Perceptions of self and other in the prisoner's dilemma: Outcome bias and evidential reasoning

JOACHIM I. KRUEGER
Brown University

MELISSA ACEVEDO
Valencia Community College

In the prisoner's dilemma, self-interest clashes with collective interest. The way players resolve this conflict affects how others view them. Cooperators are seen as more moral than defectors, and, when there is no information about the other player's choice, cooperators and defectors are seen as equally competent. However, players who are defected against are seen as less competent, especially if they themselves cooperated (Experiments 1 and 2). Similarly, cooperators see themselves as more moral, but not as less competent, than defectors do (Experiments 3). Independent of concerns about reputation and self-image maintenance, evidential reasoning contributes to cooperative behavior. Players who project their own attitudes onto others are more likely to cooperate (Experiments 3). Compared with classic game theory, a theory of reputational concerns and evidential reasoning is better equipped to explain empirical patterns of choice behavior in social dilemmas.

In social dilemmas, people must choose between their personal good and the good of the collective. Regardless of what others do, people are better off following their self-interest than acting in the collective interest. However, the collective would be better off if most individuals set aside their own self-interests (Dawes & Messick, 2000). Surprisingly many people cooperate, thus fostering the common good while making themselves vulnerable to exploitation (Komorita & Parks, 1995; Sally, 1995). This collectively efficient but individually irrational behavior creates a dilemma for scientists: How can they justify attempts to enhance cooperation when such enhancement requires the sacrifice of rationality?[1]

Several theories seek to explain why some people cooperate and how greater cooperation can be fostered. Many of these theories focus on repeated dilemmas, in which participants can build reciprocal cooperative exchanges by signaling their own trustworthiness and their willingness to punish defectors (Gintis, Bowles, Boyd, & Fehr, 2005). In one-shot dilem-

AMERICAN JOURNAL OF PSYCHOLOGY
Winter 2007, Vol. 120, No. 4, pp. 593–618

mas, where opportunities for learning and influence are blocked, fewer options are available. For these games, proposals amount to objective or subjective changes in the payoff structure or to changes in the perceived probability that others will cooperate.

Some theories assume that people are rigorously self-interested. This view implies that people avoid any behavior that is penalized. In his essay on the tragedy of the commons, Hardin (1968) championed the Hobbesian view that mutual cooperation can be attained only when mutual coercion is mutually agreed upon. Another view assumes that people are conditional reciprocators. They will cooperate if they become convinced that others will cooperate too. In support of this view, Caporael, Dawes, Orbell, and van de Kragt (1989) found that participants were more willing to cooperate after talking with one another and making promises of cooperation. Neither proposal is quite pure, however. When after changes in the objective payoffs defection no longer dominates, the dilemma is not solved but eliminated. Increases in the subjective probability that others will cooperate do not solve the dilemma either. Instead, they sharpen it. If players cooperate more, they do so in even greater defiance of the normative mandate to defect. The player who promises to cooperate but then reneges stands to gain the most.

Other theories reject the idea that all people are exclusively motivated by self-interest. Instead, many people are assumed to have significant other-regarding preferences. They care about the payoffs of others (i.e., they are benevolent), or they care about fairness (i.e., they prefer small over large differences between their own and others' payoffs). According to this view, people with significant social preferences do not interpret the dilemma in its canonical form. Instead, they transform the given payoffs into subjective utilities that take benevolence or fairness into account (Fehr & Schmidt, 1999; van Lange, 1999). Unless other-regarding preferences are implausibly strong, they do not change the prisoner's dilemma into a game in which cooperation dominates (Coleman, 2003; Krueger, 2007).

In the present article, we argue that cooperation does not necessarily arise from social preferences. Instead, we suggest that for the sake of parsimony, the possible role of a variety of self-interested motivations must be examined first. One route to self-interested cooperation recognizes people's sensitivity toward the social implications of their actions. Arguably, many people see their own cooperation as a moral choice and, likewise, consider cooperation to be moral when it is shown by others (Singer, Kiebel, Winston, Dolan, & Frith, 2004). If so, their own cooperation, even in a one-shot dilemma, may support their self-image as a moral person. Assuming that moral self-images are built from and maintained by how others evaluate one's behavior, a choice in a social dilemma has to pass muster with an internalized audience (Mead, 1934).

In this view, people choose to cooperate in part to affirm a moral self-image.

The simple route from having concerns about morality to acting morally is potentially obstructed by concerns about competence. In classic game theory, the rationality of defection is axiomatic. To understand that defection is the dominating choice in the prisoner's dilemma, one needs only to be able to subtract payoff values presented in the matrix. A cooperator must have failed at this simple task or did not even bother to look. Hence, defection can be equated with competence, and cooperation with the lack thereof. Players who understand these implications are faced**,** if not with a dilemma**,** with a conflict between which aspect of their self-image they should care more about, morality or competence. Inferring values from choices, it then seems that cooperators care more about the former and defectors about the latter.

The first goal of the present research was to examine how strongly people associate morality with cooperation and competence with defection. In Experiments 1 and 2, we examined this association among third-party observers; in Experiment 3, we focused on self-perception. The second goal (also addressed in Experiments 1 and 2) was to examine whether judgments of morality or competence are influenced by the other player's choice. In a one-shot prisoner's dilemma, players do not know each other, they do not communicate, and they do not anticipate future interactions. At the time of choice, they have no information as to what others might do. Because players lack foresight and influence over others, the choices of other players cannot reflect on the quality of their own. Nonetheless, people often fail to fully discount outcome information when that information is irrelevant as a criterion for decision quality (Allison, Mackie, & Messick, 1996; Mellers, Schwartz, & Cooke, 1998). We argue that in the prisoner's dilemma, the competence domain is particularly vulnerable to outcome biases. We hypothesize that players who meet with defection are seen as less competent, but no less moral, than players who meet with cooperation.

The third goal is to examine whether evidential reasoning (Nozick, 1969) contributes to cooperation. According to this view, people cooperate inasmuch as they assume that others are similar to them. In the prisoner's dilemma, the assumption of similarity implies that one's own choice will turn out to be matched (i.e., reciprocated) rather than mismatched by the other player. In Experiment 3, this hypothesis is tested while controlling for self-enhancement (i.e., the difference between the positivity of the self-image and the positivity of the other-image).

Morality, competence, and outcome bias

Morality and competence are the most critical domains in person perception and self-perception, respectively (Wojciszke, 2005). An action is

judged as moral inasmuch as it intentionally sets aside, or even contravenes, self-interest. In the prisoner's dilemma, cooperators enable the collective good, whereas defectors ensure that it cannot be attained. Cooperators accept the risk of being "suckered" (i.e., being exploited by a defector), whereas defectors play it safe. Therefore, cooperators should be seen as more moral than defectors.

To be sure, outcome biases can occur in the moral domain. According to Alicke's (2000) theory of culpable causation, people most strongly condemn actions resulting in tragic (i.e., unforeseen and undeserved) harm to others (see Mazzocco, Alicke, & Davis, 2004, for empirical evidence). In contrast, the harmed party in a social dilemma is the self. When cooperators meet with defection, it is only they who suffer. Judging them ex post to be less moral only adds insult to injury.

Outcome bias in the domain of competence can be expected, however, because outcomes are ecologically valid cues into a decision maker's competence (Baron & Hershey, 1992). When judging competence in terms of mastery (as, for example, when professors grade exams), people must depend on outcomes. When they do, their judgments can be understood in terms of rational belief updating (Hoffrage, Hertwig, & Gigerenzer, 2000). However, when outcomes have no normative bearing on the quality of a decision, they contaminate judgment. Baron and Hershey (1988) carefully described physicians' decisions about whether to operate and also informed participants of the outcome (i.e., the patient lived or died). Although participants were asked to evaluate the physicians' decisions based only on the information that was available to the physicians at the time of decision, they could not help being swayed by the outcomes.

In a one-shot prisoner's dilemma, outcome information is irrelevant with regard to a player's competence. A player makes a choice without knowing what the other player will do and without being able to influence the other player's choice. A defector who ends up in a situation of mutual defection should be seen as no less competent than a defector who manages to exploit a cooperator. Likewise, a cooperator who ends up realizing the common good with another should be seen as no more competent than a cooperator who ends up being exploited.

Research and theory provide some clues for why an outcome bias might affect judgments of players' competence. First, a social psychological perspective suggests that the belief in a just world implies that people get what they deserve and deserve what they get, especially when an unfortunate outcome cannot be rectified (Lerner, 2003). It can therefore be hypothesized that players who have been suckered are seen as incompetent. Second, a cognitive psychological perspective suggests that any available information will have an automatic effect on judgment. The recognition that some information is irrelevant and the removal of that information

from consideration require additional mental resources that people may be unable or unwilling to deploy (Wegner & Bargh, 1998). Third, a statistical methodological perspective suggests that a bias is established most clearly when the outcome is a chance event. In the studies conducted by Baron and Hershey (1988) and Krueger (2000), participants attributed greater skill to players who won in an uncontrollable game. In the prisoner's dilemma, being paired with a cooperator or a defector is a matter of chance.

Evidential reasoning

If outcome bias affects judgments of competence in the prisoner's dilemma, the question of how it is that many people still manage to cooperate becomes all the more poignant. Although they may feel the moral pull toward cooperation, they may be justifiably concerned about being seen as incompetent, should their opponents defect against them.

Independent of any tendency to honor a moral norm of cooperation, evidential reasoning is one way of overcoming the threat of being victimized by outcome bias. Evidential reasoning is familiar to students of Newcomb's problem (Bar-Hillel & Margalit, 1972; Nozick, 1969), but its implications for the prisoner's dilemma are still poorly understood (but see Brams, 1975, for an early effort in this direction).[2]

Evidential reasoning applies when it is assumed that people enter the prisoner's dilemma in a state of indecision. As they consider cooperation and defection in turn, they ask themselves what their own decisions imply about their opponent's choice. A Bayesian calculation suggests that whichever choice they ultimately make is more likely to be matched (i.e., reciprocated) than mismatched by the other player. By definition, most people end up in the statistical majority. Because the prisoner's dilemma is designed to maximize uncertainty about the choices of others, the player's own choice is the only available cue. It is diagnostic of the other player's choice because the other faces exactly the same decision problem. Maximum uncertainty means that all conceivable probabilities of cooperation are equally likely a priori, and therefore, the posterior probability that the player's own choice will be matched is two in three (Dawes, 1989). This probability can then be used to assess the expected value of cooperation and the expected value of defection. If the former is larger than the latter, cooperation is the desirable choice for the sake of self-interest alone.

The controversy over evidential reasoning centers on the question of whether it is rational for players to actively choose that option, which maximizes the prospective payoff. The argument against evidential reasoning is that players can certainly not *cause* others to choose as they themselves do. To think that they can would be tantamount to magical thinking (Quattrone & Tversky, 1984). The counterargument is that the diagnostic impli-

cations of one's own behavior remain valid even though they lack causal power (see Krueger & Acevedo, 2005, for a review of these arguments).[3]

It is difficult to observe in real time how individuals mentally test out the implications of their own cooperation and defection before settling on a decision. Nonetheless, there is evidence that they do perform such tests. When contemplating election outcomes, research participants who believed that their party was more likely to win if they themselves voted than if they abstained were also most likely to express the intention to vote (Acevedo & Krueger, 2004; Quattrone & Tversky, 1984). In another study, expectancies of reciprocity in a prisoner's dilemma were directly manipulated. Most players cooperated if told that their own choice would be matched with a probability of three in four (Acevedo & Krueger, 2005).

In Experiment 3 of the present research, we seek to extend this finding by testing the idea that people are likely to cooperate to the extent that they have generalized expectations that others are similar to them. Finding a positive association between individual differences in the strength of projection and willingness to cooperate would be particularly strong evidence for evidential reasoning. In contrast to our previous work, participants would not be given causal (though probabilistic) power over the other player's choice. Moreover, individual differences in the strength of projection would have to be stable enough to generalize from the judgments of general personality traits to predictions of other players' choices in the game.

EXPERIMENT 1

We assessed perceptions of morality, competence, and outcome bias from the perspective of uninvolved observers. Because observers themselves do not commit to cooperation or defection, their judgments are largely free from self-reflective considerations.

Our first hypothesis was that cooperators would be perceived as more moral and less competent than defectors. We also expected that differences in perceptions of morality would be larger than differences in perceptions of competence. This expectation was grounded in two considerations. One was that judgments of morality play a greater role in person perception than do judgments of competence (Wojciszke, 2005). The other was that the classic game theoretic equation of rationality with defection is not without its critics (Nozick, 1993; Rapoport, 2003). If some scholars refrain from equating rationality with defection, the same may be true for research participants.

Our second hypothesis was that another player's defection would make a target player seem less competent but no more or less moral. At the outset, we were agnostic about whether the magnitude of the outcome bias

would be moderated by the player's own decision. In other words, we had no reason to predict that a cooperator would be seen as less competent than a defector when being defected against by another player.

METHOD

Undergraduate students ($N = 169$), who participated in a classroom setting, read a description of a prisoner's dilemma involving two people, Joe and Jack. Participants were shown a payoff matrix with the four possible outcomes. The payoffs were \$12, \$8, \$4, and \$0, respectively, for unilateral defection, mutual cooperation, mutual defection, and unilateral cooperation. In line with the canonical description of the game, participants learned that the two players could not communicate and that they had no other prior information about what their opponent might do.

After reviewing this information, participants learned that the target player (Joe) had chosen either Option A or Option B. Although A was clearly the cooperative option and B the defecting one, these labels were withheld. Two thirds of the participants then learned the other player's (Jack's) choice, which was varied independently of the target player's choice. The remaining participants received no information about Jack's choice.

Participants were next presented with a list of 10 trait adjectives, which were chosen to reflect individual differences in morality (vs. egotism) and competence (vs. naiveté). In order of presentation, the adjectives were *intelligent, ethical, rational, egocentric, deceitful, generous, naive, trustworthy, selfish,* and *optimistic.* Participants rated the target player on each item, using a scale ranging from 1 (*total absence of the trait*) to 9 (*total presence of the trait*).

RESULTS AND DISCUSSION

To assess whether the presented adjectives tapped the constructs of morality and competence as intended, ratings were submitted to a principal component analysis with varimax rotation. The first factor included the adjectives *ethical, generous, trustworthy, egocentric, deceitful,* and *selfish.* The second factor included the adjectives *intelligent, rational,* and *naive.* The adjective *optimistic* loaded on a lone third factor and was thus dropped from analysis. Separate scale scores were computed by averaging of the ratings on the six morality items and the three competence items. The scoring of the negative items (i.e., *deceitful, naive, selfish, egocentric*) was reversed so that higher scores indicated more favorable perceptions. Both scales were reliable, with alpha coefficients of .91 for morality and .77 for competence. The low correlation between the scale scores, $r = -.15$, $p = .05$, suggested that perceptions of morality and competence were distinctive but not opposites of each other. All hypothesis tests were conducted on composite scale scores (i.e., unweighted mean ratings).

Morality

Scores of the morality scale were submitted to a 2 (target player's choice: cooperation vs. defection) by 3 (other's choice: no information vs. cooperation vs. defection) factorial analysis of variance (ANOVA). The findings, which are displayed in the top panel of Figure 1, show that the cooperating player was seen as more moral, $F(1, 163) = 153.69$, $d = 1.93$. The other player's choice had no effect, $F(2, 163) = 1.46$, and it did not qualify the effect of the player's choice, $F < 1$.

Competence

Ratings of the player's competence showed a different pattern. As shown in the bottom panel of Figure 1, participants did not perceive defection as a sign of rationality. There was only a trend for the cooperator to be seen as less competent than the defector, $F(1, 163) = 2.97$, $p = .09$, $d = .27$. However, the predicted outcome bias emerged as a significant effect of

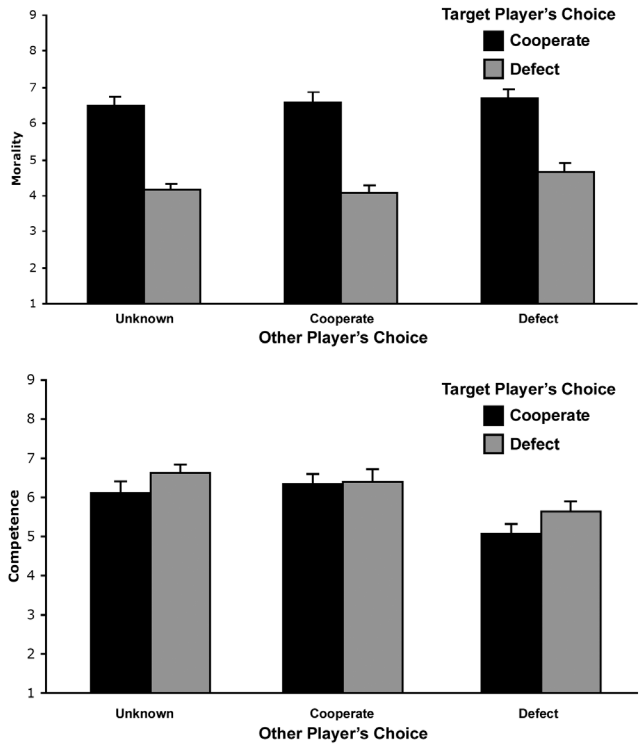

Figure 1. Cooperators' and defectors' judgments of morality (*top*) and competence (*bottom*) of self and other, Experiment 1

other's choice, $F(2, 163) = 8.99$, $p < .001$. Participants viewed the player as less competent if the other player had defected rather than cooperated or if the outcome was unknown. Post hoc comparisons performed with Tukey's honestly significant difference test showed that the other's defection reduced ratings of competence as compared with the other's cooperation ($d = .70$) and as compared with a situation in which the other's behavior was unknown ($d = .76$, both ps $< .001$). Thus, the outcome bias took only a negative form. The player's own behavior did not qualify the outcome bias, $F(2, 163) < 1$. The negative outcome bias is consistent with the pervasive tendency for negative information to exert more influence on impression formation (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001). The simplest version of this argument is that baseline impressions are near the positive ceiling so that only negative outcome information had room to modulate impressions.

EXPERIMENT 2

Experiment 2 was designed to replicate the association between a player's choice and attributions of morality. Likewise, we expected the negative outcome bias to reappear. To explore the robustness and consequences of this effect further, we introduced three additional judgment variables. First, we sought to confirm that the outcome bias can occur without attributions of foreknowledge. Consider a cooperator who meets with defection. If it is also acknowledged that one player could not anticipate the other's defection, judgments of incompetence are biased. In contrast, if it is believed that the player did anticipate defection, the perception of him being incompetent is not uniquely biased by the outcome. This player might be more appropriately characterized as a true altruist. Such a person would deliberately put another person's welfare ahead of his or her own. True altruists are rare, and we assumed that our research participants knew that they are rare.

Building on the literature on social projection (i.e., evidential reasoning), we predicted that participants would assume that players expect the other players to choose as they themselves do. Work in the prisoner's dilemma has shown that cooperators expect cooperation and that defectors expect defection (Dawes, McTavish, & Shaklee, 1977; Deutsch, 1960; Messé, & Sivacek, 1979). Of equal importance is the finding that people readily and accurately predict other people to project their responses to people in general (Krueger & Zeiger, 1993). We thus predicted that participants would expect cooperators to expect cooperation and defectors to expect defection from the other player.

Second, we agree with Baron and Hershey (1988) that evidence of out-

come bias is most persuasive when it can be shown that people understand and endorse the normative requirement that irrelevant outcomes be ignored. Assuming the outcome bias in the prisoner's dilemma to be robust and the normative rule to be evident, we predicted that participants would use negative outcome information (i.e., other player's defection) while denying that they did. Individual differences in the endorsement of the relevant norm should not be correlated with the strength of the bias.

With regard to the consequences of outcome bias, we followed a strategy used by Baron and Hershey (1988) and Krueger (2000). These investigators predicted and found that under the impression of a negative outcome, people avoid putting their own uncertain payoffs in the hands of those whose own efforts, through no fault of their own, had been thwarted by chance (or uncooperative others). We thus predicted that participants would prefer not to team up with a victim of defection in a new round of the game against a player.

METHOD

The study design paralleled that of Experiment 1 while omitting the condition in which the other player's choice remained unknown. After reviewing the instructions, the description of the game, and the information about the players' choices, 55 male and female undergraduate student participants rated the target player on nine trait adjectives. The trait of optimism was dropped because it was not clearly associated with perceived morality or competence. Participants then received three additional probes.

The first probe addressed the probability of cooperation on the part of the other player as it would have appeared to the target player. Participants were instructed,

> Turn back the clock and ask yourself what Joe might have expected Jack to do just before he, Joe, made his own choice. Enter what you think might have been Joe's estimates that Jack would choose Option A and that Jack would choose Option B.

Estimates were made as percentages under instructions that the two estimates would have to total 100%. The second probe addressed perceptions of decision quality. Participants were instructed,

> Suppose that both you and Joe join together as a team to play one round of this experimental game against a new player. As in the previous game, each player will receive a cash prize. Although both you and Joe will receive the amount indicated in the matrix below, only one of you can make the choice between Option A and B for your team. This choice will be paired with the other player's choice to determine everyone's payoff. How willing would you be to let Joe make the decision for your team? Overall, what do you think of Joe as a decision-maker (in relation to most others)?

These two ratings were made on scales ranging from 1 (*not willing at all* or *worse than most others*) to 9 (*extremely willing* or *better than most others*).

The third probe again addressed the robustness of the outcome bias. As noted

earlier, the bias would be especially troubling if participants recognized its nonnormative nature. Thus, they were asked,

> As best as you can recall, did you use information about Jack's decision when you rated the quality of Joe's decision? Do you think one should use information about Jack's decision when rating the quality of Joe's decision?

Ratings to these question could range from 1 (*certainly did not* or *absolutely not*) to 9 (*certainly did* or *most definitely*).

RESULTS

Morality and competence

The top panel of Figure 2 shows that participants rated a cooperating player as more moral than a defecting player, $F(1, 51) = 56.24$, $d = 1.91$. The other player's choice did not matter (both $F < 1$). The bottom panel of Figure 2 shows that the outcome bias was replicated in that a player meeting with defection was rated as less competent than a player meeting with cooperation, $F(1, 51) = 16.75$, $d = 1.09$. However, the bias was significant only when the player cooperated, $F(1, 24) = 13.10$, $p = .001$,

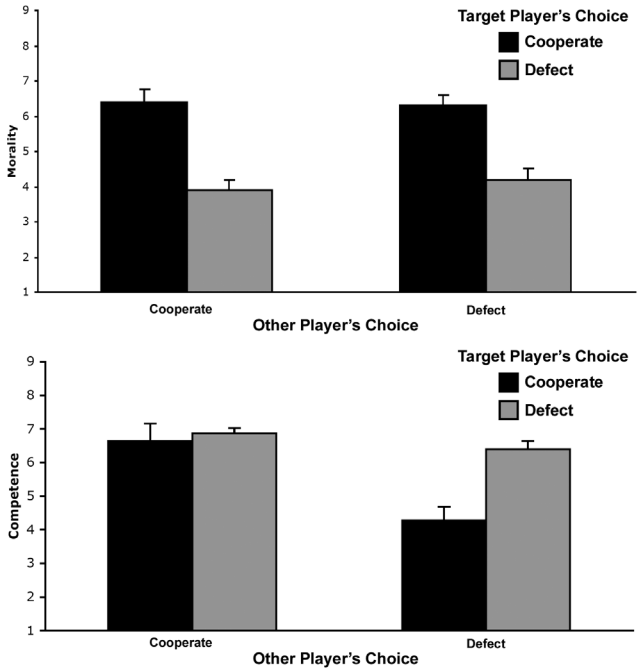

Figure 2. Cooperators' and defectors' judgments of morality (*top*) and competence (*bottom*) of self and other, Experiment 2

$d = 1.43$, but not when the player defected, $F(1, 27) = 2.48$, $p = .126$. The interaction between the player's and the other's choice indicated that these two simple effects were different in size, $F(1, 51) = 7.64$, $p = .01$.

Predicted probability of cooperation

As expected on the basis of the social projection literature, participants believed that a cooperator would provide a higher estimate of other's cooperation than a defector would, $F(1, 51) = 20.00$, $d = .49$ (Figure 3, top panel). Underscoring the robustness of the outcome bias, the other's choice did not moderate this effect, $F < 1$.

Judgments of information use

Participants' ratings of whether they used outcome information were correlated with their ratings of whether anyone should use such information, $r = .52$, and were thus averaged. These ratings hugged the middle of the scale, $M = 4.93$, $SD = 1.96$, and did not significantly vary across conditions, all Fs < 1.2. In other words, the outcome bias emerged, although participants did not ascribe any particular relevance to information indicating that the other player defected. The strongest support for the

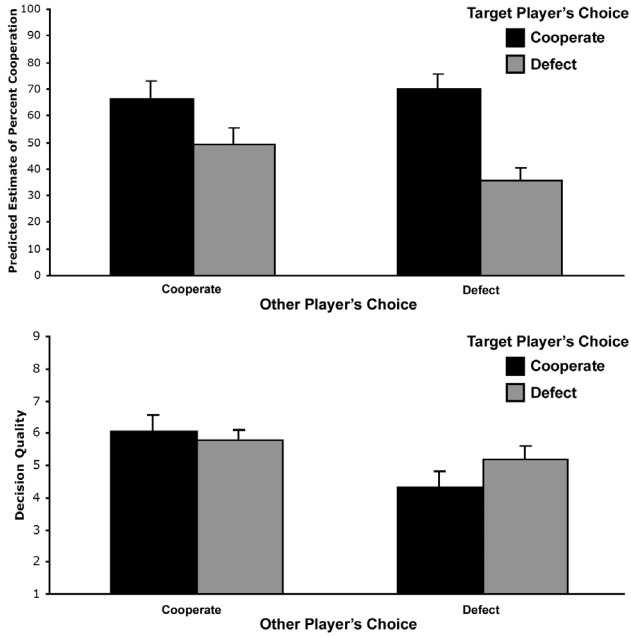

Figure 3. Predictions of cooperators' and defectors' estimates of opponent cooperation (*top*) and judgments of decision quality (*bottom*), Experiment 2

robustness of the bias would have required ratings near the bottom the scale. Although this did not happen, there was support for the prediction that ratings of information use would be independent of competence ratings, $r(26) = -.19$, $p = .35$ for other player's cooperation and $r(25) = .06$ for other's defection. Recall that the robustness of the bias would be in doubt if, given the other player's defection, low competence ratings were uniquely associated with high ratings of information use.

Decision quality

If the outcome bias is genuine, participants should take it seriously when considering their own involvement in the game. The perceived quality of the player's decision and ratings as to whether this player could be entrusted with a decision on behalf of a team including the participant were highly correlated, $r(53) = .74$, and thus averaged. As predicted, a player meeting with defection was considered a less capable decision maker than a player meeting with cooperation, $F(1, 51) = 6.88$, $p = .01$, $d = .67$ (Figure 3, bottom panel). The player's own choice did not matter, both Fs < 2.

The evidence suggests that people view victims of defection as less competent and that they would rather not have their own future payoffs depend on them. The coexistence of these two findings suggests a straightforward mediational model. Arguably, people mistrust victims of defection because they see them as incompetent. When perceptions of incompetence are statistically controlled, the association between opponent choice and judgments of decision quality (and trustworthiness) should disappear.

To test this mediational model, we first regressed the composite scores of perceived decision quality on the player's choice (cooperation = 1, defection = −1), the other player's choice (cooperation = 1, defection = −1), and the cross-product of the two. The other's choice was the only significant predictor, $\beta = .47$, $p = .016$, but this effect disappeared when competence ratings were added to the model, $\beta = .02$. In other words, players who had suffered defection were not entrusted with further choices in the prisoner's dilemma inasmuch as they had come to be seen as incompetent.

DISCUSSION

Experiment 2 replicated the morality effect observed in Experiment 1. Both effect sizes were large and of equivalent size. The outcome bias was also replicated, and the additional measures underscored its robustness. In contrast to Experiment 1, however, Experiment 2 revealed a larger outcome bias for cooperators. To address this inconsistency, we pooled the data while excluding the conditions in which the other player's choice remained unknown (Experiment 1). In this more powerful reanalysis, the interaction between the player's choice and the opponent's choice was

significant, $F(1, 163) = 5.14$, $p = .025$. Overall, then, we can conclude that the outcome bias was particularly damaging to cooperators. Cooperators who met with defection reaped a "sucker's penalty" that went beyond the downgraded competence ratings reaped by defectors who met with defection.

Three additional findings underscored the robustness of the outcome bias. Participants acknowledged that the player could not have foreseen the other's choice. Instead, they reasonably expected the player to project his own choice onto the other. In other words, participants did not believe that a player would have seen the opponent choice coming, but they believed that a player would use only his own choice as a projective cue. Participants' knowledge of the opponent choice did not affect what kind of knowledge they attributed to the player. Therefore, the outcome bias demonstrated here was not contaminated by a hindsight bias on the part of the participants. Finally, there was no evidence for the idea that the size of the outcome bias depended on participants' own appraisal of whether outcome information should be used or whether they themselves used it. Finally, the outcome bias reduced participants' confidence in victims of defection.

Before returning to the explanation of the outcome bias that we regard as being most plausible, we consider several alternatives. One possibility is that participants were sensitive to the rules of conversational logic, which demand that any available information is presented for a reason and should therefore be incorporated in judgment (Grice, 1975). This explanation is not fully convincing because it implies that participants would report that they used information about the other player, especially when defection occurred. This did not happen, however. A related possibility, alluded to earlier, is that *any* available information automatically affects judgment. The power of this explanation is reduced by the finding that participants managed to ignore outcome information when judging a player's morality. Ratings of morality thereby served as a control for the outcome bias on ratings of competence. Finally, the outcome bias might have been a mere "payoff bias," such that it was only the players' ultimate payoffs that controlled judgments of competence, not the other player's effect on payoffs. However, such a payoff heuristic would imply that a defector be seen as *more* competent than a cooperator when the other cooperated. The finding that perceptions of competence were pointedly sensitive to a player being "suckered" suggests that the outcome bias was truly social (as opposed to merely utilitarian).

This particular pattern returns us to the original hypothesis. Namely, people are used to seeing high and low levels of competence associated with positive and negative outcomes. When special circumstances arise in which outcomes are irrelevant for judgments of competence, they con-

tinue to rely on the simple and well-practiced rule of taking outcomes into account. Because negative outcomes generally loom larger than positive ones, some of the resulting judgments turn out to be irrational and unfair.

EXPERIMENT 3

The first goal of Experiment 3 was to ask whether people associate their own choices in the prisoner's dilemma with morality or competence. One possibility was that cooperators would see themselves as more moral, but not as less competent, than defectors. If so, it might be said that perceptions of others in the first two studies were partly accurate. Alternatively, own choices may not be associated with differential perceptions of morality. Participants have access to multiple sources of information that bear on their self-images. In contrast, observers in the first two studies could make personality judgments only on the basis of the players' choices (and, in the case of competence, on the basis of the other players' choices). To maintain methodological parallelism across studies, we elicited self-perceptions after participants had made their own choices.

The second goal was to test the hypothesis of evidential reasoning, namely the idea that people cooperate to the extent that they perceive others to be similar to them (Yamagishi & Kiyonari, 2000). Support for this hypothesis would suggest that players in the prisoner's dilemma arrive at a decision in part by estimating how likely it is that other players will reciprocate whatever choice they themselves make. To illustrate this idea, consider the limiting case of a player who cannot help but completely project his or her own behavioral inclinations on the other. For this player, the dilemma devolves into a choice between mutual cooperation and mutual defection. Because the former pays more than the latter, the decision does not require any regard for the opponent's payoff.

To measure perceptions of self–other similarity we used an idiographic measure of social projection, computed as the correlation between self-ratings and other ratings across the nine trait descriptors (Robbins & Krueger, 2005). Whereas idiographic projection correlations represent profile similarity, perceptions of self–other similarity can also arise as differences in elevation (Cronbach, 1955). Research on self-enhancement biases shows that most people view themselves more favorably than they view the average person (Alicke & Govorun, 2005). This bias is larger in the morality domain than in the competence domain, a finding that has been dubbed the Muhammad Ali effect (Allison, Messick, & Goethals, 1989; van Lange & Sedikides, 1998).

The inclusion of a self-enhancement measure permitted a purer test

of the evidential reasoning hypothesis. Larger projection correlations are associated with smaller correlations between self-enhancement and a third variable, such as cooperation (Krueger, in press). Assuming that the inverse is also true, it is necessary to assess the association between projection and cooperation while controlling for individual differences in self-enhancement. Put differently, to the extent that people think they are fairer (i.e., more cooperative) than others (Messick, Bloom, & Samuelson, 1985), they may hesitate to cooperate even though their attributions of fairness to others are correlated with their own.

METHOD

Female and male undergraduate students ($N = 181$) received a description of the prisoner's dilemma and a matrix illustrating the monetary payoffs for Options A (cooperation) and B (defection). After the payoff structure was explained, participants were asked to consider a single-round game against an opponent recruited from the same pool of participants. They then indicated their choice by checking Option A or B. Next, they were presented with the list of six trait adjectives from the morality domain and three trait adjectives from the competence domain and asked to rate themselves and the average student from 1 (*trait certainly absent*) to 9 (*trait certainly present*). The order of self-ratings and other ratings was counterbalanced.

RESULTS AND DISCUSSION

The proportion of cooperative choices (35%) lay within the range commonly observed in empirical work. Evidently, participants did not rush to cooperate merely because the game involved no payoffs in dollars.

Self and social perception

The composite ratings of morality and competence were obtained in a 2 (choice: cooperation vs. defection) by 2 (trait domain: morality vs. competence) by 2 (target: self vs. other) design in which the first variable was between participants. The means and standard deviations of the eight conditions are displayed in Figure 4. A mixed-model ANOVA revealed that self-ratings were more favorable than other ratings, $F(1, 176) = 80.70$, and that competence ratings were higher than morality ratings, $F(1, 176) = 56.31$. These effects were superseded by an interaction between target and domain, $F(1, 176) = 55.41$, whose shape revealed the anticipated Muhammad Ali effect. Self–other differences were larger in the morality domain, $F(1, 177) = 175.17$, $d = .99$, than in the competence domain, $F(1, 177) = 6.05$, $p = .015$, $d = .18$.

In Experiments 1 and 2, cooperation was associated with greater morality, but not greater competence, than defection. This effect was replicated

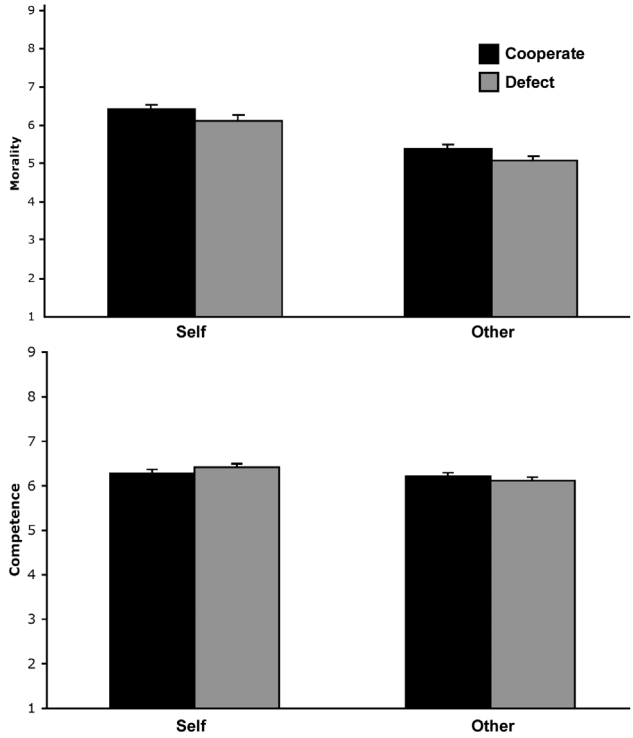

Figure 4. Cooperators' and defectors' judgments of morality (*top*) and competence (*bottom*) of self and other, Experiment 3

by the interaction between choice and domain, $F(1, 176) = 5.87$, $p = .018$. The absence of a triple interaction, $F = 1$, meant that cooperators rated both themselves and the average other higher on morality traits. Observers' perceptions in the first two studies were thereby socially validated. The finding that cooperators also viewed the average person as being more moral than defectors suggested the presence of social projection. We now turn to this phenomenon and its relevance for evidential reasoning.

Evidential reasoning

To index social projection, we computed for each participant the correlation between self-ratings and other ratings across traits. These correlation coefficients were transformed to Fisher's Z scores, averaged, and then transformed back to correlation coefficients. The average correlation ($M = .40$, $t(180) = 10.33$) was close to the meta-analytic benchmark obtained for projection to social ingroups (Robbins & Krueger, 2005). The test of the evidential reasoning hypothesis consisted of the correlation

between individual differences in projection and choice, $r(178) = .15$. The significance of this correlation, $p = .02$, one-tailed, suggested that stronger projection was associated with greater cooperation. Stated differently, cooperators projected more strongly ($M = .49$) than defectors did ($M = .35$).

Idiographic self-enhancement scores were computed from four averages, namely the average self-rating on morality traits, the average self-rating on competence traits, the average other rating on morality traits, and the average other rating on competence traits, where ratings for all negative traits were reverse scored. The self-enhancement scores, which were the differences between the summed self-related averages and the summed other-related averages, were negatively correlated with social projection, $r = -.37$. Participants who projected more self-enhanced less (Krueger, 2002). Individual differences in self-enhancement did not predict cooperation, $r = .05$, and therefore could not reduce the effect of evidential reasoning, partial $r = .14$.

The evidential reasoning effect was small to medium, corresponding to a difference of .3 in standard units. To better appreciate the regularity and significance of this finding, we ranked the participants according to their projection scores and examined the probability of cooperation for each quartile. As Figure 5 shows, the probability of cooperation rose linearly from .27 in the lowest quartile of projection to .45 in the highest quartile. As the probability of cooperation increases, so does the sum of all payoffs. If this probability were zero, the expected payoff would be that for mutual defection (here, \$4). If participants were grouped according to the strength of their projection, members of the lowest quartile could expect a payoff of \$5.08, whereas members of the highest quartile could expect \$5.80. The percentage of cooperators in a group corresponds exactly to the distance covered between the payoff for mutual defection and the payoff for mutual cooperation (here, \$5.80 is 45% of the distance between \$4 and \$8).

Payoff matrices differ in the readiness with which they enable cooperation. These differences are captured by the K statistic, which is the ratio of the difference between the mutual cooperation payoff and the mutual defection payoff over the difference between the unilateral defection payoff and the unilateral cooperation payoff (Rapoport, 1967). In this research, $K = (8 - 4)/(12 - 0) = 1/3$. A higher K value would have increased the expected value of the game, and mostly so for groups comprising high projectors. Moreover, higher K values make it more likely that individual players cooperate. Whenever the subjective probability that another player will reciprocate one's own choice is larger than $1/(1 + K)$, the expected value of cooperation is greater than the expected value of defection (Acevedo & Krueger, 2005). In the present research, $K = 1/3$ means that the

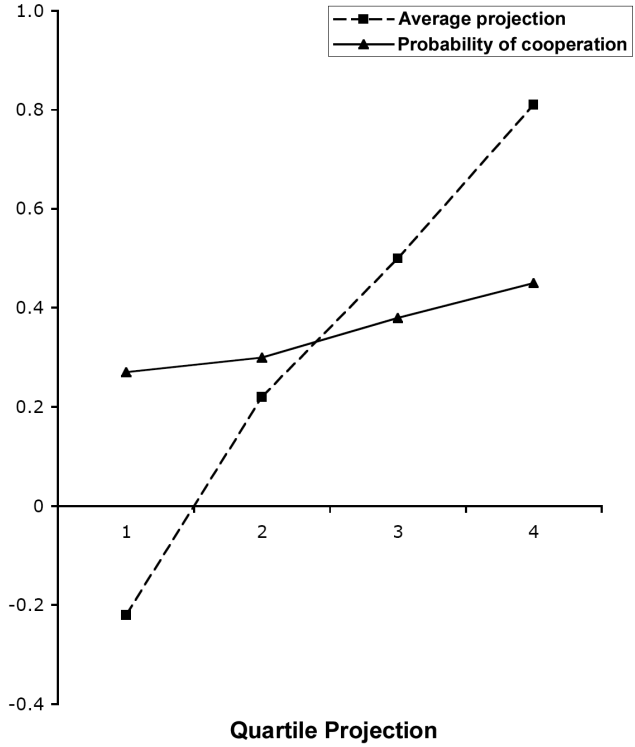

Figure 5. Relationship between the strength of projection and probability of cooperation, Experiment 3

probability of reciprocation had to be greater than .75, which corresponds to $r = .50$. Had a matrix with $K = .5$ been used, for example, a projection coefficient of $r = .33$ would have been enough to entice cooperation.

The correlational nature of the findings raises the possibility of reverse causation. Perhaps the decision to cooperate strengthened projection, or the decision to defect weakened it. We think that this possibility was unlikely. There is no theoretical model that would predict it. Likewise, the empirical data on postchoice projection contradicted it. As noted before, cooperators and defectors in most studies expect their opponents to choose as they themselves did. When there is a difference, it is the defectors who project more (Kelley & Stahelski, 1970). We therefore conclude that the individual differences in projection were in place *before* the game was presented and that they induced some participants to expect reciprocity and thus to cooperate.

GENERAL DISCUSSION

People associate morality more strongly with cooperation than with defection, they commit an outcome bias when judging victims of defection (and cooperators in particular) as being less competent, and to the extent that they project their own responses onto others they are more likely to cooperate. These findings shed new light on why people cooperate and how rates of cooperation may be further increased. Concerns about the integrity of one's self-image and the potential use of evidential reasoning do not play a role in classic game theory. The traditional assumption is that defection is motivated by the hope of exploiting other players or by the fear of being exploited. Cooperation is seen as an anomalous altruistic act that can be performed only when these powerful self-serving motivations are held at bay (Dawes & Thaler, 1988).

The value of explaining an altruistic act with reference to an altruistic motive is obviously limited, and the present analysis suggests that no such correspondent inference should be made without corroborating evidence. We suggest that such corroboration can be obtained when it is recognized that people care not only about monetary rewards but also about how their choices reflect on their personalities. Because cooperation satisfies a social norm of responsible behavior, people judge themselves and others as more moral when they cooperate rather than defect. Concerns about social prestige enable both cooperative behavior and the willingness to punish norm violators at a personal cost (Gintis et al., 2005).

Evidential reasoning means that people simulate other people's possible choices in their own minds before deciding what to do. Hence, those who project their own choices most strongly to others are also most likely to cooperate. Evidential reasoning can help explain how cooperation is possible in the absence of altruistic motives. In this sense, evidential reasoning is consistent with the classic game theoretic assumption that people are primarily self-regarding. This is not to deny that many people care about others and about fairness. Prosocial people prefer mutual cooperation payoffs over the unilateral defection payoff (van Lange, 1999). Still, the attribution of cooperation to a prosocial attitude is incomplete without an element of evidential reasoning. To coordinate their choices, they still need to project their own choices onto the other (Acevedo & Krueger, 2005).[4]

Social preference theories are incomplete in another sense. They assume that prosocials assign a certain weight to the other player's payoff and to fairness and that this weight is a stable expression of their attitude. In contrast, the theory of evidential reasoning recognizes that the degree to which people project from themselves to others is sensitive to context, the most robust finding being that people strongly project to members of

their ingroups and project only weakly to members of outgroups (Buchan, Croson, & Dawes, 2002; Robbins & Krueger, 2005).[5]

When strategic behavior is demanded, the contextualization of social projection helps explain why evidential reasoning in social dilemmas is consistent with the theory of inclusive fitness (Hamilton, 1964). According to that theory, the probability of helping increases with the degree of genetic relatedness. When people help their kin, they may not benefit directly as individuals, but they increase the probability that their genes will be propagated. Altruism and egotism thus merge. Because degrees of kinship sometimes are difficult to assess, humans have evolved to use social ecological cues as proxies. Phenotypic similarity, spatial proximity, and shared group membership are among the relevant variables. By stimulating projection to similar others, evidential reasoning enables cooperation where it matters the most, in small local groups.

Evidential reasoning is also compatible with reciprocal altruism (Trivers, 1971). Reciprocal altruism consists of ongoing, mutually beneficial exchange relationships. The question is how these relationships get started. Computer simulations show that a strategy of tit for tat optimizes both individual and collective welfare (Axelrod & Hamilton, 1981). The strategy can be summarized by two simple rules: Cooperate on the first move and then reciprocate on each subsequent move. Evidential reasoning provides a rationale for why a player might cooperate on the first move (unless, however, the payoff matrix is very difficult, i.e., K is very low).

The present research makes several novel recommendations for how cooperation might be increased in social dilemmas. These recommendations can be implemented without changing objective payoffs and thereby defining the dilemma away. Most importantly, the theory of evidential reasoning frees investigators from their own psychological conflict of having to induce individuals to act on behalf of the collective good in violation of their own self-interest (i.e., to act irrationally). People can truthfully be told that whatever they will choose to do, they will probably end up with the majority choice. This intervention was experimentally tested by Acevedo and Krueger (2005). Most participants in that study were willing to cooperate when told that there was a 75% chance that the other would reciprocate, and almost everyone cooperated when the probability of reciprocity was 1. This strategy does not amount to a self-fulfilling prophecy because participants are not told that they should cooperate or that others would cooperate. It is left to them to identify the statistical implications of their own choices.

Evidential reasoning brings the dangers of direct exhortations into focus. If people were admonished to cooperate, and if they knew that others were admonished in the same way, their temptation to defect might even

increase. Indeed, there is some evidence that rates of cooperation go down when people know the probability of cooperation or when they know that the other had cooperated (Acevedo & Krueger, 2005, Study 2; Shafir & Tversky, 1992). In other words, the effect of knowing the probability of cooperation can be self-eliminating, whereas the effect of knowing the probability of reciprocity is not.

Much as evidential reasoning frees investigators from the problem of finding ways to make participants act irrationally, it also solves the question of how to give coherent advice. An investigator relying on classic game theory cannot, in good conscience, advise anyone to cooperate. To be coherent, such an investigator would have to steer each player to defection as the "sure thing" and thus to the collectively aversive Nash equilibrium of mutual defection. In contrast, an investigator relying on evidential reasoning can remind players that in the end, they will probably be in the majority. To stimulate strong projection, the investigator can remind the players of social categories that include other players (Krueger & Clement, 1996). The investigator can then leave it to the players to draw their own conclusions.

In social cognitive research, errors and biases often are summarily assumed to be bad, although their effects on behavior are not considered (Krueger & Funder, 2004). Research on social projection grew out of studies on the false consensus effect. Because it has been shown that projection need not be false but a bias compatible with Bayesian induction, the time has come to consider its adaptive implications for interpersonal behavior. In contrast, the outcome bias amounts, under certain conditions, to a genuine cognitive error. Although outcomes tend to be correlated with decision quality in the social world, failure to neglect truly irrelevant outcomes creates social costs, such as the unfair denigration of unfortunate decision makers.[6] The present research suggests that, inasmuch as people can foresee the implications of the outcome bias for themselves, they may refrain from cooperative behavior. In this sense, the socially beneficial heuristic of projection is pitted against the detrimental error of outcome bias. In order to induce cooperation, evidential reasoning not only must be statistically strong, it must overcome the perceived risk of putting one's reputation in the competence domain on the line.[7]

Notes

Melissa Acevedo is now at Westchester Community College. We thank Robyn Dawes, Theresa DiDonato, Jordan Robbins, and Judith Schrier for their helpful comments on a draft version of this manuscript.

Correspondence about this article should be addressed to Joachim I. Krueger, Department of Psychology, Brown University, Box 1853, 89 Waterman Street,

Providence, RI 02912 (e-mail: Joachim_Krueger@Brown.edu). Received for publication January 12, 2006; revision received June 9, 2006.

1. It should be noted that cooperation is desirable for the collective only in the dilemmas typically presented to research participants. The valence of cooperation is reversed occasionally, as when individuals form coalitions to wage war on others or when local populations aim to outbreed competing populations (Krueger, 2007).

2. A Newcomb player faces two boxes. Box A contains $1 million if a nearly but not perfectly infallible demon predicted that the player would take only Box A. Box B contains $1,000 regardless of whether the player takes only Box A or both boxes. Taking both boxes is the dominating choice, whereas evidential reasoning suggests that taking only Box A postdicts—but does not cause—the demon's prediction and thus the presence of riches.

3. According to the view that defection is the only rational (i.e., dominating) choice, evidential reasoning is magical, and the belief that cooperation maximizes utility can only be illusory. In contrast, the view that evidential reasoning is reasonable suggests that choice-dependent utility estimates are akin to multistable figures known from visual perception (Attneave, 1968). Much as the Necker cube permits different and mutually exclusive interpretations of the same stimulus display, the prisoner's dilemma affords two different but equally reasonable predictions of what others do. Because neither prediction is objectively privileged, why should people choose the one that offers less money?

4. Taking the view that choices based on evidential reasoning are irrational, Robyn Dawes (in a signed review) raised the question of whether magically thinking players would also need to project their own magical thinking onto others. We submit that there is no need to do that. Evidential reasoning requires only that people recognize that they are more likely to end up in the majority than in the minority. They need not make any assumptions about *how* their fellow majority members got there. However, if they did project their own tendency to project (and their tendency to project their own projections ad infinitum), the probability of cooperation should ultimately approach 1.

5. The moderating effect of social categorization can also explain why groups are less cooperative with one another than individuals are (Wildschut, Pinter, Vevea, Insko, & Schopler, 2003). When individuals interact, a common group membership often can be found. When groups interact, however, social categorization is highly salient. Therefore, members of one group cannot project their own intended choices onto members of the other group.

6. Participants who decline to let a recent sucker play on their behalf betray an illusion of control because their choice cannot improve their chances to win.

7. To be fair, hopes of increasing one's prestige in the moral domain are aligned with evidential reasoning in their pull toward cooperation.

References

Acevedo, M., & Krueger, J. I. (2004). Two egocentric sources of the decision to vote: The voter's illusion and the belief in personal relevance. *Political Psychology, 25,* 115–134.

Acevedo, M., & Krueger, J. I. (2005). Evidential reasoning in the prisoner's dilemma. *American Journal of Psychology, 118,* 431–457.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126,* 556–574.

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. Dunning, & J. I. Krueger (Eds.), *The self in social perception* (pp. 85–106). New York: Psychology Press.

Allison, S. T., Mackie, D. M., & Messick, D. M. (1996). Outcome biases in social perception: Implications for dispositional inference, stereotyping, attitude change, and social behavior. *Advances in Experimental Social Psychology, 28,* 53–93.

Allison, S. T., Messick, D. M., & Goethals, G. R. (1989). On being better but not smarter than others: The Muhammad Ali effect. *Social Cognition, 7,* 275–295.

Attneave, F. (1968). Triangles as ambiguous figures. *American Journal of Psychology, 81,* 447–453.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science, 211,* 1390–1396.

Bar-Hillel, M., & Margalit, A. (1972). Newcomb's paradox revisited. *British Journal of the Philosophy of Science, 23,* 295–304.

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology, 54,* 569–579.

Baron, J., & Hershey, J. C. (1992). Judgment by outcomes: When is it justified? *Organizational Behavior and Human Decision Processes, 53,* 89–93.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323–370.

Brams, S. J. (1975). Newcomb's problem and prisoner's dilemma. *Journal of Conflict Resolution, 19,* 596–612.

Buchan, N. R., Croson, R. T. A., & Dawes, R. M. (2002). Swift neighbors and persistent strangers: A cross-cultural investigation of trust and reciprocity in social exchange. *American Journal of Sociology, 108,* 168–206.

Caporael, L. R., Dawes, R. M., Orbell, J. M., & van de Kragt, A. J. (1989). Selfishness examined: Cooperation in the absence of egoistic incentives. *Behavioral and Brain Sciences, 12,* 683–739.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences, 26,* 139–198.

Cronbach, L. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52,* 177–193.

Dawes, R. M. (1989). Statistical criterion for establishing a truly false consensus effect. *Journal of Experimental Social Psychology, 25,* 1–17.

Dawes, R. M., McTavish, J., & Shaklee, H. (1977). Behavior, communication, and assumptions about other people's behavior in a commons dilemma situation. *Journal of Personality and Social Psychology, 35,* 1–11.

Dawes, R. M., & Messick, D. M. (2000). Social dilemmas. *International Journal of Psychology, 35,* 111–116.

Dawes, R. M., & Thaler, R. T. (1988). Anomalies: Cooperation. *Journal of Economic Perspectives, 2,* 187–197.

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human Relations, 13,* 123–139.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114,* 159–181.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life.* Cambridge, MA: MIT Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics,* Volume 3: *Speech acts* (pp. 41–58). New York: Academic Press.

Hamilton, W. D. (1964). The genetical evolution of social behavior I and II. *Journal of Theoretical Biology, 7,* 1–32.

Hardin, G. (1968). The tragedy of the commons. *Science, 162,* 1243–1248.

Hoffrage, U., Hertwig, R., & Gigerenzer, G. (2000). Hindsight bias: A by-product of knowledge updating? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 566–581.

Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology, 16,* 66–91.

Komorita, S. S., & Parks, C. D. (1995). Interpersonal relations: Mixed motive interactions. *Annual Review of Psychology, 46,* 183–207.

Krueger, J. (2000). Distributive judgments under uncertainty: Paccioli's game revisited. *Journal of Experimental Psychology: General, 129,* 546–558.

Krueger, J. I. (2002). On the reduction of self-other asymmetries: Benefits, pitfalls, and other correlates of social projection. *Psychologica Belgica, 42,* 23–41.

Krueger, J. I. (2007). From social projection to social behavior. *European Review of Social Psychology, 18,* 1–35.

Krueger, J. I. (in press). The robust beauty of simple associations. In J. I. Krueger (Ed.), *Rationality and social responsibility: Essays in honor of Robyn M. Dawes.* Mahwah, NJ: Erlbaum.

Krueger, J. I., & Acevedo, M. (2005). Social projection and the psychology of choice. In M. D. Alicke, D. Dunning, & J. I. Krueger (Eds.), *The self in social perception* (pp. 17–41). New York: Psychology Press.

Krueger, J., & Clement, R. W. (1996). Inferring category characteristics from sample characteristics: Inductive reasoning and social projection. *Journal of Experimental Psychology: General, 125,* 52–68.

Krueger, J. I., & Funder, D. C. (2004). Towards a balanced social psychology: Causes, consequences and cures for the problem-seeking approach to social behavior and cognition. *Behavioral and Brain Sciences, 27,* 313–376.

Krueger, J., & Zeiger, J. S. (1993). Social categorization and the truly false consensus effect. *Journal of Personality and Social Psychology, 65,* 670–680.

Lerner, M. J. (2003). The justice motive: Where social psychologists found it, how they lost it, and why they may not find it again. *Personality and Social Psychology Review, 7,* 388–399.

Mazzocco, P. J., Alicke, M. D., & Davis, T. L. (2004). On the robustness of outcome bias: No constraint by prior culpability. *Basic and Applied Social Psychology, 26,* 131–146.

Mead, G. H. (1934). *Mind, self, and society.* Chicago: University of Chicago Press.

Mellers, B. A., Schwartz, A., & Cooke, A. D. J. (1998). Judgment and decision making. *Annual Review of Psychology, 49,* 447–477.

Messé, L. A., & Sivacek, J. M. (1979). Predictions of others' responses in a mixed-motive game: Self-justification or false consensus? *Journal of Personality and Social Psychology, 37,* 602–607.

Messick, D. M., Bloom, S. J. P., & Samuelson, C. D. (1985). Why we are fairer than others. *Journal of Experimental Social Psychology, 21,* 480–500.

Nozick, R. (1969). Newcomb's problem and two principles of choice. In. N. Rescher (Ed.), *Essays in honour of Carl G. Hempel* (pp. 114–146). Dordrecht, The Netherlands: Reidel.

Nozick, R. (1993). *The nature of rationality.* Princeton, NJ: Princeton University Press.

Quattrone, G. A., & Tversky, A. (1984). Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology, 46,* 237–248.

Rapoport, A. (1967). A note on the index of cooperation for prisoner's dilemma. *Journal of Conflict Resolution, 11,* 101–103.

Rapoport, A. (2003). Chance, utility, rationality, equilibrium. *Behavioral and Brain Sciences, 26,* 172–173.

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review, 9,* 32–47.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5,* 296–320.

Sally, D. (1995). Conversation and cooperation in social dilemmas. *Rationality and Society, 7,* 58–92.

Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology, 24,* 449–474.

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain responses to the acquired moral status of faces. *Neuron, 41,* 653–662.

Trivers, R. L. (1971). *Social evolution.* Menlo Park, CA: Benjamin/Cummings.

van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology, 77,* 337–349.

van Lange, P. A. M., & Sedikides, C. (1998). Being honest but not necessarily more intelligent than others: Generality and explanations for the Muhammad Ali effect. *European Journal of Social Psychology, 28,* 675–680.

Wegner, D. M., & Bargh, J. A. (1998). Control and automaticity in social life. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. I, pp. 446–496). New York: McGraw-Hill.

Wildschut, T., Pinter, B., Vevea, J. L., Insko, C. A., & Schopler, J. (2003). Beyond the group mind: A quantitative review of the interindividual–intergroup discontinuity effect. *Psychological Bulletin, 129,* 698–722.

Wojciszke, B. (2005). Morality and competence in person- and self-perception. In W. Stroebe & M. Hewstone (Eds.), *European review of social psychology* (Vol. 16, pp. 155–188). New York: Taylor & Francis.

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly, 63,* 116–132.