen by engaging in acts of piety, devotion to duty, hard work, and self-denial such as they would be expected to manifest if they had been chosen and, according to Weber, that that is why capitalism developed more quickly and more strongly in Protestant than in Catholic countries. If this interpretation is right, and there is empirical evidence to support it (McClelland, 1961), then this is further evidence that ordinary people can be seduced by evidential decision making, but it is even easier to see the fallacy in this case.

Concluding Comments

Although cooperation is irrational in unrepeated (one-shot) social dilemmas, and there is no possibility of a normative solution that mandates or permits cooperation, it is not difficult to understand why decision makers cooperate in many experimental and everyday social dilemmas. We have reviewed the most important and interesting explanations. Simpson's paradox (Chater, Vlaev, & Grinberg, 2008), which is discussed in the target article, and metagame theory (N. Howard, 1971, 1974, 1987), which is not, are also interesting, but they are formal explanations, and space constraints prevent us from covering them here. Theories of morality, discussed at length in the target article, are more relevant to discussions of how we *ought* to act than to explanations of how we *actually do* act. We pass over the theory of random error without comment.

Lastly, Krueger et al. (this issue) make much of Rapoport's (1967) cooperation index $K = (R - P)/(T - S)$ as a criterion for evaluating theories of cooperation. They claim that a good theory ought to be able to explain the higher levels of cooperation that are observed in Prisoner's Dilemma games that are "nice" (ones in which K is large) than in those that are "nasty" (in which K is small). In our opinion, it is obvious why K correlates strongly with cooperation. First, if we fix the denominator $T - S$, then a large numerator $R - P$ implies that the individual reward for joint cooperation is large relative to the individual punishment for joint defection, and this means that payoff-maximizing players will be relatively strongly attracted to cooperation, and that the

value of K will also be relatively large. Second, if we fix $R - P$, then a large $T - S$ denominator implies that the temptation to defect is great, relative to the risk of the sucker's payoff from cooperating, therefore payoff-maximizing players will be relatively less inclined to cooperate, and the value of K will be correspondingly small, because the denominator is large. The correlation of K with cooperation therefore seems to follow directly from common sense, without the need for any special theory.

Footnote

¹ The “last-minute intrigue” experimental data in the target article are interpreted in terms of theoretical predictions that begin with “a player who has chosen to cooperate, who believes that the other player will *probably* cooperate too” [emphasis added]. This player is given the opportunity to switch unilaterally to defection and “now faces a choice between the value of $p_c R$ and $p_c T$.” The player actually faces a choice between expected payoffs of $p_c R + (1 - p_c)S$ and $p_c T + (1 - p_c)P$, respectively, and correct predictions require expressions of this form. However, from that point on, the probabilities disappear entirely and the analysis proceeds, in effect, with the tacit assumption that the other player chooses the same strategy *with certainty*, so that, given a last-minute opportunity for a bilateral switch: “Now, the projection hypothesis advises against switching because it would suggest trading the anticipated R payoff for the P payoff,” and so on. A bilateral switch from joint cooperation to joint defection is thus a reduced game consisting of just the top-left and bottom-right cells of Figure 1, and a rational (individual payoff-maximizing) player will obviously decline to switch, because the choice is simply between R and P , and $R > P$. Similarly, for Player I, a unilateral switch from joint cooperation reduces the game to the left-hand column in Figure 1, and a rational Player I will switch, because $T > R$. For similar reasons, a rational player will accept a bilateral switch and decline a unilateral switch from joint defection. These are the choices that most of the experimental participants predicted, but it is not correct to say that “the results uniquely supported the projection hypothesis,” because they support game theory (and common sense).

Acknowledgments

We gratefully acknowledge support from the Leicester Judgment and Decision Making Endowment Fund and from the University of Leicester for granting study leave to the second author.

Note

Address correspondence to Andrew M. Colman, School of Psychology, University of Leicester, Leicester LE1 7RH, United Kingdom. E-mail: amc@le.ac.uk

References

- Alexander, R. D. (1987). *The biology of moral systems*. New York: Aldine de Gruyter.
- Anand, P. (1990). Two types of utility: An experimental investigation into the prevalence of causal and evidential utility maximisation. *Greek Economic Review*, 12, 58-74.
- Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.
- Bacharach, M. (1999). Interactive team reasoning: A contribution to the theory of co-operation. *Research in Economics*, 53, 117-147. doi: 10.1006/reec.1999.0188
- Bacharach, M. (2006). *Beyond individual choice: Teams and frames in game theory* (N. Gold, & R. Sugden, Eds.). Princeton, NJ: Princeton University Press.
- Balliet, D., Mulder, L., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, 137, 594-615. doi: 10.1037/a0023489
- Batson, C. D., & Ahmad, N. (2001). Empathy-induced altruism in a Prisoner's Dilemma II: What if the target of empathy has defected? *European Journal of Social Psychology*, 31, 25-36. doi: 10.1002/ejsp.26
- Batson, C. D., & Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry*, 2, 107-122. doi: 10.1207/s15327965pli0202_1
- Binmore, K. (1998). *Just playing: Game theory and the social contract II*. Cambridge, MA: MIT Press.
- Chater, N., Vlaev, I., & Grinberg, M. (2008). A new consequence of Simpson's paradox: Stable cooperation in one-shot Prisoner's Dilemma from populations of individualistic learners. *Journal of Experimental Psychology: General*, 137, 403-421. doi: 10.1037/0096-3445.137.3.403

- Colman, A. M. (1998). Rationality assumptions of game theory and the backward induction paradox. In M. Oaksford & N. Chater (Eds), *Rational models of cognition* (pp. 353-371). Oxford: Oxford University Press.
- Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, 26, 139-153. doi: 10.1017/S0140525X03000050
- Colman, A. M., Browning, L., & Pulford, B. D. (2012). Spontaneous similarity discrimination in the evolution of cooperation. *Journal of Theoretical Biology*. Advance online publication. doi: 10.1016/j.jtbi.2011.05.022
- Colman, A. M., Körner, T. W., Musy, O., & Tazdaït, T. (2011). Mutual support in games: Some properties of Berge equilibria. *Journal of Mathematical Psychology*, 55, 166-175. doi: 10.1016/j.jmp.2011.02.001
- Colman, A. M., Pulford, B. D., & Rose, J. (2008). Collective rationality in interactive decisions: Evidence for team reasoning. *Acta Psychologica*, 128, 387-397. doi: 10.1016/j.actpsy.2007.08.003
- Danielson, P. (1992). *Artificial morality: Virtuous robots for virtual games*. New York: Wiley.
- Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90, 980-994. doi: 10.1257/aer.90.4.980
- Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137-140. doi: 10.1038/415137a
- Feinberg, M., Willer, R., Stellar, J. & Keltner, D. (2012). The virtues of gossip: Reputational information sharing as prosocial behavior. *Journal of Personality and Social Psychology*. Advance online publication. doi: 10.1037/a0026650
- Gauthier, D. (1986). *Morals by agreement*. Oxford: Oxford University Press.
- Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, 206, 169-179. doi: 10.1006/jtbi.2000.2111
- Gintis, H., Bowles, S., Boyd, R. T., & Fehr, E. (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*. Cambridge, MA: MIT Press.
- Hamilton, W. D. (1964). The genetical evolution of social behaviour. (Parts I and II). *Journal of Theoretical Biology*, 7, 1-16, 17-52. doi: 10.1016/0022-5193(64)90038-4
- Hamilton, W. D. (1970). Selfish and spiteful behaviour in an evolutionary model. *Nature*, 228, 1218-1220. doi: 10.1038/2281218a0
- Howard, J. V. (1988). Cooperation in the Prisoner's Dilemma. *Theory and Decision*, 24, 203-213. doi: 10.1007/BF00148954
- Howard, N. (1971). *Paradoxes of rationality: Theory of metagames and political behavior*. Cambridge, MA: MIT Press.
- Howard, N. (1974). "General" metagames: An extension of the metagame concept. In A. Rapoport (Ed.), *Game theory as a theory of conflict resolution* (pp. 261-283). Dordrecht-Holland: D. Reidel.
- Howard, N. (1987). The present and future of metagame analysis. *European Journal of Operational Research*, 32, 1-25. doi: 10.1016/0377-2217(87)90267-0
- Imhof, L. A., & Nowak, M. A. (2010). Stochastic evolutionary dynamics of direct reciprocity. *Proceedings of the Royal Society B*, 277, 463-468. doi: 10.1098/rspb.2009.1171
- Imhof, L. A., & Nowak, M. A. (2010). Stochastic evolutionary dynamics of direct reciprocity. *Proceedings of the Royal Society B*, 277, 463-468. doi: 10.1098/rspb.2009.1171
- Luce, R. D., & Raiffa, H. (1957). *Games and decisions: Introduction and critical survey*. New York: Wiley.
- McClelland, D. C. (1961). *The achieving society*. New York: Van Nostrand.
- McClintock, C. G. (1972). Social motivation: A set of propositions. *Behavioral Science*, 17, 438-454. doi: 10.1002/bs.3830170505
- Messick, D. M., & McClintock, C. G. (1968). Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology*, 4, 1-25. doi: 10.1016/0022-1031(68)90046-2
- Nowak, M. A., & Sigmund, K. (1998). The dynamics of indirect reciprocity. *Journal of Theoretical Biology*, 194, 561-574. doi: 10.1038/nature04131
- Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature*, 437, 1291-1298. doi: 10.1038/nature04131
- 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. doi: 10.1037/0022-3514.46.2.237
- Quine, W. V. (1962). Paradox. *Scientific American*, 206(4), 4-96. doi: 10.1038/scientificamerican0462-84
- Rapoport, A. (1967). A note on the index of cooperation for Prisoner's Dilemma. *Journal of Conflict Resolution*, 11, 100-103. doi: 10.1177/002200276701100108
- Rusbult, C. E., & Van Lange, P. A. M. (2003). Interdependence, interaction, and relationships. *Annual Review of Psychology*, 54, 351-375. doi: 10.1146/annurev.psych.54.101601.145059

- Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7, 58-92. doi: 10.1177/1043463195007001004
- Savage, L. J. (1972). *The foundations of statistics* (2nd ed.). New York: Dover. (Original work published 1954)
- Sugden, R. (1993). Thinking as a team: Towards an explanation of nonselfish behaviour. *Social Philosophy and Policy*, 10, 69-89. doi: 10.1017/S0265052500004027
- Sugden, R. (2005). The logic of team reasoning. In N. Gold (Ed.), *Teamwork: Multi-disciplinary perspectives* (pp. 181-199). Basingstoke: Palgrave Macmillan.
- Trivers, R. L. (1971). The theory of reciprocal altruism. *Quarterly Review of Biology*, 46, 25-57. doi: 10.1086/406755
- Trivers, R. (2005). Reciprocal altruism: 30 years later. In C. P. van Schaik & P. M. Kappeler (Eds.), *Cooperation in primates and humans: Mechanisms and evolution* (pp. 67-83). Berlin: Springer-Verlag.
- Van Lange, P. A. M. (2000). Beyond self-interest: A set of propositions relevant to interpersonal orientations. *European Review of Social Psychology*, 11, 297-331. doi: 10.1080/14792772043000068
- Weber, M. (1904-1905). Die protestantische Ethik und der "Geist" des Kapitalismus [The protestant ethic and the spirit of capitalism]. *Archiv für Sozialwissenschaft und Sozialpolitik*, 20, 1-54, 21, 1-110.