Psychological Inquiry, 23: 1–27, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1047-840X print / 1532-7965 online
DOI: 10.1080/1047840X.2012.641167

Psychology Press
Taylor & Francis Group

TARGET ARTICLE

Social Projection Can Solve Social Dilemmas

Joachim I. Krueger
Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, Rhode Island

Theresa E. DiDonato
Department of Psychology, Loyola University of Maryland, Baltimore, Maryland

David Freestone
Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Providence, Rhode Island

Evidence for cooperation in social dilemmas is empirically robust, socially desirable, and theoretically controversial. We review theoretical positions offering normative or descriptive accounts for cooperation and note the scarcity of critical tests among them. We then introduce a modified prisoner's dilemma to perform a critical test of the social projection hypothesis. According to this hypothesis, people cooperate inasmuch as they believe others respond to the situation as they themselves do. The data from three illustrative studies uniquely support the projection hypothesis. We make the analytical case for the social projection hypothesis in the context of the theory of evidential decision making. We review and rebut critical arguments that have been leveled against this theory. We note that a meta-theoretical benefit of evidential decision making is that the rationality of cooperators in social dilemmas is restored without appeals to murky notions of "collective rationality."

The immense majority of even the noblest persons' actions have self-regarding motives, nor is this to be regretted, since, if it were otherwise, the human race could not survive.—Bertrand Russell (1930)

The study of social dilemmas is an important enterprise within the social sciences. In the broadest sense, a social dilemma is a situation in which a person's self-interest is incompatible with the welfare of the group (Hardin, 1968). If all individuals put their self-interest first, the group does poorly, and so does each individual. From the individual's point of view, acting on the group's behalf while being the only one to do so, is even worse. To the individual, it would be best if all others acted selflessly. When all cooperate, the outcome is favorable for both the individual and the group. Of course, collective cooperation always raises the temptation to defect for any individual, and hence all. In short, the study of social dilemmas is concerned with the question of how coordinated collective action is possible in the face of stubborn selfishness.

There are many types of social dilemma. Some involve entire societies, with arms races, whaling, and climate control being familiar examples. Other dilemmas involve only two people, such as negotiators, parent and child, and the proverbial prisoner's themselves. Cooperative action may involve the contribution of resources or restraint in removing them (Dawes, 1980). Interactions may be repeated or not (Axelrod, 1980), and they may be symmetrical (as in the prisoner's dilemma; Flood & Drescher, 1952) or not (as in the trust game; Berg, Dickhaut, & McCabe, 1995). The choice of strategy may be performed in public or in private. The players may know one another or not. They may communicate or not, and so forth.

The fundamental question posed by social dilemmas is why anyone would want to abandon self-interest. Why do people sacrifice or forego resources by acting on behalf of others or the group? Let us begin with the starkest case of the prisoner's dilemma (PD), which involves two players who act anonymously and only once. The stubborn finding is that many of them cooperate. Cooperation is a welcome empirical fact, but it is also an inconvenient truth from a normative perspective (Luce & Raiffa, 1957; von Neumann & Morgenstern, 1947). Cooperation is welcome because it is perceived as the moral, other-regarding choice (Krueger & DiDonato, 2010). With the cost to the cooperator being

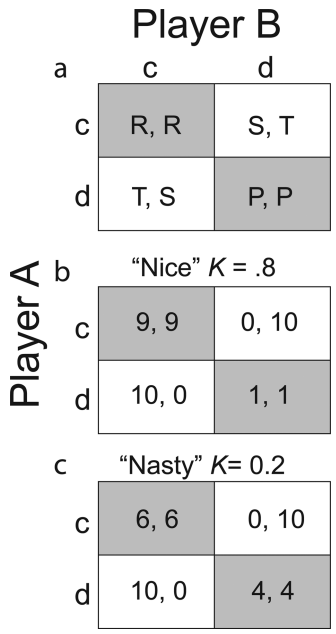

Figure 1. Payoff matrices for the prisoner's dilemma. Row (column) player's payoffs are shown to the left (right) of the comma.

smaller than the benefit to the other player, cooperation increases the overall efficiency of the game, that is, the sum of all individual earnings becomes larger, provided that the sum of the mutual cooperation payoffs is larger than the sum of the unilateral defection payoff and the unilateral cooperation payoff. However, cooperation is also inconvenient because it contradicts the theory that rational individuals should—and will—favor the dominating strategy of defection.[1] In the PD, defection dominates cooperation because it yields a higher payoff irrespective of the other player's behavior. Consider the payoff matrices displayed in Figure 1. The top panel shows the game in its symbolic form. If Player B cooperates by choosing strategy "c," Player A has a choice between the "Reward" payoff R, which is obtained if she too cooperates, and the "Temptation" payoff T, which is obtained if she defects by choosing strategy "d" (notation after Rapoport, 1967). As T > R, and if Player A would rather have more than less of a valued currency, she will defect. Conversely, if Player B defects, Player A has a choice between the "Sucker's" payoff S, obtained from cooperation, and the "Penalty" payoff P, obtained from defection. As P > S, defection is more rewarding. Put together, these two scenarios show that A does not even need to know B's choice. She is better off defecting, no matter what. This is why defection is the normative solution, although the Nash equilibrium it creates is deficient. Mutual defection is an equilibrium because neither player has an incentive to switch to cooperation (Howard, 1988); it is deficient because every individual, and thus the group as a whole, would be better off if they were to switch to cooperation en masse. The center panel and the bottom panel of Figure 1 show two numerical examples of payoffs in the PD. The relevance of the differences between the two matrices will become apparent soon.

From the normative point of view, the empirical evidence for cooperation is an anomaly (Dawes & Messick, 2000), and it is a theoretical paradox that a pair of rational players should be less prosperous than a pair of irrational players. Rapoport and Chamnah (1965) noted that, "confronted with this paradox, game theorists have no answer" (p. 29). To make matters more intriguing, the average probability of cooperation, as seen in meta-analysis, is quite large, approaching .5 (Sally, 1995). As anomalies accumulate, they reach a point at which they can no longer be ignored (Kuhn, 1962). A normative theory that cannot explain pertinent empirical data must be repaired or consigned to the sterility of an idealized abstraction. Several general proposals have been made to enrich classic game theory with assumptions, concepts, and findings from psychology and behavioral economics (Camerer, 2003; Colman, 2003; Pothos & Busemeyer, 2009). Among these efforts, there are several specific proposals to explain cooperation in social dilemmas and in the PD in particular. We review six such proposals, note some of their problems, and then present a seventh approach based on the theory of evidential decision-making (Jeffrey, 1983) and social-psychological research on projection (Dawes, 1989; Krueger, 1998). To evaluate the theoretical merit of this approach, we review common critiques put forth to question the normative claims of evidential decision making. To address the empirical merit of this approach, we introduce the "last-minute-intrigue paradigm" and present evidence that uniquely supports the evidential view. We then return to a discussion of normative issues and conclude with notes on the pragmatic value of evidential decision making.

Attempts to Account for Cooperation

The six conventional attempts to explain cooperation fall into two general classes: social normative and cognitive bias. Among the social normative approaches, we distinguish between morality-based theories, reciprocity, social value orientation, and team reasoning (see Kerr, 1995, for a similar classification). Among the cognitive-bias approaches, we distinguish between the error approach and the Simpson's paradox approach.

[1] In this article, we assume that game theory prescribes the use of the dominating strategy. Some game theorists, however, avoid prescriptive statements. Reinhard Selten, for example, asserted that "game theory is for proving theorems, not for playing games" (as cited in Goeree & Holt, 1999, p. 10564).

Morality

From the assumption that most people recognize cooperation as the just and compassionate strategy, it is a short step to attribute cooperative behavior to personal attitudes or dispositions reflecting such values. According to this account, some people cooperate because they want to, perhaps in an effort to placate an uncompromising conscience (Campbell, 1975). Notice that this conclusion is arrived at by way of exclusion, not with positive corroboration. Although this speculation may turn out to be valid, it runs the risk of being a correspondence error, which is an explanation that appeals to the disposition of which the behavior is a representative instance (Gilbert & Malone, 1995).

There are two types of morality-based explanation. These explanations differ in how purely they suggest people are motivated by concerns for morality. The purest conception of morality is deontological, which demands the desired behavior categorically, without regard to costs or benefits (Kant, 1785/1998; White, 2006). To explain the empirical data of about 50% cooperation, this account needs to make the implausible assumption that roughly every other person has internalized a categorical norm of cooperation. Further, the categorical morality hypothesis cannot explain why the probability of cooperation varies as a function of the "difficulty" of the game. The term *difficulty* does not refer to a subjective experience or the objective hardness of the task. Instead, it refers to the difference between the two payoffs for matching choices (R-P) relative to the difference between the payoffs for mismatching choices (T-S). Rapoport (1967) recommended the ratio $K = \frac{R-P}{T-S}$ as an index of difficulty and early research showed that it predicts cooperation in the PD (Jones, Steele, Gahan, & Tedeschi, 1968; Murnighan & Roth, 1983; Steele & Tedeschi, 1967).

The center panel of Figure 1 shows a game that is easy or "nice" (Chater, Vlaev, & Grinburg, 2008) because it has a high *K* ratio. Conversely, the game displayed in the bottom panel of Figure 1 is difficult or "nasty" because it has a low *K* ratio. A successful theory must explain why more people cooperate in a "nice" game than in a "nasty" game without begging the question. The categorical morality hypothesis would need additional assumptions to explain why morality should depend on game's *K* index. Contrary to the empirical findings one could argue that the moral high ground would be reached if people cooperated most robustly in nasty games.

When social norms demanding cooperation are merely known but not internalized, they can be followed strategically to bolster a person's reputation as someone who is moral and trustworthy (Krebs, 2008). This version of morality is utilitarian, if only in a Machiavellian sense (cf. Mill & Bentham, 1987). The question is whether this version of the morality hypothesis has more success in explaining cooperation. Cooperative behavior can serve as an instrument for building a reputation as a person with whom others seek exchanges, presumably to mutual benefit (Levitt & List, 2007) and perhaps also with a view toward future exploitation. The Machiavellian eye toward the future is crucial here. A moral reputation may be pleasant in the moment, but its value for material interests lies in exchanges still to be had. It is therefore difficult to reconcile the reputation-utilitarian approach with cooperation in one-shot games (Cooper, DeJong, Forsythe, & Ross, 1996). To overcome this difficulty, it is necessary to return to the idea that the social norm is internalized. If so, cooperators can derive moral satisfaction from their choice by regarding themselves as individuals who are doing the right thing. Sustaining a moral self-image is tantamount to maintaining a moral reputation in one's own eyes. With this assumption, the morality hypothesis returns to the tautology of the correspondence error. People cooperate because they want to.

The reputational-gain hypothesis is also unable to explain the nice-versus-nasty effect. Although the reputation gains would be greatest in nasty games, it is here that people cooperate the least.

Reciprocity

According to classic theory, players do not need to calculate expected values to realize that defection dominates cooperation. When it is enough to know that $T > R$ and that $P > S$, the probability that the other person will cooperate, p_c, is irrelevant. Indeed, there is no value of p_c at which the expected value of cooperation is equal to or greater than the expected value of defection. The expected value of cooperation, EVC, $= p_cR + (1 - p_c)S$ and the expected value of defection, EVD, $= p_cT + (1 - p_c)P$. It follows that EVC = EVD only if $P_c = \frac{P-S}{R-S-T+P}$, which cannot yield a value between 0 and 1.

According to the reciprocity hypothesis, some individuals care about p_c; they cooperate if they believe its value is high. The notion of *expected* reciprocity is an extension of a more general normative principle. A person who has received a favor is under pressure to reciprocate (Bicchieri, Duffy, & Tolle, 2004; Cialdini & Goldstein, 2004; Gouldner, 1960; Trivers, 1971). In sequential games, some players resist the temptation to defect after learning of the other player's cooperation (Berg et al., 1995; Gneezy & List, 2006; Krueger, Massey, & DiDonato, 2008; Pillutla, Malhotra, & Murnighan, 2003). The critical difference between the expected-reciprocity account and the canonical reciprocity account is, well, expectation. Players who cooperate because they think others cooperate reciprocate ex ante, before they have evidence that others cooperate.

At first glance, it may seem that the expected-reciprocity hypothesis can account for the nice-versus-nasty effect. People cooperate in nice games, where the *K* ratio is high, because they expect that others will cooperate, and they defect in nasty games, where the *K* ratio is low, because they expect that others will defect. Unfortunately, this view begs the question of why people expect the *K* value to affect others' decisions before it affects their own.

Social Values

Bertrand Russell (see epigraph) warned that we ignore human selfishness at our own peril. Hardin (1977) urged even more pointedly that we "never ask a person to act against his own self-interest." To many, this position seems cynical and unrealistic. Since the days of Adam Smith and David Hume, moral philosophers and psychologists (e.g., Miller, 1999) have doubted that people are entirely selfish (or "self-regarding"). These scholars argue that many people care (somewhat) about the welfare of others and that they are upset by large differences between their own welfare and the welfare of others. According to social preference models, other-regarding preferences, such as benevolence (caring about others) and inequality-aversion (fairness), are important elements of people's utility functions (Bolton & Ockenfels, 2000; Fehr & Schmitt, 1999). Following ideas proposed by Kelley and Thibaut (1978), van Lange (1999) suggested that the formula for the expected value of a strategy can be rewritten by taking other-regarding preferences into account. A benevolent player, for example, will multiply the other player's payoff with a weight *w* and add the product to his or her own payoff (see also Kollock, 1998).

To a benevolent player, who values both her own and the other player's payoff, the effective (i.e., transformed) value of cooperation is $R + S + w\,(T + R)$ and the effective value of defection is $T + P + w\,(P + S)$, and $0 \leq w \leq 1$. Cooperation is the dominating strategy if $w > \frac{T+P-R-S}{T+R-P-S}$. The benevolence hypothesis can explain why people are more willing to cooperate in nice than in nasty games. The explanation lies in the fact that over games, the critical threshold value of *w* is strongly and inversely correlated with the *K* index. In other words, when *K* is high (i.e., the game is nice), a small benevolence weight is sufficient to tip the scale.

This type of analysis can also be performed for the social value of fairness. Kerr (1995) wrote that "it seems very plausible that one of the reasons a sizable fraction of subjects in nearly every social dilemma study choose not to take advantage of clear free-riding opportunities is that such free-riding violates the equity norm, although I know of no research which confirms this directly" (p. 39). In van Lange's (1999) notation, the fairness weight *w* refers to the negative utility of self-other differences in payoff. However, no *w* is large enough to make cooperation the dominating strategy. At best, a concern with fairness can transform the PD into an assurance game or trust dilemma (also known as "stag hunt" after Rousseau, 1755/1992) with a payoff ranking of $R > T > P > S$. In this game, both players desire the efficient outcome of mutual cooperation but worry that the other player may want to settle for unilateral defection, in which case they would be suckered. When *w* is positive for both benevolence and fairness, the main effect of benevolence is diluted (Krueger, 2007).

At this point, the benevolence hypothesis (particularly when undiluted by concerns with fairness) has emerged as the only credible repair of the classic view of the exclusively self-interested person. By being sensitive to the relative weight of other-regarding preferences, this account differs sharply from categorical demands for moral behavior; by excluding expectations with regard to others, it distinguishes itself from the reciprocity hypothesis (Smith, 2003). Yet, some problems remain, which we will address in the discussion section.

Team Reasoning

A recent addition to the suite of repair models is the idea of team reasoning (Bacharach, 1999; Colman, Pulford, & Rose, 2008a; Sugden, 2000; see Krueger, 2008, or Schmid, 2003, for critical discussions). Team reasoning is presented as a radical departure from both the classic approach, which only considers self-regard as a source of utility, and social preference models, which introduce other-regarding preferences. The theory of team reasoning is an attempt to break with the tradition of methodological individualism. According to the hypothesis of team reasoning, people use the collective's perspective instead of their own individual perspective to assess utilities. Instead of asking, What do I want and what do I have to do to get it? they ask, What do we want and what can I do to help get it?

Taking a social-psychological view, Dawes, van de Kragt, and Orbell (1988) suggested that people may cooperate out of group solidarity and in the absence of any overt or hidden advantages to the self. Taking an evolutionary view, Roughgarden, Oishi, and Akçay (2006) suggested that at least for some species of animal, parental teams have greater success in passing on their genes to the next generation than selfish procreators do. "With courtship and intimate physical contact, they can synchronize activities and play as a team instead of as individuals" (p. 967).

Applied to the anonymous one-shot PD, team reasoning requires that the suspension of self-interest is generalized to a highly ambiguous situation. With team reasoning, players focus not 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. In this case, the player knows that without her cooperation 2R cannot happen. 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.

Likewise, the uniqueness of team reasoning is threatened by the possibility that its prediction can be reconstructed as a special case of the benevolence hypothesis. Individuals reasoning for the team do not discriminate between their own payoffs and the payoffs of others. They simply sum them up, which means that they act as if their benevolence is as strong as their self-regard.[2] With cooperation being the dominating strategy for *any* PD, the theory cannot explain the nice-versus-nasty effect.

The PD, as we consider it here, is symmetrical, that is T–R = P–S, which means that 2R > T+S. If an asymmetry is introduced such that 2R < T+S, fairness at the individual level is pitted against the collective good. A team reasoner would be troubled because there are now two competing interpretations of what is good for the team. If she wishes to attain the maximum sum of payoffs, she would want to cooperate if the other player defects and defect if the other player cooperates. In other words, an asymmetrical nasty PD would turn into a volunteer's dilemma (Diekmann, 1985).

Error

The first of the two cognitively oriented repair theories focuses on the role of error. When behavior is observed that according to a normative theory should not exist, the (apparently) simplest explanation is that a random measurement or performance error has occurred and that the behavior cannot be counted as evidence against the theory (Funder, 1987; Goeree & Holt, 1999). If defection is the normatively correct response in the PD, then cooperation is an error. Unless errors are built into the system, as they are in certain visual illusions that at first are impossible not to see, practice with the task coupled with swift and accurate feedback, should lead to an increase in normative responding. There is some evidence for this ideal. As repeated play in the PD erodes cooperation, players seem to gradually err less as they come to understand that defection is the dominating strategy (Dal Bó, 2005).

Still, the error hypothesis faces some difficulties. One difficulty is uncertainty as to the source of the error. By one account, the error lies within the cooperating participants who fail to understand the rules of the game (Andreoni, 1995). By another account, the error lies within a research design that fails to describe the game clearly (Binmore, 1999; Gintis, 2009). Although it has a long tradition in social psychology, the separation of person and situation effects is now somewhat anachronistic (Reeder, 2008). To say that the task is too difficult or that participants are not mentally equipped to solve it is ultimately the same thing (Krueger, 2009).

Another difficulty is that error is not distinguished from systematic bias. When interval-scaled measures contain random error, the aggregation of observations is a remedy. In his famous "vox populi" paper, Galton (1907) illustrated this principle by showing that the average estimate of the weight of an ox at a farm fair was astonishingly close to its measured weight, although the farmers' individual estimates varied greatly (see Larrick, Manes, & Soll, 2012, for a recent update). When responses are categorical, however, and only one response is deemed normatively correct, error is asymmetrical. It can go only in one direction and is therefore indistinguishable from systematic bias, which also can only point in this direction. Unlike bias, however, random error has an upper bound. With the normative probability of cooperation, p_c, being 0, a small nonzero value of p_c might stem from error or bias, but a value greater than .5 must reflect at least some bias.

Another version of the error hypothesis holds that cooperation in a one-shot anonymous game is not a random event but the result of a maladaptive "spillover" from experience in iterated games. In iterated games, players may learn that sustained mutual cooperation may be achieved, especially among players who pursue a strategy of reciprocity such as tit-for-tat. In ancestral environments, one-shot exchanges may have been rarer than they are now, and so "human cooperative mechanisms are not in equilibrium with our environment" (Burnham & Johnson, 2005, p. 130).

Like other revisionist accounts, neither form of the error hypothesis can explain why there is more cooperation in nice games than in nasty games.

Simpson's Paradox

Another recent cognitive hypothesis suggests that the error of cooperation in the PD is not the result of randomness or the lack of a psychological process but

[2]Colman (2003) and Colman et al. (2008a) rejected the idea that team reasoning is a special case of other-regarding social preferences. They pointed out that in a High–Low coordination game (e.g., when both players earn 10 points if they both choose "Left," both earn 5 points if both choose "Right," and earn nothing if their choices mismatch) subjective payoff transformations with *w* do not alter the structure of the game. They only increase payoffs. To solve a coordination game, it is necessary to estimate p_c. Neither social preference theories nor team reasoning provide a mechanism for this estimation (Krueger, 2008).

that it is the predictable result of a specific but faulty process. From this perspective, cooperation is a cognitive (or rather, behavioral) illusion. Chater et al. (2008) asserted that cooperators fall into the "statistical trap [of] Simpson's (1951) paradox" (p. 403). Simpson's paradox characterizes a situation in which an association that exists globally does not exist, or is reversed, locally (E. H. Simpson, 1951). Stated more concretely for the context of the PD, people have observed over time that their cooperative choices were followed by larger payoffs than their defecting choices. Remembering this, they now choose to cooperate in order to capitalize on this correlation. They believe, apparently, that the conditional probability of the other person cooperating given own cooperation is greater than the conditional probability of the other person cooperating given own defection.

How is it that individuals have observed a positive correlation between cooperating (vs. defecting) and reaping high (vs. low) payoffs? A necessary condition for this correlation to occur (and to be perceived) is that there is variation in the games' *K* ratios. As noted earlier, games with a high *K* ratio are "nice" and most people cooperate, whereas games with a low *K* ratio are nasty and most people defect. In any particular game, however, it is not true that the conditional probability of the other cooperating given the player's own cooperation is higher than the conditional probability of the other cooperating given the player's own defection. A single game lacks the source of variation (i.e., in *K*) that produces the correlation in the long run. For this reason, the past association between own and others' behavior is irrelevant and should be ignored, and the player should defect, honoring the sure-thing principle (Savage, 1954). Those who fail to ignore it are the victims of a cognitive illusion.

Chater et al. (2008) take a behaviorist approach to explain the illusion. Using a simple version of Thorndike's (1898) law of effect, they note that because there are more cooperators in nice than in nasty games, most individuals learn over time that their own behavior is that of the majority. When they cooperate, they are often rewarded by the cooperation of others; when they defect, they are often punished by the defection of others. Chater et al. then argue that people fail to mentally correct for the influence of confounding variables (i.e., game difficulty). Instead, they will simply repeat rewarded behaviors.

The Achilles heel of this learning hypothesis is that it tries to explain present cooperation with past cooperation without explaining why rates of cooperation differ between nice and nasty games in the first place. As a consequence, the Simpson argument involves the logical error of confusing the *explanandum* with the *explanans*. The argument then commits the psychological error of drawing a correspondent inference from cooperative behavior to the "cooperativeness" of the game (although the direction of this bias is opposite to the correspondence bias plaguing theories of morality).

Summary of the Review

Aiming to repair classic game theory by explaining cooperation in social dilemmas, the models reviewed here succeed only partially. The social-normative models are stumped by the anonymous one-shot PD. The validity of norm-based explanations is threatened by the possibility that people act only as if they had internalized these norms and by the possibility that these explanations are mere correspondent inferences from behavior to attitude. The validity of cognitive-bias explanations is threatened by the by the lack of process models convincingly showing that cooperation is an error or an illusion. The one problem that affects most models is that they fail to explain why people are more likely to cooperate in "nice" (high *K* ratio) than in "nasty" games (low *K* ratio). This failure at the theoretical plane is poignant because to the untutored player it is rather obvious that cooperation is a better strategy in the former type of game than in the latter type. Using this common intuition as evidence for any of these theories cannot work because it begs the question of *why* this intuition is so compelling to so many. It is the very phenomenon that needs to be explained.

In what follows, we focus on symmetrical, anonymous, two-person one-shot social dilemmas and the PD in particular. We propose a basic and general cognitive process that can predict a person's chosen strategy, be it cooperation or defection. If we succeed, we can appeal to the principle of parsimony. Yet, aside from Occam's razor, there is another reason for why a compelling model of choice in an anonymous one-shot game has merit. Consider a comparison between two scenarios: the simple scenario is an anonymous, two-person one-shot game and the complex scenario is an "onymous," multiple-person, iterated game with communication, promises, and opportunities for punishment. Suppose a compelling theoretical explanation exists for each. The explanation for the simple scenario is simple; the explanation for the complex scenario is complex. That is to say, the scope of the theory (*explanans*) matches the scope of the phenomenon (*explanandum*).

Now we ask how well either theory would do in a cross-matched scenario. Our contention is that there is a basic asymmetry. The simple explanation will likely account for at least a portion of the empirical data in the complex scenario, whereas the complex explanation will get bogged down in ambiguity when applied to the simple scenario. Because the complex explanation was fitted to an equally complex scenario many of its assumptions will be useless, like keys without a lock. Of those assumptions that do appear applicable, it is difficult to know which assumptions actually do the

explanatory work. It is rather like taking a model of the entire brain to explain how the amygdala works.

Social Projection: Inductive Reasoning at Work

From Postchoice Projection to Prechoice Decision Making

In our review of extant repair models, we presented the learning model (Simpson's paradox) last because it taps into an important psychological and statistical fact, namely, people's sensitivity to the association between their own choices and the choices of others. Drawing on the theory of evidential decision making (Davis, 1977; Grafstein, 1991, 2002; Jeffrey, 1983; Rapoport, 1966, 2003) and research on social projection (Ames, 2004; Ames, Weber, & Zou, 2011; DiDonato, Ullrich, & Krueger, 2011; Epley, Keysar, & van Boven, 2004; Krueger, 1998), we have developed an account of choice in experimental games that has both normative appeal and empirical fit (Acevedo & Krueger, 2004, 2005; Krueger, 2007; Krueger & Acevedo, 2005, 2007, 2008).

The projection model begins with the robust finding that cooperators and defectors both tend to believe that most others (about 60%) choose as they themselves do (Dawes, McTavish, & Shacklee, 1977; Messé & Sivacek, 1979). When this effect was first observed, the prevailing view was that social perceivers are victims of a "false consensus effect" (Ross, Greene, & House, 1977). According to this view, social projection is necessarily egocentric, self-serving, and irrational. Dawes (1989) refuted this idea. He proved that when information about others is lacking, individual players rationally believe that their own choice is that of the majority. Dawes's critical insight was that individuals who have only one piece of evidence (i.e., their own behavior), and whose evidence is different, *should* make different estimates regarding the prevalence of that behavior.

Bayes's Theorem offers a formal rationale for social projection (see Dawes, 1990, or Hoch, 1987, for how a regression-based account leads to the same conclusion). In the idealized situation, in which all prior values of p_c are the same (i.e., the principle of insufficient reason [to favor any one particular hypothesis over any other]; Laplace, 1783/1953), the aggregated posterior probability of the observed behavior is 2/3 (as given by the ratio $\frac{k+1}{n+2}$, where k represents the number of "successes," e.g., cooperation, and where n represents the size of the sample; see Dawes, 1989, for a mathematical derivation). Although it may be hard for individuals to see their own behaviors as random samples of one, that is what they are in statistical terms. Bayes's Theorem states that the first observation has an impact on probability revision that is larger than any that follow. Henceforth, we refer to the expected probability that the other person's choice (whichever it will be) will match the player's own choice as the probability of reciprocity or p_r.

Dawes's (1989) reconstruction of the consensus effect is now generally accepted. Both the cooperator and the defector are making rational estimates of p_c *given what they know*. The fact that their estimates differ, and that therefore at least one must be wrong, is a necessary implication of Bayesian induction. It does not negate their individual rationality; instead, it corroborates it. Dawes's original account does not explain, however, why the players chose the way they did. This is where the theory of evidential decision-making comes into play. It offers an answer by extending Dawes's argument from postchoice projection to prechoice projection.

Consider a player who has not yet selected a strategy in the PD. This player may ask two questions. First, if I cooperate, what will be my rational estimate of p_c cooperation? Second, what will be my estimate if I defect? An intuitive Bayesian will realize that her estimate of the probability of cooperation will be higher if she cooperates than if she defects. Assuming uniform priors before choice, the expected values of p_c are 2/3 and 1/3, respectively for assumed own cooperation and own defection. The player is entitled to believe that both estimates are rational. The question confronting the player is how to rationally choose between two strategies that yield discrepant estimates.

The theory of evidential decision making suggests that the player now assesses the expected value of cooperation and the expected value of defection in light of the two anticipated Bayesian estimates of p_c. Assuming correctly that the final decision—whichever it may turn out to be—will probably be the decision reached by most others, the player can assess expected values by multiplying the payoffs of the game with the probability of reciprocity, p_r, and its complement, 1-p_r. Specifically, $EVC = p_rR + (1 - p_r)S$; $EVD = p_rP + (1 - p_r)T$. We have reached our first substantive conclusion: the projection hypothesis can explain the nice-versus-nasty effect. As the *K* ratio increases, a lower level of projection is sufficient to yield an expected value of cooperation that exceeds the expected value of defection. The indifference point lies at $\frac{T-S}{T-S+R-P}$ or $\frac{1}{1+K}$ (Acevedo & Krueger, 2005; Fischer, 2009).

Illustrating the General Properties of the Model

The top panel (a) of Figure 2 shows how the indifference point declines as *K* increases. The bottom panel (b) shows the how the relative value of cooperation (i.e., the difference between the expected value of cooperation and the expected value of defection)

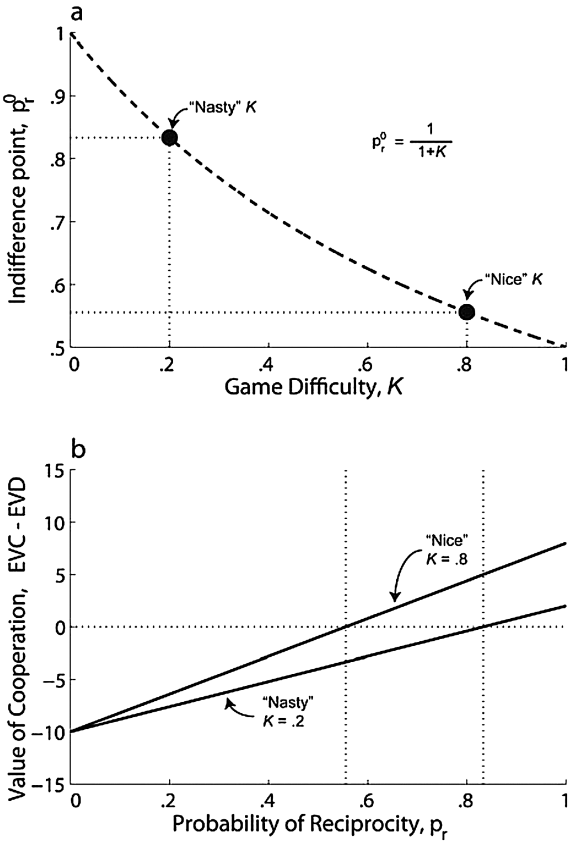

Figure 2. The indifference point of p_r as a function of game difficulty (K) [panel a] and the relative expected value of cooperation (EVC – EVD) as a function of the expected probability of reciprocity, p_r [panel b].

increases with greater expectations of reciprocity, and that the point of indifference is passed for the nice game before it is passed for the nasty game.

Figure 3 displays four quantitative properties of the model with illustrative and plausible parameter settings. In panel a, the expected probability of reciprocity, p_r, is shown as distributed over individuals. The distribution is normal, with a mean of .75, a standard deviation of .2, and it is censored[3] at 1. The probability of cooperation is represented by the area under the curve to the right of the indifference point. When $K = .5$, the indifference point is $p_r^0 = .6\overline{6}$. The probability of cooperation, p_c, is represented by the shaded region in the figure. It is the proportion of individuals whose strength of social projection exceeds the indifference point. The value of p_c is computed by evaluating the integral over the distribution of p_r from the indifference point to 1. That is, $p_c = \int_{x=p_r^0}^{\infty} p_r(x)\,dx$. In the present example, $p_r^0 = .67$. Assuming no change in any variable except the game's difficulty, panel b shows a plot of p_c against p_r. As the games become nastier (i.e., as K decreases and the indifference point p_r^0 increases), cooperation diminishes. Returning to the parameter default settings for the distribution of p_r, panel c shows that p_c increases with the expected probability, p_r. Finally, panel d shows that p_c decreases with increases in the standard deviation of p_r.

Finally, Figure 4 displays the most important implication of the evidential decision model. As the mean strength of projection increases, so does the total expected yield of the game (i.e., summed over the four possible outcomes), and it does so more steeply as the games payoff structure changes from nasty to nice. A group of strong projectors will both individually and collectively outperform a group of weak projectors, and they do so without caring about one another's welfare.

The Normative and Descriptive Appeal of the Model

In our model, the statistical equivalence of prechoice and postchoice projection justifies the former as a decision rule. If postchoice projection is rational, prechoice projection must also be rational. We consider this a matter of logical necessity. The fact that the former involves two different players (one cooperator and one defector), whereas the latter involves a single player in two potential future states, is irrelevant. By explaining the nice-versus-nasty effect, the projection model also explains the overall main effect of cooperation; it is simply the aggregate of the two extremes, or nice plus nasty divided by two.

It is important to stress that the social projection hypothesis does not predict that everyone will cooperate. Only individuals who project strongly enough to pass the indifference point will cooperate. Several studies have supported this proposition. Acevedo and Krueger (2005) directly manipulated the value of p_r and found that increasing it increased the rate of cooperation. Krueger and Acevedo (2007) found that individual differences in the strength of projection predicted cooperation in the PD. Acevedo and Krueger (2004) and Krueger and Acevedo (2008), respectively, found that individual differences in projection predicted voting intentions and intentions to contribute to a public good. Fischer (2009) manipulated the self-other similarity of players in a PD and found that cooperation increased as a function of that similarity.

These differences are important because early conceptualizations of the similarity hypothesis overstated its power. Rapoport (1973) reasoned that no player has grounds to infer that the other will act differently because the PD is symmetrical with regard to the effects each player can have on the other and because both players are assumed to be rational and to *know* that both are assumed to be rational. Given these assumptions, only mutual cooperation and mutual defection remain as reasonable outcomes, and mutual cooperation is clearly preferable. Empirically, the radical symmetry argument is falsified just as decisively as

[3]Each value greater than 1 is set to be equal to 1.

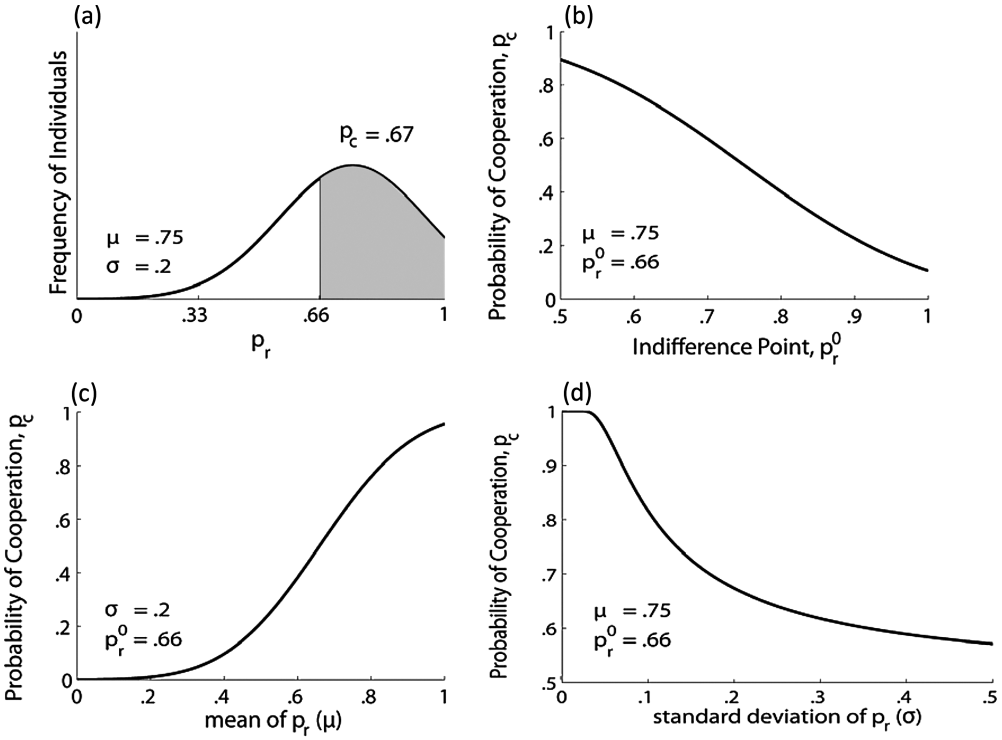

Figure 3. Four properties of the social projection model.

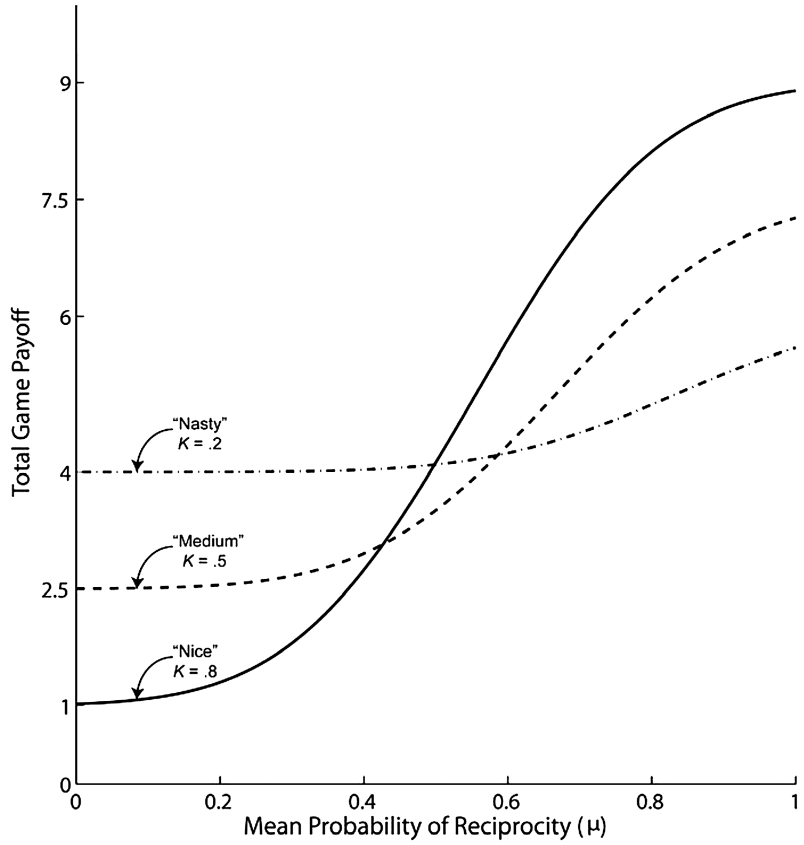

Figure 4. The total expected value of the PD as a function of the expected probability of reciprocity, p_r.

the classic defection-dominates argument. Like most other theories, it cannot explain the nice-versus-nasty effect. The social projection hypothesis presented here—and Fischer's SERS (subjective expected relative similarity; see also Anand, 1989)—solves the PD because it can explain this effect.

Critiques and Challenges

The rationale of evidential decision making has been questioned (e.g., Lewis, 1981), but the theory has never been refuted to our satisfaction. We now review and evaluate three common objections. The first objection is that there is only one true probability of cooperation. The player may not know it, but no matter. This is why according to the classic analysis, the player only needs to know that defection pays more than cooperation does irrespective of p_c. The rejoinder is that the player's own behavior is statistically tied to the behavior of others. A behavior is more probable if there is evidence for it than if there is no evidence for it. Once this implication of Bayes's Theorem is accepted, the process of decision making can proceed by estimating and comparing the expected values of cooperation and defection.

The argument that one must defect because defection dominates cooperation is analytically true, but so is the argument that one must choose whichever strategy has the larger expected value. Nozick (1969) stumbled upon this contradiction when discussing Newcomb's problem (to which we return later), and he responded to it as if it were a zen kōan. He cheerfully acknowledged that he had no idea how to deal with it. In our view, a new criterion must be introduced if two analytical truths cannot be reconciled. To do this, we turn to the idea of the long-run consequences of each choice strategy. Over the objection that the individual cooperator might be suckered, we find that cooperation increases the efficiency of the game.[4]

The second objection is that a person who generates two different probabilities of others' cooperation conditional on own behavior, and who then selects that behavior which permits the more favorable forecast, is guilty of magical (i.e., wishful) thinking (Quattrone & Tversky, 1984). In some contexts, the charge of magical thinking sticks (cf. Hastie & Dawes, 2010, chap. 2). In jury decision making, for example, verdicts are (and we better hope that this is true) positively correlated with guilt versus no guilt. Once jurors know this, they might decide to convict or acquit—whatever their punitive preferences demand—to infer that the conditional probability of guilt is respectively high or low. This inference betrays an unwarranted causal belief. In reality, a defendant's guilt or innocence cannot be altered by a jury's decision. In paternity testing, there is a correlation between men's confidence in being a child's father and actually being the father. Most men whose confidence is high receive positive test results, and the rate of positive results drops as confidence goes down. But, for a man with low confidence, not taking the test cannot increase the probability of being the father (Anderson, 2006).[5]

In the context of predicting what others will do in a social dilemma, however, the magical-thinking argument does not have traction. Players' inability to influence one another does not erase the statistical interdependence of their behaviors.[6] Consider the conditional statement that "if one player can cause another player to select the same strategy, then the players' strategies will be correlated." With *modus tollens* the antecedent (i.e., causation) is refuted by the denial the consequent (i.e., the correlation). Conversely, the consequent is not refuted by the denial of the antecedent. As any student of statistics must learn, correlation does not imply causation; correlation can occur without causation. Analytically and empirically, the existence of the self-other correlation is an inescapable fact. If one is to demand that individual players ignore the diagnostic power of their own choices—thereby accepting lower payoffs in the long run—one must find an argument other than the lack of causation.

The third objection is that if a player generates two discrepant probabilities of others' cooperation, these probabilities cancel each other out. The mindful player realizes that the estimate of p_c is higher when considering cooperation than when considering defection. An agnostic would say that both estimates are equally likely to be correct (or incorrect) and that therefore their average will minimize the estimation error. As the average is no longer conditional on the player's presumptive choice, defection returns as the dominating strategy. The problem with this argument is that averaging amounts to a negation of evidence and thus the nullification of the entire logic of induction.[7]

[4]We want to be clear that we are appealing to increases in efficiency only as an auxiliary argument to break the analytical stalemate. As noted in our discussion of morality-based decisions and team reasoning, we do not believe that the efficiency argument can stand on its own.

[5]When done considering all of our arguments, the gentle reader will understand why we believe that evidential thinking is magical in these examples but not in the context of social dilemmas.

[6]One must distinguish between two types of correlation. The correlation over players after choice is zero because the players act independently. However, the expected correlation within a player between own cooperation versus own defection with other's cooperation versus other's defection is positive. Likewise, the diagnostic ratio $\frac{p(\text{other's cooperation} \mid \text{own cooperation})}{p(\text{other's cooperation} \mid \text{own defection})} > 1$.

[7]Averaging one's own discrepant estimates increases predictive accuracy when both estimates can be modeled as the sum of a single true score plus random error (Herzog & Hertwig, 2009; Vul & Pashler, 2008). In prechoice projection, however, the discrepancy between the estimates is the systematic result of differential conditioning, not random error.

According to Bayes's Theorem, evidence taken from a single case can take only one form—cooperation or defection. The two behaviors cannot be sampled from the same individual who can act only once. Therefore, these estimates cannot be averaged.

To illustrate why we draw this conclusion, consider the averager's predicament. After generating an optimistic ($p_c = .6\bar{6}$) and a pessimistic ($p_c = .3\bar{3}$) forecast when respectively contemplating cooperation and defection, the player concludes that $p_c = .5$ is the best estimate. Now the player defects and wonders if this information can be used to revise the estimate. That is, the player has entered the postdecision stage, in which, as everyone knows, social projection is a rational strategy (Dawes, McTavish, & Shaklee, 1977). If the averaging argument is to be coherent, the answer must be "No!" because there is no new evidence. The player already considered both cooperation and defection in the predecision stage, and by averaging eliminated the diagnosticity of own choice. To bring this evidence back in the postdecision stage would be incoherent.

What happens when this player learns that another player, randomly drawn from the same population, has cooperated? By standard Bayesian logic, a defector should now believe that $p_c = .5$ because one cooperative and one defecting behavior are in evidence. Conversely, a cooperator should now believe that $p_c = .75$. If, however, the player had already chosen to ignore the diagnosticity of her own choice and assumed that $p_c = .5$, she must also ignore the other player's choice lest she be allocentric. By what rationale can one say, "I will discount my own behavior as evidence because I know I could have chosen differently; but I will generalize your choice to the majority because I have no reason to think that you could have chosen differently?" Such reasoning claims free will for the self while denying it to others, which makes it oddly egocentric. This kind of reasoning seems more magical than the reasoning described by Bayes's Theorem. If, for the sake of coherence, the choice of the sampled player's choice is ignored, then the choice of any other must also be ignored. A player committed to this kind of coherence takes a vow not to learn. We conclude that this type of coherence is irrational.

The (Ecological) Impossibility of Last-Minute Intrigue

Most disagreements over normative issues involve questions of theoretical coherence. Reviewing critiques of evidential decision making, we have focused on arguments raised from the classic game-theoretic point of view. We later return to normative questions and consider a broader sample of hypotheses. But coherence is only one criterion by which to evaluate a theory. Another criterion is whether a theory can make predictions that do not duplicate predictions made by competing theories and that receive unique empirical corroboration. In our view, far too much theorizing to explain cooperation in the anonymous one-shot PD has been post hoc. Indeed, the canonical game is so simple by design that it cannot discriminate between alternative theoretical accounts. The game must be modified to accomplish this task. In what follows, we develop a "last-minute-intrigue" variant of the PD to do just that, and briefly report the findings from three empirical studies. The modified game is designed to invite predictions from each of the reviewed theories and to yield a pattern of results predicted by social projection that does not duplicate the pattern generated by any other theory. The evidence will lead us to conclude that only the social projection hypothesis can account for the full range of the data.

The Modified Prisoner's Dilemma

The projection hypothesis (i.e., evidential decision making) starts with the premise that what an individual ultimately decides is probably the choice of the majority. This premise can be characterized as a truism—"most people are members of a majority"—but its implications are not obvious. Some may believe that the logic of induction can be tricked, that a resourceful player may find a way to slip into the statistical minority. Consider a player who has contemplated cooperation, and who has generated a high estimate of p_c. When this player decides to cooperate, she may be tempted to outsmart the implications of her own reasoning. What if she could switch from the intention of cooperation to the behavioral choice of defection so fast that her estimate of p_c would remain high? With a swiftly executed switch, she could perhaps hope to grab the Temptation payoff.

This idea is only hypothetical because the logic of inductive reasoning does not provide a time lag between evidence and inference. Consider a perceptual analogy. A motorist might wish to shift her gaze so rapidly that she can catch her own image in the rearview mirror with her eyes still on the road. Although logically impossible, the idea of last-minute intrigue (Brams, 1975) points to an opportunity for a unique empirical test of the projection hypothesis. In an experiment it is possible to create conditions that cannot occur ecologically.

Imagine a player who has chosen to cooperate, who believes that the other player will probably cooperate too, and who believes that the experiment is over. To her surprise, the experimenter offers a bonus option. The player may change her strategy, and she is assured that the other player's choice will remain whatever it is at this moment. In this bonus round, the player now faces a choice between the value of p_cR and the value of p_cT.

Table 1. *Prisoner's Dilemma: Predictions of Nine Theories in Four Conditions.*

	Initial Choice			
	Cooperation		Defection	
Type of Switching Theory	Unilateral	Bilateral	Unilateral	Bilateral
Classic	1	1	0	0
Morality	0	0	1	1
Reciprocity	0	0	0	1
Social values (benevolence)	0	0	.5	1
Team reasoning	0	0	1	1
Error	1	1	0	0
Simpson's paradox	1	1	0	0
Projection	1	0	0	1

Note. 1 = switch, 0 = hold.

As the latter is higher than the former, an exclusively self-regarding person will switch to defection.

This prediction is not unique; it is shared by classic theory, the error hypothesis, and Simpson's paradox. Table 1 lists the predictions derived from all hypotheses discussed here. Column 1 refers to the decision of unilateral switching after initial cooperation. The four hypotheses that advise against switching are the ones that seek to repair game theory by curtailing self-interest through the adherence to social norms (i.e., the hypotheses of morality, reciprocity, social values, and team reasoning).

Now imagine another player who, like the first one, has chosen to cooperate and who expects the other player to cooperate. This player also receives a bonus opportunity to switch, but is told that if she exercises the option, the other player's choice, whatever it is, will also be changed. The change will be automatic and will not require the other player's consent. Now, the projection hypothesis advises against switching because it would suggest trading the anticipated R payoff for the P payoff. In this case, the projection hypothesis is joined by the four social-normative hypotheses (Table 1, column 2). Classic theory, the error hypothesis, and Simpson's paradox advise defection no matter what.

The third case is analogous to the first in that a unilateral switch is offered. The player has, however, initially selected defection. Here, the projection hypothesis advises against switching because it implies trading in P for S. This advice is shared by most hypotheses (Table 1, column 3). The social-normative hypotheses are split. Morality, especially its deontological variant, demands cooperation. In contrast, switching would violate the tit-for-tat mindset of the reciprocity hypothesis. The social value (benevolence) hypothesis would encourage switching only for high values of the weight *w*, which is given to the other's payoff. Hence, the value of .5 entered in Table 1 expresses ambivalence with regard to switching. The team-reasoning hypothesis is aligned with the morality hypothesis. If switching is good for the team (i.e., if $T + S > 2P$), the player is expected to make the sacrifice.

The fourth case is analogous to the second. A bilateral switch is offered, but the initial choice is defection. The projection hypothesis recommends switching as a strategic choice. The expected change from 2P to 2R is not only good for the collective, but it is also good for the individual. This is a critical condition because it shows how a rational, self-interested player can support the collective interest. The four social-normative hypotheses agree with the choice of switching because they prize the collective good (Table 1, column 4). Classic theory, the error hypothesis, Simpson's paradox advise against switching because they favor defection no matter what.

Inspection of Table 1 reveals five discrete sets of predictions: Set 1 comprises the hypotheses calling for defection no matter what (classic, error, Simpson). Set 2 comprises the hypotheses calling for cooperation no matter what (morality, team reasoning). Sets 3 and 4 comprise one hypothesis each, namely, expected reciprocity and social value. Both these hypotheses predict that switching is rare. Finally, Set 5 contains only the social-projection hypothesis.

Chickens and Stags

All theories discussed here lay claim to explaining strategic decision making in more than one context. Although the PD is the most widely studied experimental game and perhaps the game with the greatest ecological significance, other games must be considered if only for the purpose of theory evaluation. The game of chicken (also known as the snowdrift or hawk-and-dove game; Maynard Smith & Price, 1973; Russell, 1959) and the stag hunt are similar to the PD, but there are important differences.

The distinguishing characteristic of the game of chicken is that the Penalty payoff is more catastrophic than the Sucker's payoff so that $T > R > S > P$ (Rapoport & Chamnah, 1966). This inequality leads to a reversal of the predictions derived from classic theory, the error hypothesis, and the Simpson's paradox (see Table 2). Although cooperation does not dominate defection, it does not render the worst outcome if things go badly. Players deciding by the minimax principle (the way of the chicken) cooperate to avoid the worst. Alternatively, they can cooperate with a probability of .5 by flipping a mental coin. Doing this, they maximize the total expected payoff in the game. Finally, if players feel they have enough information or intuition to estimate of probability of other's cooperation, they can choose to cooperate if $p_c = \frac{P-S}{R-S-T+P}$.

The social-normative hypotheses continue to favor cooperation, which means that in the game of chicken these hypotheses cannot be distinguished from the

Table 2. *Game of Chicken: Predictions of Nine Theories in Four Conditions.*

	Initial Choice			
	Cooperation		Defection	
Type of Switching Theory	Unilateral	Bilateral	Unilateral	Bilateral
Classic	0	0	1	1
Morality	0	0	1	1
Reciprocity	0	0	0	1
Social values (benevolence)	0	0	1	1
Team reasoning	0	0	1	1
Error	0	0	1	1
Simpson's paradox	0	0	1	1
Projection	1	0	1	1

Note. 1 = switch, 0 = hold.

classic theory they are designed to repair. Again, however, the projection hypothesis makes a unique prediction. It calls for unilateral switching after initial cooperation. The reciprocity hypothesis is also unique in that it discourages unilateral switching after initial defection.

The distinguishing characteristic of the stag hunt is that the Reward payoff is more desirable than the Temptation payoff so that R > T > P > S. Compared with the predictions derived from the PD, the stag hunt payoff ranking leads to no change in the predictions of the competing theories except that a person with social preferences (i.e., benevolence) will more readily accept the opportunity to perform a unilateral switch from defection to cooperation (see Table 3). The projection hypothesis no longer predicts a unilateral switch after cooperation, as it did in the PD and in the game of chicken. Hence, the projection hypothesis is aligned with the reciprocity hypothesis and, as a consequence, does not predict a unique pattern. Therefore, the stag hunt is not suitable for an empirical test.

Before moving on to review the evidence obtained from studies with the last-minute-intrigue paradigm,

Table 3. *Stag Hunt: Predictions of Nine Theories in Four Conditions.*

	Initial Choice			
	Cooperation		Defection	
Type of Switching Theory	Unilateral	Bilateral	Unilateral	Bilateral
Classic	1	1	0	0
Morality	0	0	1	1
Reciprocity	0	0	0	1
Social values (benevolence)	0	0	1	1
Team reasoning	0	0	1	1
Error	1	1	0	0
Simpson's paradox	1	1	0	0
Projection	0	0	0	1

Note. 1 = switch, 0 = hold.

three additional points bear noting. First, only the projection hypothesis, by its sensitivity to K, predicts that if the same numerical payoffs are used, there should be less cooperation in the PD than in the game of chicken or the stag hunt. In the latter two games, the rate of cooperation should be the same. We know of no comparative studies that have directly tested this hypothesis, let alone a meta-analysis. Hence, we submit this hypothesis for future research.

Second, when the payoff matrix is symmetrical in the sense that T – R = P – S, the projection hypothesis predicts a nice-versus-nasty effect in the PD, but not in the game of chicken (where T – R = S – P) or the stag hunt (R – T = P – S). Figure 5 shows a simple numerical arrangement to illustrate this point with a nice game displayed in panel (a) and a nasty game displayed in panel (b). If, however, the matrix is asymmetrical, the projection hypothesis yields more predictions. In panels (c) and (d), the highest payoff is doubled, which makes the PD and the game of chicken nastier while making the stag hunt nicer. We now also see a nice-versus-nasty differential within each game. The PD differential is positively correlated with the chicken differential but negatively correlated with the stag differential. It pays to look at the specifics of the payoff matrix. Literally.

Third, social distance has a strong and well-documented effect on cooperation and other forms of behavior beneficial to others. To date, the best-fitting description of this phenomenon is that the rate of benefitting others drops off hyperbolically, that is, it falls sharply at first and more shallowly later (Jones & Rachlin, 2006, 2009). This drop-off in giving or investing is typically interpreted from a moral point of view, be it inspired divinely or by Darwin (Burnstein, Crandall, & Kitayama, 1994). People care less about distant others than about close others because they are just too far away, not genetically related enough to the self, not available for mutually gainful trade or exchange, or simply because there are too many of them (e.g., there may be 100 people living on my street but 100,000 in my town). As distance grows, and with it the number of people who inhabit it, so grows the number of others who might cooperate with or help those others. Cooperation and altruism become somebody else's problem, or, from a game-theoretic perspective, a volunteer's dilemma (Diekmann, 1985; Fischer et al., 2011; Krueger & Massey, 2009).

The classic theory, the error hypothesis, and learning theory (Simpson's paradox) are easy prey to the social distance effect. They do not see it coming, and they claim it should not exist. People should, after all, defect no matter what. They cannot cooperate with the rate of less then nothing as social distance grows. The social normative theories favor cooperation and with some additional adjustments, some of them can accommodate the social distance effect. Pure morality, of course,

Moderate Temptation

a "Nice"

Payoff	Game		
	Prisoner's Dilemma	Chicken	Stag
T	10	10	9
R	9	9	10
P	1	0	4
S	0	1	1
K	$.8$	1	1
p_r^0	$.5\bar{5}$	.5	.5

b "Nasty"

Payoff	Game		
	Prisoner's Dilemma	Chicken	Stag
T	10	10	6
R	6	6	10
P	4	0	4
S	0	4	0
K	$.2$	1	1
p_r^0	$.8\bar{3}$	.5	.5

Extreme Temptation

c "Nice"

Payoff	Game		
	Prisoner's Dilemma	Chicken	Stag
T	20	20	9
R	9	9	20
P	1	0	4
S	0	1	1
K	$.4$	$.47$	$.2\bar{1}$
p_r^0	$.71$	$.67$	$.32$

d "Nasty"

Payoff	Game		
	Prisoner's Dilemma	Chicken	Stag
T	20	20	6
R	6	6	20
P	4	0	4
S	0	4	0
K	$.1$	$.37$	$2.\bar{6}$
p_r^0	$.9$	$.72$	$.27$

Figure 5. Nice and nasty games of chicken and stag hunt.

cannot account for the effect because it presumes to be universal. Reciprocity can point to the reasonable expectation that interaction becomes less likely with increasing social distance. The social value hypothesis must find a way to predict why benevolence (or concern about fairness) decreases with increasing social distance. In other words, this hypothesis must find an additional principle from outside its own purview to account for the result. Team reasoning may raise the ecological point that most teams are local and that team reasoning, by definition, does not extend beyond the team. This is, of course, an inelegant solution, because it begs the question, Why not include everyone in the team?

The problem of the social normative theories is that they can explain only how social norms sustain moral behavior within the group. They cannot explain why or how people choose certain degrees of inclusiveness (from the neighborhood to the world). Explanations for how and why people do this may be imported, but they tend to be post hoc and not part of the theory itself. To date, the conflict between parochial and universal morality remains a hot-button issue (Singer, 1993).

The social projection hypothesis offers a simple and ecologically viable explanation for the social distance effect. As social distance increases, or as categories become more inclusive, a random other person will become less similar to the self (Krueger & Zeiger, 1993). Hence, the probability that the other person will behave like the self, p_r, shrinks, which raises the threshold for cooperation. Social distance thus acts directly on the key parameter of the projection hypothesis. Importing extraneous mechanisms is unnecessary. An implication of the social-distance hypothesis is that people project less to outgroups than to ingroups (DiDonato et al., 2011; Robbins & Krueger, 2005). This differential in projection correctly predicts that people are more willing to trust and cooperate with members of

an ingroup than with members of an outgroup (Brewer, 2008; Buchan, Croson, & Dawes, 2002; Dawes et al., 1988; Foddy, Platow, & Yamagishi, 2009; Tanis & Postmes, 2005; Yamagishi & Kiyonari, 2000).

Evidence for the Last-Minute Intrigue Hypothesis

To test the unique predictions of the social projection hypothesis in the last-minute-intrigue paradigm, we set up four types of scenario by varying target player's reported choice (cooperation vs. defection) and the type of switching offer (unilateral vs. bilateral) orthogonally. Participants took the role of observer and advisor, predicting whether the player would accept the offer and judging whether she should take the offer. Half the participants made judgments in the context of the prisoner's dilemma, and the other half made judgments in the context of the game of chicken. We reasoned that the projection hypothesis would be supported if its predictions provided the best fit with the empirical data in both contexts. We conducted three studies, the results of which we sketch here. A more complete report can be obtained from the authors.

In the first study, the participants received the following vignette:

> Suppose two players are involved in a game, in which they can win various amounts of money. Each player has a coin, which he can place Heads up or Tails up. The $ amount he receives depends on his own choice and on the other player's choice. The players do not know each other, they cannot see each other, and they cannot communicate with each other. There are four possible outcomes of the game.

For the PD, participants were told that the player "P" and that the other player "O" would each get $9 if both chose Heads, that P would get $0 and that O would get $12 if P chooses Heads and O chooses Tails, that P would get $12 and that O would get $0 if P chooses Tails and O chooses Heads, and that P and O would both get $3 if they both choose Tails. For the game of chicken, the payoffs were modified as required.

The next paragraph stated that P "has chosen HEADS (TAILS)" for cooperation or defection, respectively, and that "he thinks there is an 80% chance that [player] O chose as he himself did. He is waiting for his payoff to be announced. Now the experimenter appears and offers P the opportunity to switch from Heads to Tails [Tails to Heads]." In the condition of unilateral switching, the description then stated that "the experimenter assures P that O does not have the opportunity to switch. Whatever O chose is now locked in." In the condition of bilateral switching, the description stated that "the experimenter explains that if P switches, O will switch too. Whatever O has chosen will be reversed." Using a 4-point scale, participants then judged whether they thought P would switch and whether he should switch. Responses to these two questions formed a composite index because they were highly correlated. Participants responded to all four vignettes of the design.

The results uniquely supported the projection hypothesis. Recall (and see in Table 1) that for the PD only the projection hypothesis predicted a full cross-over interaction between initial choice and type of switch. This was the obtained pattern, $\eta_p^2 = .28$, $p < .001$. Switching was endorsed by 62.5% of the respondents in the cooperation/unilateral condition, by 27.5% in the cooperation/bilateral condition, by 45% in the defection/unilateral condition, and by 72.5% in the defection/bilateral condition.

For the game of chicken, the projection hypothesis called for an asymmetrical interaction, and this was found, $\eta_p^2 = .22$, $p = .002$. As predicted, the dichotomized data showed that switching was endorsed by 70% of the respondents in the cooperation/unilateral condition, by 32.5% in the cooperation/bilateral condition, by 77.5% in the defection/unilateral condition, and by 82.5% in the defection/bilateral condition.

These results suggest that participants understood and let themselves be guided by the complex interplay of a player's expectations, his initial choice, and the type of switching offered. Expectations could have been easily ignored because they were invariant over conditions, and the type of switching could have been ignored or confused because it was too complicated. In other words, the strong demands made on the participants and the multiple opportunities to foul up stacked the deck against the projection hypothesis by making it more likely for participants to fall back on simple heuristics such as "Always cooperate!" or "Always defect!"

In the second study, the level of projection was no longer provided; the respondents had to estimate it themselves. Our reasoning was that providing a fixed $p_r = .8$ makes it likely to detect effects of projection *if* people are attuned to the implication of projection at all. A test of the projection hypothesis is more conservative if $p_r < .8$, and it is riskier if participants provide an estimate of projection themselves. We anticipated, however, that participants would attribute social projection to others whose choices they knew (Krueger & Zeiger, 1993). People intuitively understand that cooperators expect others to cooperate and that defectors expect others to defect (Krueger & Acevedo, 2007). Given these considerations, our central hypothesis was that the interaction between initial choice and type of switching would reappear but that the effect size would be smaller than in the first study.

The pattern predicted by the projection hypothesis emerged again ($p = .001$) and, as predicted, its effect size was reduced (from .28 to .12). Switching

was endorsed by 75% of the respondents in the cooperation/unilateral condition and by 45.5% in the cooperation/bilateral condition. Conversely, switching was endorsed by 35.9% in the defection/unilateral condition and by 46.5% in the defection/bilateral condition.

We had also predicted that the average expected projection would be greater than .5 (i.e., the point of no projection) but smaller than .8. All four of the mean estimates were close to the .6 mark, and all were significantly larger than .5 (cooperation/unilateral: .57, cooperation/bilateral: .58, defection/unilateral: .60, defection/bilateral: .59).

In a final study, we sought to open a new empirical window into the rational and moral aspects of decision making in the PD. The theories considered here differ not only in their predictions regarding rational choice but also in how they construe the relation between rationality and morality. According to classic game theory, rationality and morality are incompatible (Danielson, 1998). The PD forces a choice between the two. The rational agent defects, whereas the moral agent cooperates. Incompatibilism means that agents cannot claim both virtues at the same time. Consideration of the four payoffs illustrates the conflict. Over payoffs, cooperation is negatively correlated with a player's own payoffs (–.45) and positively (and more strongly) correlated with the other player's payoffs (.89). Hence, the two players' payoffs are negatively correlated (–.8) with each other, which signals their conflict of interest.

The social-normative theories take a compatibilist view by making no clear distinction between rationality and morality. Instead, they prioritize cooperation as the moral choice, and seek to justify this choice by rationalizing it (Pinker, 2008). When both players cooperate, they are lauded not only for their morality but also for aligning themselves with "collective rationality" *sensu* Durkheim (1895/1982; see Segre, 2008, for a recent review of Durkheim's work and its relevance for rational choice theory).

The social-projection hypothesis is also compatibilist, but it prioritizes individualist rationality. As previously noted, this hypothesis assumes that people seek to maximize their individual gains, and they do so by rationally expecting others to be like themselves. When both players cooperate, they might do so because they have estimated the expected value of cooperation to be higher than the expected value of defection. They rationally choose the strategy that is also regarded as the moral one. This view may be taken as a response to Russell's and Hardin's position that self-interest ought not be discounted too lightly as the main motivational principle. With social projection, egocentrism and self-interest remain of primary relevance; people can cooperate and be rational and moral at the same time.

The final study approached the question of rationality and morality from the observer's perspective. Earlier research supports classic incompatibilism. People judge cooperators as more moral than defectors and defectors as more rational than cooperators (Krueger & Acevedo, 2007). Moreover, the difference in judged morality is larger than the difference in judged rationality, a finding that is consistent with contemporary research on the two-factor theory of social judgment (Cuddy, Fiske, & Glick, 2008) and with the revisionist view that people care more about collective than individual outcomes. Against this empirical background, the predictions derived from the social-projection hypothesis are risky (and thus scientifically useful; Krueger & Funder, 2004; Popper, 1963). This hypothesis assumes that people associate rationality, but not necessarily morality, with a specific pattern of switching (or holding onto) one's choice.

The vignettes resembled the ones used before. The game was described, complete with information about payoffs, the level of projection, and the type of switching bonus available to the player. The additional piece of information was whether Player P accepted or rejected the offer to switch. Participants were asked to contemplate what they had learned and then to judge P on a series of trait adjectives reflecting rationality or morality. This design afforded an opportunity to study the relationship between perceptions of rationality and perceptions of morality in the context of the PD. More important, we were able to test specific hypothesis as derived from the contending theories of choice.

Hypotheses About Perceived Rationality

The theory of evidential decision making suggests that four players would be seen as rational: a player who (a) switches unilaterally to defection, (b) does not switch bilaterally to defection, (c) does not switch unilaterally to cooperation, or (d) switches bilaterally to cooperation. Notice that these are the four decisions that participants endorsed in the first two studies. In the other four conditions, judgments of rationality should be low.

Considering all theories at play, we identified five discrete patterns of predictions. The first pattern is unique to the social-projection hypothesis. The second pattern comprises classic game theory, the error hypothesis, and Simpson's paradox. These hypotheses state that judgments of rationality should be high for any player whose final choice is defection. The third pattern comprises moral norms theory and team reasoning. These hypotheses state that judgments of rationality should be high for any player whose final choice is cooperation. The fourth pattern represents the predictions of expected reciprocity. This pattern is similar to previous one, except that reciprocity does not demand a unilateral switch to cooperation. The fifth and final pattern represents the social value of benevolence. This pattern resembles the other two norm hypotheses

(Patterns 3 and 4), except that a unilateral switch from defection to cooperation depends on the strength of benevolence.

Hypotheses About Perceived Morality

As all theories under contention seek to explain rational choice in the PD, the critical hypothesis tests refer to judgments of rationality. Yet we also explore what the theories imply for judgments of morality. A heuristic way of generating hypotheses is to consider whether a theory regards morality as compatible or as incompatible with rationality. On the grounds of theories falling into the incompatibilist camp (classic, error, Simpson's), one might expect a negative correlation between judgments of morality and judgments of rationality. In contrast, one might expect a positive correlation on the grounds of theories in the compatibilist camp (morality, social values, team reasoning, reciprocity). The social-projection hypothesis is compatibilist in a weak sense. This hypothesis merely suggests that the correlation between judgments of morality and judgments of rationality is not negative. The correlation may be positive but it could be weak. On this view, the morality of cooperation is merely a welcome by-product of its rationality rather than an evaluation that is equally important. An important corollary of this hypothesis is that across conditions, the variation of judgments of rationality will be greater than the variation of judgments morality. Note that this prediction contravenes a prediction derived from the two-factor theory of social perception (Cuddy et al., 2008).

Participants received vignettes describing the PD as in the first study with information about the player's decision to accept of reject the switching offer added. Then they rated the player on these trait-descriptive adjectives: "ethical," "intelligent," "naive," "rational," "selfish," and "trustworthy." The words *intelligent, naive* (reverse scored), and *rational* formed a scale tapping into rationality, whereas the words *ethical, selfish* (reverse scored), and *trustworthy* formed a brief scale tapping into morality.

Findings

The social projection hypothesis specified aninteractive pattern involving all three independent variables(initial choice, type of switching, and final decision. The empirical pattern conformed to the predicted one, $\eta_p{}^2 = .18$, $p < .001$. An initial cooperator who accepted a unilateral opportunity to switch ($M = 4.71$) or who rejected a bilateral opportunity to switch ($M = 5.56$) received high ratings of rationality. Likewise, an initial defector who rejected a unilateral switch ($M = 5.59$) or who accepted a bilateral switch ($M = 4.69$) was seen as rational. Conversely, an initial cooperator who rejected a unilateral switch ($M = 4.62$) or who accepted a bilateral switch ($M = 3.25$) was rated low on rationality. Likewise, an initial defector who accepted a unilateral switch ($M = 3.73$) or rejected a bilateral switch ($M = 4.20$) received low rationality ratings. An analysis of perceptions of morality only yielded an effect of final decision, $\eta_p{}^2 = .11$, $p < .001$. Players who held ($M = 4.04$) were perceived as more moral than players who switched ($M = 3.26$).

Recall that the classic hypothesis is incompatibilist in that it assumes rationality and morality to be negatively related. In contrast, the social-norm hypotheses assume a positive association. Across participants, the correlation between the two dimensions of judgment did not differ significantly from zero ($r = –.14$). Social projection is the only hypothesis consistent with this finding (which can count, admittedly, only as weak additional support for this theory).

To quantify the goodness-of-fit of each theory with the average judgments of rationality, we translated each hypothesis into a pattern of 1s (if the predicted judgment was high) and 0s (if the predicted judgment was low). Four of these patterns were duplicates of others, leaving five distinct patterns. The first, representing classic game theory, the error hypothesis, and Simpson's paradox, did not predict average rationality judgments ($r = –.12$). The second pattern, representing the morality and the team reasoning hypotheses, failed as well ($r = .12$). The reciprocity hypothesis did better ($r = .72$), and the social value hypothesis had some success ($r = .48$). In contrast, the association between the mean ratings and the social-projection pattern (and the descriptive version of Simpson's paradox) was nearly perfect ($r = .91$). We supplemented these analyses by asking whether the player's response to the switching offer predicted perceptions of rationality. The result was a medium effect of players upholding their initial choice being perceived as more rational ($r = .39$).

As the predicted patterns showed some degree of overlap with one another, we simultaneously regressed average judgments on the five hypothetical patterns. Now, the social-projection pattern emerged as the only significant predictor ($\beta = .71$).

We also found evidence for the idea that average judgments of morality would be less variable across conditions ($SD = .44$) than average judgments of rationality ($SD = .84$). This finding runs counter to the commonly observed primacy of the morality dimension, but it is consistent with our expectation that the possibility of last-minute intrigue would stimulate observers to ask which decision would make the most sense for the player.

Other hypotheses regarding morality judgments failed. The finding that neither initial cooperation ($r = –.12$) nor final cooperation ($r = .12$) predicted judgments of morality casts further doubt on the social norm hypotheses. The only variable that predicted perceived morality was the player's response to the offer

to switch. Those who stuck to their initial choice were perceived as more moral than players who switched ($r = .80$). In hindsight, it appears that people follow an intuition that says "one ought not change one's mind in an interdependent situation when the other person does not have the same opportunity." This intuition applies even to the condition of bilateral switching because only Player P, but not Player O, had the luxury to take or to reject the offer.

After the Evidence

Because the failure of classic game theory to account for cooperation in the prisoner's dilemma (and other games) has become undeniable, a variety of revisionist theories have been proposed. Most of these theories seek to overcome the conflict between rational and moral choice by redefining the former as a case of the latter. We have expressed doubt concerning these efforts. Social norm theories may do more to explain the problem of cooperation away than to explain why people cooperate.

To make a fresh start, we propose that the social-projection hypothesis, as an instance of the theory of evidential decision making, provides a perspective with both normative and descriptive appeal. Our main theoretical point is that, because social projection is recognized as a rational inference strategy for individuals who have made a decision, it must also be acknowledged that social projection is a rational strategy for individuals who are still in the process of deciding. We argue that the statistical logic for a single individual who anticipates making one of two decisions in the future is exactly the same as it is for two individuals who have made different decisions. It follows that the rejection of single-person predecision projection would entail the rejection of two-person postdecision projection. This, in turn, would amount to a rejection of Bayes's Theorem.

We found a pattern of results that was uniquely predicted by the social-projection hypothesis. We found that participants are sensitive to the joint implications of social projection, initial choice, and the laterality (uni- vs. bi-) of the switching option. Their judgments revealed close adherence to the principle of payoff maximization. To recapitulate a key result, consider the condition of initial defection and bilateral switching. Most participants recommended switching. Given projection, the revised expected outcome was mutual cooperation. From the classic perspective, however, any switching from defection to cooperation is irrational. The collectivist theories recommend a switch, but for the sake of morality. Yet, these theories also recommend unilateral switching to cooperation. A rational person would not switch in this condition—which is what we found.

What is the proper role of morality in two-person games? We caution against one-to-one inferences from socially desirable behavior (cooperation) to correspondent moral intentions or dispositions. People can act in socially desirable ways for egocentric or even selfish reasons (Maner et al., 2002). Inferences of morality should be discounted if other personal attributes (i.e., rationality) also explain the outcome (Krueger, 2009; McClure, 1998). Inasmuch as the individual actions are what matters to collective welfare, such discounting of morality is just as well. On this view, moral intentions are an added psychological benefit.

Back to the Normative Question

In the opening section, we offered a normative justification of evidential decision making by raising and disputing three objections from the perspective of classic game theory. After showing that only the social-projection hypothesis, but not the classic hypothesis or the social norm hypotheses, passes empirical tests, we return to the task of analytic theory evaluation. We begin with comparisons between the classic and the evidential hypotheses and continue with comparisons involving other repair hypotheses.

Taking Determinism Seriously

A critical difference between the classic and the evidential approach to social dilemmas lies in the conceptualization of "choice." Under the classic view, choice is free. The player is free to contemplate the sure-thing principle and act on it. He is not constrained by the statistical linkage between his choice and the choice of the majority. Even a player who knows that he belongs to a population consisting of 99% cooperators is presumed "free" to put himself in the 1% minority. Statistics, as they say, do not apply to the individual. By contrast, the evidential view is that acts of will cannot undo this statistical dependency. To repeat: The evidential view does not claim that individuals can cause noninteracting others to choose as they themselves do. Instead, this view assumes a common-cause model (Reichenbach, 1956). The behaviors of different individuals who find themselves in the same situation are correlated because of a common causal influence the situation exerts on them. Indeed, the identification of these forces is the research program of social psychology (Krueger, 2009). Whether a player believes to have free choice is beside the point.

Newcomb's problem, which has been called a one-player PD (Brams, 1975; Lewis, 1979), poignantly illustrates the polarity of determinism and free will. A Newcomb player is asked to open either one or both of two boxes. If he opens only one box, and if a near-omniscient demon predicted that he would, that demon

placed $1 million in that box. If the demon predicted that the player would open both boxes, he left that box empty. The other box always contains $1 thousand. According to the classic view, the player should—"freely choose to"—open both boxes. The logic is the sure-thing principle. Whatever the demon predicted, a two-boxer is better off by $1 thousand than a one-boxer.[8]

The classic view treats the near-omniscience of the demon as irrelevant. The player is advised to open both boxes because he has no causal power—retroactive or otherwise—over the demon's prediction. In contrast, the theory of evidential decision making does not ignore the demon's record of forecasting. The player is advised to open one box because this choice reveals, although it does not cause, the demon's prediction.

The notion of determinism is writ large in Newcomb's problem. Many scientists may envy the demon's predictive powers, but most embrace the idea of determinism (see Baumeister, 2008, for a dissenting view). His tendency to project from himself to others notwithstanding, Poincaré's (1914/1996) stated the case for determinism lucidly in his *Science and Method*: "Every phenomenon, however trifling it be, has a cause, and a mind infinitely powerful and infinitely well-informed concerning the laws of nature could have foreseen it from the beginning of the ages" (pp. 64–65). Poincaré drew an important distinction between determinism and lay ideas of causation by noting that the former is bidirectional: "The laws of nature link the antecedent to the consequent in such a way that the antecedent is determined by the consequent just as much as the consequent is by the antecedent" (p. 70). With bidirectional determinism (see also Ayer, 1956; Russell, 1913), a player cannot presume to choose independently of the demon's predictions just as he cannot presume that the demon can make predictions independent of the player's choices (Bar-Hillel & Margalit, 1972).

As in the PD, about half of Newcomb players follow the logic of evidential decision making (Krueger & Acevedo, 2005; Shafir & Tversky, 2004). Of interest, the evidential choice is not confounded with the moral choice in Newcomb's problem. Opening only one box does not contribute to a collective good and it is not presumed to be a kindness shown to the demon (e.g., to save him money). If there is a moral overtone, it is reversed with respect to the PD. If two-boxing is attributed to the exercise of "free will," and if free will is a prerequisite of morality, then two-boxing is both classically rational and moral. Unlike social norm theories, the theory of evidential decision making has no trouble giving a coherent normative and descriptive account for both games.[9]

Giving Advice

As coherence is a core criterion of rationality (Dawes, 1998; Krueger, 2012), it can guide further explorations into the normative status of evidential decision making. According to the classic view, principled defection in the PD (and 2-boxing in Newcomb's problem) is coherent, and thus rational. A player who knows that he will be better off regardless of what others do must defect before knowing what others do. If he cooperated, he would be in violation of the sure-thing principle. Choices derived from social projection are also coherent, however, because they satisfy Bayes's Theorem. One ought not ignore evidence even if it consists of a single observation. To be coherent, one must apply this logic to comparisons between two different players postchoice, and to a single player in two different states prechoice. In short, the coherence criterion per se does not appear to break the impasse.

The difference between the classic and the evidential view comes into focus if one considers the implications for advice giving. Both theories generate prescriptions for how a player should choose. As the PD is usually construed as a problem of individual decision making, the normative prescription is typically framed as advice given to an individual. If the individual defects, she is better off because $T > R$ and $P > S$. Coherence demands that the same advice be given to all players. If all defect, the result is an inefficient Nash equilibrium ($2P < 2R$). Empirical results confirm the self-defeating nature of classic advice. Students trained in classic game theory are more likely to defect than naive students (Frank, Gilovich, & Regan, 1993). If the trained students play among themselves, they do not do well individually or collectively. To paraphrase Bertrand Russell, Hell holds a special place for game theorists, where they are condemned to play one-shot PDs with one another, forever.

The reason why game-theoretically savvy players are likely to defect is that they are no longer naïve about the value of p_c. Assume that the probability of cooperation is low a priori, a contemplation of their own potential cooperation has little effect, and the expected value of defection remains greater than the expected value of cooperation. In contrast, the premise of the evidential view is that players are ignorant of p_c. In

[8]Newcomb's problem can be seen a time-reversed trust game (Evans & Krueger, 2009). The player opens one box if she trusts that the demon filled it. In a regular trust game, the trustee may choose to reciprocate trust, in part, because she honors the trustor's decision to accept vulnerability. In Newcomb, this path to reciprocity is blocked. Trusting a demon should thus be harder than trusting a person. The common cause model, of course, does not care.

[9]To players beliefs matter. Vohs and Schooler (2008) found that belief in determinism (vs. free will) increased the likelihood of cheating on a test, presumably because these participants felt exempted from the moral responsibility they associated with free choice. The social-projection hypothesis suggests the opposite in noncooperative experimental games. Here, belief in determinism should help players to feel exempted from the demands of classic rationality.

other words, the evidential theory models the PD in its pure form, which allows no information aside from the payoffs.

Now recall that Newcomb's problem takes the premise that the demon rarely makes a mistake. A player who opens both boxes (i.e., defects) may think he is honoring the sure-thing principle, but given the premise of demonic foresight, he must conclude that the demon predicted his cleverness and accordingly left one box empty. Given the description of the problem, the player cannot, on his own, capitalize on classic rationality without negating part of the definition of the problem. In terms of the last-minute-intrigue paradigm, the player cannot unilaterally defect.

What about advice giving in Newcomb's problem? Suppose the player contemplated the task and decided to open only one box (i.e., to cooperate). Now the classic advisor urges him to switch and open both boxes. Again we need to distinguish advice given to one player from advice to everyone. If the advice is given to only one player, and if one assumes that the advisor's intrusion into the process was not foreseen by the demon (i.e., assuming that arrangements were made to enable unilateral switching), the player may gain. However, the advice to open both boxes must be given to everyone if it is to be coherent. This maneuver will negate the premise of the game, that is, the demon's phenomenal accuracy.

The top panel of Figure 6 shows plausible numbers for an intact game. Of 100,000 players, half open one box and half open both boxes. Within each group, 99% of the demon's predictions are correct. The conditional probability of a particular choice C given the corresponding prediction P is .99. The demon's overall accuracy is also 99% because of the symmetry of the marginal frequencies. The correlation between choice and prediction is $\Phi = .98$.

The bottom panel shows the effects of a classic advisor's intervention. Suppose all but 100 of the original one-boxers have been persuaded to switch to two-boxing. If the switchers are randomly sampled, the expected result it that 99 one-boxers remain whose choice the demon correctly predicted. The largest increase is in the group of two-boxers whose choice the demon did not foresee (from 500 to 49,901). The probability of one-boxing given the demon's prediction of one-boxing is now .00198, although the probability of two-boxing given the demon's prediction of two-boxing is .99998. The overall proportion of correct predictions has dropped to .5005, and the correlation between choice and prediction is $\Phi = .03$.

The demon's accuracy is the conditional probability of a player's choice C given the demon's corresponding prediction P, or p(C|P). Yet the player is trying to estimate the inverse conditional probability, namely, the probability that the demon's prediction will turn out to match the player's is choosing, or p(P|C). Bayes's

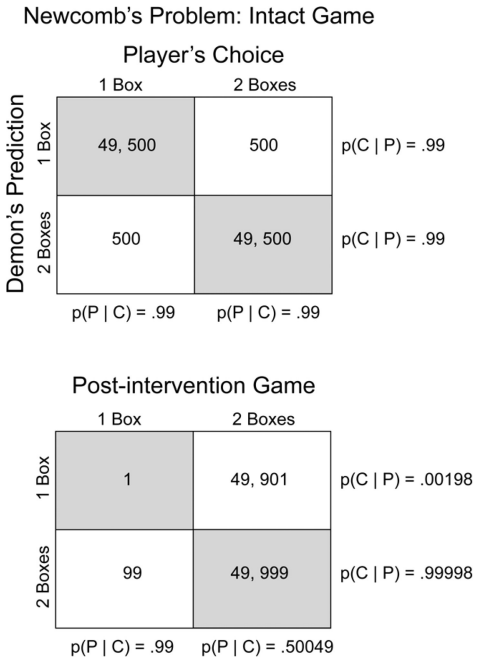

Figure 6. Newcomb's problem before and after advice giving.

Theorem entails that the two inverse conditional probabilities are the same only if the two base rates are the same. Might it not be possible that even though the demon is highly accurate, p(P|C) is much lower? People often equate inverse conditional probabilities without regard to differences in the base rates (Dawes, Mirels, Gold, & Donahue, 1993; Krueger, 1996; Levi, 1975). This bias could falsely boost players' optimism.

Upon reflection, these worries are unfounded if p(C|P) is assumed to be very high for both one- and two-boxing. Suppose $p(C_1|P_1) = p(C_2|P_2) = .99$, but also suppose the demon is heavily biased toward two-boxing, that is, $p(P_2) = .9$. A player who opens only one box does so, presumably, because he believes his choice is diagnostic of the demon's prediction. In other words, he believes that $p(P_1|C_1)$ is high. From Bayes's Theorem, we learn that $p(P_1|C_1)=\frac{p(P_1)p(C_1|P_1)}{p(P_1)p+(C_1|P_1)+p(P_2)p(C_1|P_2)}$, or $p\left(P_1|C_1\right) = \frac{.1\cdot.99}{.1\cdot.99+.9\cdot.01} = .917$. In other words, the demon's bias reduces the accuracy of the player's prediction, but only marginally.[10] The problem of inverse conditional probabilities does not even arise in the PD because the two players are interchangeable. The conditional probability of Player O's choice given Player P's choice is the same as its inverse. The labels are arbitrary.

[10] In fact, the demon's bias increases the two-boxer's accuracy from .99 to .999.

Challenges From Repair Models

Seeing that the evidential model passes the coherence test, whereas the classic approach does not, is not enough. Confidence in a model requires responses to challenges from other approaches as well. We now consider critical arguments arising from reciprocity (trust), social-value orientation, team reasoning, and the learning hypothesis (Simpson's paradox).

Reciprocity and Trust

The expected-reciprocity hypothesis assumes that people cooperate inasmuch as they think the probability of others' cooperating to be high. Willingness to cooperate when there remains a risk of being suckered may be seen as an expression of trust. Recently, the concept of "trust" has gained currency in psychology and economics (Evans & Krueger, 2009, 2011; Krueger et al., 2008; Luhmann, 2000; J. Simpson, 2007). Trust facilitates social exchanges that would otherwise remain unrealized given that people often lack complete information about the preferences of others and given that many exchanges cannot be regulated by laws or contracts. Kenneth Arrow (1974) called trust "a lubricant for social systems" (p. 23). What is it—if anything—that the evidential model can explain that trust-as-expected-reciprocity cannot?

First, appeals to trust can be circular. In the standard trust game (Berg et al., 1995), a player's investment is both the action to be explained and the measure of trust. This problem is partly addressed by evidence showing that individual differences in the propensity to trust predict behavior in the game (Bicchieri et al., 2004; Evans & Revelle, 2008). Second, trust-based explanations require additional mediator variables. Yamagishi and Yamagishi (1994) proposed that people trust inasmuch as they expect others to reciprocate. In other words, trust requires a high expectation regarding p_c. The question is how individuals generate this expectation. If they expect p_c to be high for reasons other than projection, they trust because of expected reciprocity. As we have seen, the expected-reciprocity hypothesis cannot fully account for the current findings. Specifically, the reciprocity hypothesis fails to predict that people want to unilaterally switch to defection in the PD and to accept any unilateral switch in the game of chicken.

Brewer (2008) proposed the concept of "depersonalized trust" (see also Yamagishi & Kiyonari, 2000). Her model lacks specificity because it admits several mediators of trust (anticipated reciprocity, social projection, and social identity). Given the specificity of the social-projection hypothesis and its fit with the data, we conclude that projection is a more effective lubricant of social exchange than trust is. Indeed, social projection appears to serve as a powerful antecedent to both, trust and cooperation.

Why Not Benevolence?

We noted that one derivate of the social-value-orientation framework, individual differences in benevolence, can account for the differences in cooperation between nice and nasty games. The social-value orientation approach also has the advantage of being theoretically plausible and of providing a straightforward measurement model. Again, however, unilateral switching to defection in both the PD and the game of chicken is a troubling result for this theory.

A model gains credence if it can account for behavior in diverse contexts. Colman (2003) pointed out that the social-value approach cannot solve simple coordination games. Suppose two players receive \$10 each if they both choose heads and receive \$5 each if they both choose tails. If their choices mismatch, neither gets anything. It is clear that no weight placed on the other person's gain (*benevolence*) will turn the choice of heads (or tails, for that matter) into a dominating strategy. Yet, naive players are not troubled by this game. They choose heads, cheerfully and correctly assuming that others will too.

A Farewell to Methodological Individualism?

Colman, Pulford, and Rose (2008b) believed that coordination games trip up not only the social-value hypothesis but also the social-projection hypothesis. They claim that

> players have no rational justification for assuming that others act as they do [and that evidential decision-theory] leads to absurdities because both players have free will and make their decisions independently [and therefore] a player's own decision cannot affect the probability that the other player will choose [the same strategy]. (p. 410)

Our response is clear. The theory of evidential decision making does not accord one player causal power over the choice of another. If it did, how could both be able to influence each other? As we have noted, we assume a common-cause model. The players' choices are similar because they are determined by the same antecedents (e.g., the "difficulty" of the game). The idea that "free will" deflates the evidential hypothesis is a red herring.

Colman et al. (2008b) suggested that there are games that only team reasoning can solve. In the "Ball Gown Game" (BGG), each of two women owns a blue and a red dress. Both prefer it if Lady B appears in her blue gown while Lady R appears in her red gown. Another acceptable, but less welcome, outcome is that

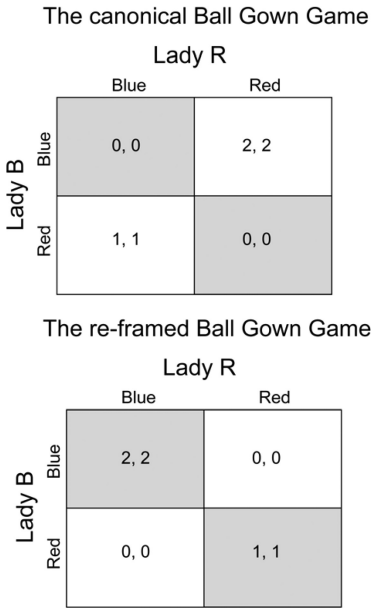

Figure 7. The Ball Gown Game in its canonical and modified form.

Lady B wears red while Lady R wears blue. Neither wants to be seen wearing the same gown as the other. The preference ranking is displayed in the top panel of Figure 7. Colman et al. (2008b) suggested that

> intuition and team reasoning predict the [optimizing, namely, the B, R] outcome . . . unambiguously, [whereas] evidential decision theory suggests that either strategy [i.e., choosing the blue or the red gown] is as bad as the other and is foredoomed to failure in any case because, as Krueger claims, "most people have a strong expectation that members of their own groups will act as they themselves do." (p. 410).

Notice two things. First, the BGG offers no support for team reasoning. The two women can solve their problem simply by wearing their favorite gowns. The collectively optimal outcome is simply a by-product of two individual self-regarding decisions. Second, a reframing of the BGG shows that evidential reasoning works well. Each lady needs to know only her own and the other's preference, and then ask herself what the other will do if she herself dons her favorite gown. With social projection, she will conclude that the other lady will also don her favorite gown, and the ball will be a success. The preferences displayed in the bottom panel of Figure 7 represent this reframing.

Again, for good measure, it must be noted that team reasoning did not predict the empirical results of the three studies.

Rational Judgment Is Future Oriented

We have argued that the empirical results uniquely support the social-projection hypothesis. But what about Simpson's paradox? Here, the idea is that people have observed a correlation between their own choices and payoffs over time, and that they overgeneralize this correlation to a particular game in the present. To use Reichenbach's phrase, the correlation that players have learned from past experience is "screened off" within individual games. Hence, letting one's choices be governed by that correlation is considered irrational (Chater et al., 2008). In contrast, the social-projection hypothesis is derived from the analytical truth that even within games, a positive correlation exists. For this correlation to appear, each player must consider both available choice strategies.

The correlation comprised by Simpson's paradox is rooted in past choices and payoffs. In the PD and other games, this correlation is never causal. So even if the correlation is not spurious, it must be ignored. In contrast, evidential decision making is future oriented. No information from the past enters the judgment and decision process. Instead, players' reasoning is strictly future oriented, as Dawes (1988; see also Hastie & Dawes, 2010) argued it should be (see Krueger, 2000, for an empirical examination). Perhaps the most common cause of irrational choice is mistaken respect for past behaviors, outcomes, and events (as, e.g., in outcome bias or the sunk cost fallacy). By this criterion, evidential decision making passes the bar of rationality.

The Pragmatic Value of Evidential Decision-Making: Improving Collective Outcomes Without Even Trying

A strong theory can account for phenomena beyond the particular context in which the theory's predictions were tested. In the case of the theory of evidential decision making (and the specific social-projection hypothesis), the broader claim is that the theory accounts for the general problem of collective action. The prototype of the collective-action problem is voting. Why do (some) individuals vote when they should realize that in a large election, an individual vote is wasted (Aldrich, 1993; Meehl, 1977). When voting is cast as a cooperative act, and when it is assumed (and shown) that people project their own inclination to vote (or abstain) more strongly to supporters of their own party than to supporters of the opposing party, it can be shown that the expected value of voting exceeds the expected value of abstaining (Acevedo & Krueger, 2004; Krueger & Acevedo, 2008).

Another puzzle is why so many social and economic exchanges take place when certain psychological biases appear to inhibit such activity. Consider a person

hoping to sell a good. A prospective seller can use a variety of cues to estimate the good's market value. One of these cues is the number of other sellers. If information is incomplete or uncertain, an individual willing to sell can projectively infer that many others are also willing to sell. As the attainable price is tied to the product's scarcity, and as projection works against perceptions of scarcity, this seller will likely set a lower price. Hence, projection facilitates trade. Likewise, a prospective buyer may projectively believe that the product is in great demand. This inference should increase his willingness to pay a higher price. Our model suggests that projection increases willingness to buy and sell, respectively, when buying and selling are being contemplated. Hence, projection facilitates action before choice instead of being a postchoice rationalization.

In combination, these two projective tendencies increase the probability of mutually beneficial social exchange.[11] They compensate for "empathy gaps" (van Boven, Dunning, & Loewenstein, 2000), which work against the execution of trades. Empathy gaps refer to the tendency of people to underestimate the endowment effect (Kahneman, Knetsch, & Thaler, 1991), that is, the finding that people who currently own a good value it more than people who do not currently own it. If the existence of the endowment effect, combined with a lack of understanding of this effect, leads sellers to ask more than buyers are willing to pay, no trade will take place. Social projection curtails the inhibitory effect of this bias.

A similar logic applies to mate selection. A person hoping to attract others with his or her desirable features will projectively assume that these features are in good supply in the market. Hence, to attract a mate, this person will make fewer demands on the other. Likewise, a person wishing to approach an attractive other will projectively assume that there are other suitors. Hence, this person will try harder, for example, by improving his or her own suite of features. Again, projection lowers the barriers of exchange. Without it, more people would end up alone than is currently the case.

Conclusion

The theory of evidential decision making has been controversial, and the debate over its merits has been mostly carried out by philosophers with an interest in Newcomb's problem (cf. R. Campbell & Sowden, 1985). We have reviewed some of their arguments, and we introduced new ones from a psychological point of view. Our goal is to revitalize interest in this theory by showing how it can benefit from research on social projection. Most important, we have derived and found support for a unique set of empirical predictions. The evidence shows that evidential decision making accounts for people's strategic behavior and their perceptions of other agents better than any of several competing theories. Specifically, the theory can explain differences in cooperation between nice and nasty games, it can predict under what conditions agents will switch their decisions if given the opportunity, and it can contribute to our understanding of some social behaviors such as voting, trading, and mating. Almost as an aside, the theory of evidential decision making overcomes the nagging conflicts between the presumed rationality and morality of research subjects and between a theory's own normative and descriptive aims.

Yet the theory does not provide a panacea for the reconciliation of individual and collective interests. Its major characteristic (and perhaps limitation) is that its predictions depend on the individual agents being ignorant with regard to what others will do. Only under conditions of social ignorance can their own personal choices lead to Bayesian inductive beliefs that favor cooperation over defection. Once people gather behavioral information from others, this information crowds out the single observation provided by sampling their own behavior. The inevitable consequence is that estimates of p_c stabilize; they no longer vary with the person's own assumed choices. As defection is increasingly recognized as the dominating strategy, rates of cooperation drop off. This drop-off is a robust empirical result and the theory of evidential decision making is consistent with it.

A note of hope comes from the finding that rates of cooperation rebound when agents are reconvened in new groups (Andreoni, 1988; Croson, 1996). The theory can explain this result. Categorization of humans into novel groups restores social ignorance and thus the benefits brought by social projection. As individual group members expect one another to be similar to themselves (Krueger & DiDonato, 2008), they have a sense of shared identity, which enables mutual cooperation—until, that is, they get to know one another better.

Another and somewhat paradoxical note of hope comes from findings suggesting that some people some of the time project too much. Projection is egocentric when people weight evidence from their own behavior more strongly than evidence from others' behavior (Alicke & Largo, 1995; Krueger & Clement, 1994). A person who egocentrically overestimates the diagnosticity of her own behavior may still cooperate when a more balanced person would choose to defect. With strong egocentrism, social projection becomes aligned with irrational, although it still supports the collective welfare. To rephrase that which seems ironic: Our theory predicts that the most self-involved individuals can end up as pillars of society.

[11] We thank Thorsten Meiser for this suggestion.

Acknowledgments

Part of this work was completed when the first author was visiting at the University of Marburg, Germany, as an Alexander-von-Humboldt Research Prize winner. We thank Johannes Ullrich and the Social Psychology team at the University of Frankfurt for stimulating discussions and Anthony Evans, Steven Guglielmo, Andrew Monroe, and Jan Rummel for comments on a draft of this manuscript. Christina Wehrli helped with data collection, and David Rand provided valuable references.

Note

Address correspondence to Joachim I. Krueger, Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Box 1853, 190 Thayer Street, Providence, RI 02912. E-mail: Joachim_Krueger@Brown.edu

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. doi:10.1111/j.1467-9221.2004.00359.x

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

Aldrich, J. H. (1993). Rational choice and turnout. *American Journal of Political Science, 37*, 246–278. doi:10.2307/2111531

Alicke, M. D., & Largo, E. (1995). The role of self in the false consensus effect. *Journal of Experimental Social Psychology, 31*, 28–47. doi:10.1006/jesp.1995.1002l

Ames, D. R. (2004). Inside the mind reader's tool kit: Projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology, 87*, 340–353. doi: 10.1037/0022-3514.87.3.340

Ames, D. R. Weber, E. U., & Zou, X. (2012). Mind-reading in strategic interaction: The impact of perceived similarity on projection and stereotyping. *Organizational Behavior and Human Decision Processes, 117*, 96–110. doi:10.1016/j.obhdp.2011.07.007

Anand, P. (1989). Two types of utility: An experimental investigation into the prevalence of causal and evidential utility maximisation. *Greek Economic Review, 12*, 58–74.

Anderson, K. G. (2006). How well does paternity confidence match actual paternity: Evidence from worldwide paternity rates. *Current Anthropology, 47*, 513–520. doi:10.1086/504167

Andreoni, J.,(1988). Why free ride? *Journal of Public Economics, 37*, 291–304. doi:10.1016/0047-2727(88)90043-6

Andreoni, J. (1995). Cooperation in public-goods experiments: Kindness or confusion? *American Economic Review, 85*, 891–904.

Arrow, K. (1974). *The limits of organization*. New York, NY: Norton.

Axelrod, R. (1980). Effective choice in the prisoner's dilemma. *Journal of Conflict Resolution, 24*, 3–25.

Ayer, A. J. (1956). *The problem of knowledge*. Baltimore, MD: Penguin.

Bacharach, M. (1999). Interactive team reasoning: A contribution to the theory of cooperation. *Research in Economics, 53*, 117–147. doi:10.1006/reec.1999.0188

Bar-Hillel, M., & Margalit, A. (1972). Newcomb's paradox revisited. *British Journal of the Philosophy of Science, 23*, 295–304. doi:10.1093/bjps/23.4.295

Baumeister, R. F. (2008). Free will in scientific psychology. *Perspective on Psychological Science, 3*, 14–19. doi:10.1111/j.1745-6916.2008.00057.x

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social-history. *Games and Economic Behavior, 10*, 122–142. doi:10.1006/game.1995.1027

Bicchieri, C., Duffy, J., & Tolle, G. (2004). Trust among strangers. *Philosophy of Science, 71*, 286–319. doi:10.1086/381411

Binmore, K. (1999). Why experiments in economics? *Economic Journal, 109*, 16–24. doi:10.1111/1468-0297.00399

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and cooperation. *American Economic Review, 90*, 166–193. doi:10.1257/aer.90.1.166

Brams, S. J. (1975). Newcomb's problem and prisoner's dilemma. *Journal of Conflict Resolution, 19*, 596–612. doi:10.1177/002200277501900402

Brewer, M. B. (2008). Depersonalized trust and ingroup cooperation. In J. I. Krueger (ed.), *Rationality and social responsibility: Essays in honor of Robyn Mason Dawes* (pp. 215–232). New York, NY: Taylor & Francis.

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. doi:10.1086/344546

Burnham, T. C., & Johnson, D. D. P. (2005). The biological and evolutionary logic of human cooperation. *Analyse & Kritik, 27*, 113–135.

Burnstein, E., Crandall, C., & Kitayama, S. (1994). Some Neo-Darwinian decision rules for altruism: Weighing cues for inclusive fitness as a function of the biological importance of the decision. *Journal of Personality and Social Psychology, 67*, 773–789. doi:10.1037/0022-3514.67.5.773

Camerer, C. F. (2003). *Behavioral game theory*. Princeton, NJ: Princeton University Press.

Campbell, D. T. (1975). On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist, 30*, 1103–1126. doi: 10.1037/0003-066X.30.12.1103

Campbell, R., & Sowden, L. (1985). *Paradoxes of rationality and cooperation: Prisoner's dilemma and Newcomb'sproblem*. Vancouver, CA: University of British Columbia 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

Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*, 591–621. doi:10.1146/annurev.psych.55.090902.142015

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

Colman, A. M., Pulford, B. D., & Rose, J. (2008a). Collective rationality in interactive decisions: Evidence for team reasoning. *Acta Psychologica, 128*, 387–397. doi:10.1016/j.actpsy.2007.08.003

Colman, A. M., Pulford, B. D., & Rose, J. (2008b). Team reasoning and collective rationality: Piercing the veil of obviousness. *Acta Psychologica, 128*, 409–412. doi:10.1016/j.actpsy.2008.04.001

Cooper, R., DeJong, D. V., Forsythe, R., Ross, T. W. (1996). Cooperation without reputation: Experimental evidence from prisoner's dilemma games. *Games and Economic Behavior, 12*, 187–218. doi:10.1006/game.1996.0013

Croson, R. T. A. (1996). Partners and strangers revisited. *Economic Letters, 53*, 25–32. doi:10.1016/S0165-1765(97)82136-2

Cuddy, A., Fiske, S., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS Map. *Advances in Experimental Social Psychology, 40*, 61–149. doi:10.1016/S0065-2601(07)00002-0

Dal Bó, P. (2005). Cooperation under the shadow of the future: Experimental evidence from infinitely repeated games. *American Economic Review, 95*, 1591–1604. doi:10.1257/000282805775014434

Danielson, P. A. (1998). *Modeling rationality, morality, and evolution*. New York, NY: Oxford University Press.

Davis, L. H. (1977). Prisoners, paradox, and rationality. *American Philosophical Quarterly, 14*, 319–327.

Dawes R. M. (1980). Social dilemmas. *Annual Review of Psychology, 31*, 169–193. doi:10.1146/annurev.ps.31.020180.001125

Dawes, R. M. (1988). *Rational choice in an uncertain world*. San Diego, CA: Harcourt Brace Javanovich.

Dawes, R. M. (1989). Statistical criteria for establishing a truly false consensus effect. *Journal of Experimental Social Psychology, 25*, 1–17. doi:10.1016/0022-1031(89)90036-X

Dawes, R. M. (1990). The potential nonfalsity of the false consensus effect. In R. M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 179–199). Chicago, IL: University of Chicago Press.

Dawes, R. M. (1998). Behavioral decision making and judgment. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1, pp. 497–548). New York, NY: McGraw-Hill.

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. doi:10.1037/0022-3514.35.1.1

Dawes, R. M., & Messick, D. M. (2000). Social dilemmas. *International Journal of Psychology, 35*, 111–116. doi:10.1080/002075900399402

Dawes, R. M., Mirels, H. L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science, 4*, 396–400. doi:10.1111/j.1467-9280.1993.tb00588.x

Dawes, R. M., van de Kragt, A. J. C., & Orbell, J. M. (1988). Not me or thee but we: the importance of group identity in eliciting cooperation in dilemma situations: experimental manipulations. *Acta Psychologica, 68*, 83–98. doi:10.1016/0001-6918(88)90047-9

DiDonato, T. E., Ullrich, J., & Krueger, J. I. (2011). Social perception as induction and inference: An integrative model of intergroup differentiation, ingroup favoritism, and differential accuracy. *Journal of Personality and Social Psychology, 100*, 66–83. doi:10.1037/a0021051

Diekmann, A. (1985). Volunteer's dilemma. *Journal of Conflict Resolution, 29*, 605–610. doi:10.1177/0022002785029004003

Durkheim, E. (1982). *Règles de la méthode sociologique* [The rules of sociological method]. (S. Lukes, Ed.; W. D. Halls, Trans.). New York, NY: Free Press. (Original work published 1895)

Epley, N., Keysar, B., & van Boven. L. (2004). Perspective taking as egocentric anchoring and adjustment. *Journalof Personality and Social Psychology, 87*, 327–339. doi:10.1037/0022-3514.87.3.327

Evans, A. M., & Krueger, J. I. (2009). The psychology (and economics) of trust. *Social and Personality Compass: Intrapersonal Processes, 3*, 1003–1017. doi:10.1111/j.1751-9004.2009.00232.x

Evans, A. M., & Krueger, J. I. (2011). Elements of trust: Risk taking and expectation of reciprocity. *Journal of Experimental Social Psychology, 47*, 171–177. doi:10.1016/j.jesp.2010.08.007

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality, 42*, 1585–1593. doi:10.1016/j.jrp.2008.07.011

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

Fischer, I. (2009). Friend or foe: Subjective expected relative similarity as a determinant of cooperation. *Journal of Experimental Psychology: General, 138*, 341–350. doi:10.1037/a0016073

Fischer, P., Krueger, J. I., Greitemeyer, T., Vogrinic, C., Kastenmüler, A., Frey, D., . . . Kainbacher, M. (2011). The bystander effect: A meta-analytic review on bystander intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin, 137*, 517–537. doi:10.1037/a0023304

Flood, M., & Drescher, M. (1952). *Some experimental games. Research memorandum RM–789*. Santa Monica, CA: Rand Corporation.

Foddy, M., Platow, M. J., & Yamagishi, T. (2009). Group-based trust in strangers: The role of stereotypes and expectations. *Psychological Science, 20*, 419–422. doi:10.1111/j.1467-9280.2009.02312.x

Frank, R. H., Gilovich, T., & Regan, D. T. (1993). Does studying economics inhibit cooperation? *Journal of Economic Perspectives, 7*, 159–171.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgment. *Psychological Bulletin, 101*, 75–90. doi:10.1037/0033-2909.101.1.75

Galton, F. (1907). Vox populi. *Nature, 75*, 509–510. doi:10.1038/075509e0

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*, 21–38. doi:10.1037/0033-2909.117.1.21

Gintis, H. (2009). *The bounds of reason: Game theory and the unification of the behavioral sciences*. Princeton, NJ: Princeton University Press.

Gneezy, U., & List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica, 74*, 1365–1384. doi:10.1111/j.1468-0262.2006.00707.x

Goeree, J. K., & Holt, C. A. (1999). Stochastic game theory: For playing games, not just for doing theory. *Proceedings of the National Academy of Sciences, 96*, 10564–10567. doi:10.1073/pnas.96.19.10564

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review, 25*, 161–178. doi:10.2307/2092623

Grafstein, R. (1991). An evidential decision theory of turnout. *American Journal of Political Science, 35*, 989–1010. doi:10.2307/2111503

Grafstein, R. (2002). What rational political actors can expect. *Journal of Theoretical Politics, 14*, 139–165. doi:10.1177/095169280201400201

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 243–248.

Hardin, G. (1977). *The limits of altruism: An ecologist's view of survival*. Bloomington, IN: University of Indiana Press.

Hastie, R., & Dawes, R. M. (2010). *Rational choice in an uncertain world* (2nd ed). Thousand Oaks, CA: Sage.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20*, 231–237. doi:10.1111/j.1467-9280.2009.02271.x

Hoch, S. J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53*, 221–234. doi:10.1037/0022-3514.53.2.221

Howard, J. V. (1988). Cooperation in the prisoner's dilemma. *Theory and Decision, 24*, 203–214. doi:10.1007/BF00148954.

Jeffrey, R. (1983). *The logic of decision* (2nd ed.). Chicago, IL: University of Chicago Press.

Jones, B. A., & Rachlin, H. (2006). Social discounting. *Psychological Science, 17*, 283–286. doi:10.1111/j.1467-9280.2006.01699.x

Jones, B. A., & Rachlin, H. (2009). Delay, probability and social discounting in a public goods game. *Journal of the Experimental Analysis of Behavior, 91*, 61–93. doi:10.1901/jeab.2009.91-61

Jones, B., Steele, M., Gahagan, J., & Tedeschi, J. (1968). Matrix values and cooperative behavior in the prisoner's dilemma game. *Journal of Personality and Social Psychology, 8*, 148–153. doi:10.1037/h0025299

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *Journal of Economic Perspectives, 5*, 193–206.

Kant, I. (1998). *Grundlegung zur Metaphysik der Sitten* [The Groundwork of the Metaphysics of Morals] (M. Gregor, Trans.). New York, NY: Cambridge University Press. (Original work published 1785)

Kelley, H. H., & Thibaut, J. W. (1978). *Interpersonal relations: A theory of interdependence*. New York, NY: Wiley.

Kerr, N. L. (1995). Norms in social dilemmas. In D. Schroeder (Ed.), *Social dilemmas: Perspectives on individuals and groups* (pp. 31–48). Westport, CT: Praeger.

Kollock, P. (1998). Transforming social dilemmas: Group identity and co-operation. In P. A. Danielson (Ed.), *Modeling rationality, morality, and evolution* (pp. 185–209). New York, NY: Oxford University Press.

Krebs, D. L. (2008). Morality: An evolutionary account. *Perspectives on Psychological Science, 3*, 149–172. doi:10.1111/j.1745-6924.2008.00072.x

Krueger, J. (1996). Probabilistic national stereotypes. *European Journal of Social Psychology, 26*, 961–980. doi:10.1002/(SICI)1099-0992(199611)26:6<961::AID-EJSP799>3.0.CO;2-F

Krueger, J. (1998). On the perception of social consensus. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 30, pp. 163–240). San Diego, CA: Academic Press. doi:10.1016/S0065-2601(08)60384-6

Krueger, J. (2000). Distributive judgments under uncertainty: Paccioli's game revisited. *Journal of Experimental Psychology: General, 129*, 546–558. doi:10.1037/0096-3445.129.4.546

Krueger, J. I. (2007). From social projection to social behavior. *European Review of Social Psychology, 18*, 1–35. doi:10.1080/10463280701284645

Krueger, J. I. (2008). Methodological individualism in experimental games: Not so easily dismissed. *Acta Psychologica, 128*, 398–401. doi:10.1016/j.actpsy.2007.12.011

Krueger, J. I. (2009). A componential model of situation effects, person effects and situation-by-person interaction effects on social behavior. *Journal of Research in Personality, 43*, 127–136. doi:10.1016/j.jrp.2008.12.042

Krueger, J. I. (2012). Rationality: Variations on a theme. Italy: In-Mind.

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, NY: Psychology Press.

Krueger, J. I., & Acevedo, M. (2007). Perceptions of self and other in the prisoner's dilemma: Outcome bias and evidential reasoning. *American Journal of Psychology, 120*, 593–618.

Krueger, J. I., & Acevedo, M. (2008). A game-theoretic view of voting. *Journal of Social Issues, 64*, 467–485. doi:10.1111/j.1540-4560.2008.00573.x

Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67*, 596–610. doi:10.1037/0022-3514.67.4.596

Krueger, J. I., & DiDonato, T. E. (2008). Social categorization and the perception of groups and group differences. *Social and Personality Psychology Compass: Group Processes, 2*, 733–750. doi:10.1111/j.1751-9004.2008.00083.x

Krueger, J. I., & DiDonato, T. E. (2010). Person perception in (non)interdependent games. *Acta Psychologica, 134*, 85–93. doi:10.1016/j.actpsy.2009.12.010

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. doi:10.1017/S0140525×04000081

Krueger, J. I., & Massey, A. L. (2009). A rational reconstruction of misbehavior. *Social Cognition, 27*, 785–810. doi:10.1521/soco.2009.27.5.786

Krueger, J. I., Massey, A. L., & DiDonato, T. E. (2008). A matter of trust: From social preferences to the strategic adherence of social norms. *Negotiation & Conflict Management Research, 1*, 31–52. doi:10.1111/j.1750-4716.2007.00003.x

Krueger, J., & Zeiger, J. S. (1993). Social categorization and the truly false consensus effect. *Journal of Personality and Social Psychology, 65*, 670–680. doi:10.1037/0022-3514.65.4.670

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Laplace, M. (1953). *Essay on probability*. Mineola, NY: Dover. (Original work published 1783)

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Social judgment and decision making* (pp. 225–240). New York, NY: Psychology Press.

Levi, I. (1975). Newcomb's many problems. *Theory and Decision, 6*, 161–175. doi:10.1007/BF00169104

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives, 21*, 153–174. doi:10.1257/jep.21.2.153

Lewis, D. K. (1979). Prisoner's dilemma is a Newcomb problem. *Philosophy & Public Affairs, 8*, 235–240.

Lewis, D. K. (1981). Causal decision theory. *Australasian Journal of Philosophy, 59*, 5–30. doi:10.1080/00048408112340011

Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York, NY: Wiley & Sons.

Luhmann, N. (2000). Familiarity, confidence, trust: Problems and alternatives. In D. Gambietta (Ed.), *Trust: Making and breaking cooperative relations*, (pp. 94–107). Oxford, UK: University of Oxford Press.

Maner, J. K., Luce, C. L., Neuberg, S. L., Cialdini, R. B., Brown, S., & Sagarin, B. J. (2002). The effects of perspective taking on motivations for helping: Still no evidence for altruism. *Personality and Social Psychology Bulletin, 28*, 1601–1610. doi:10.1146/annurev.ps.45.020194.001213

Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature, 246*, 15–18. doi:10.1038/246015a0

McClure, J. (1998). Discounting causes of behavior: Are two reasons better than one? J*ournal of Personality and Social Psychology, 74*, 7–20. doi:10.1037/0022–3514.74.1.7

Meehl, P. E. (1977). The selfish voter paradox and the thrown-away vote argument. *The American Political Science Review, 71*, 11–30. doi:10.2307/1956951

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. doi:10.1037/0022–3514.37.4.602

Mill, J. S., & Bentham, J. (1987). *Utilitarianism and other essays* (A. Ryan, Ed.). New York, NY: Penguin Books.

Miller, D. T. (1999). The norm of self-interest. *American Psychologist, 54*, 1053–1060. doi:10.1037/0003-066X.54.12.1053

Murnighan, J. K., & Roth, A. E. (1983). Expecting continued play in prisoner's dilemma games: A test of several methods. *Journal of Conflict Resolution, 27*, 279–300. doi:10.1177/0022002783027002004

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.

Pillutla, M. M., Malhotra, D., Murnighan, J. K. (2003). Attributions of trust and the calculus of reciprocity. *Journal of Experimental Social Psychology, 39*, 448–455. doi:10.1016/S0022-1031(03)00015-5

Pinker, S. (2008, January 13). The moral instinct. *New York Times Magazine*, p. 32.

Poincaré, H. (1996). *Science and method*. Bristol, UK: Thoemmes Press. (Original work published 1914)

Popper, K. (1963). *Conjectures and refutations*. New York, NY: Routledge.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of 'rational' decision theory. *Proceedings of Biological Sciences, 276*, 2171–2178. doi:10.1098/rspb.2009.0121

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. doi:10.1037/0022-3514.46.2.237

Rapoport, A. (1966). *Two-person game theory*. Ann Arbor, MI: University of Michigan Press.

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

Rapoport, A. (1973). *Two-person game theory: The essential ideas*. Ann Arbor, MI: University of Michigan Press.

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

Rapoport, A., & Chammah, M. (1965). *Prisoner's dilemma: A study in conflict and cooperation*. Ann Arbor, MI: Ann Arbor Press.

Rapoport, A., & Chammah, M. (1966). The game of chicken. *American Behavioral Scientist, 10*, 10–28. doi: 10.1177/000276426601000303

Reeder, G. D. (2008). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological Inquiry, 20*, 1–18. doi:10.1080/10478400802615744

Reichenbach, H. (1956). *The direction of time*. Berkeley, CA: University of California Press.

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. doi: 10.1207/s15327957pspr0901_3

Ross, L., Greene, D., & House, P. (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology, 13*, 279–301. doi:10.1016/0022-1031(77)90049-X

Roughgarden, J., Oishi, M., & Akçay, E. (2006). Reproductive social behavior: Cooperative games to replace sexual selection. *Science, 311*, 965–969. doi:10.1126/science.1110105

Rousseau, J.-J. (1755/1992). *Discourse on the origins of inequality*. Indianapolis, IN: Hackett.

Russell, B. W. (1913). On the notion of cause. *Proceedings of the Aristotelian Society, 13*, 1–26.

Russell, B. W. (1930). *The conquest of happiness*. London, UK: Unwin.

Russell, B. W. (1959). *Common sense and nuclear warfare*. London, UK: Unwin.

Sally, D. (1995). Conversation and cooperation in social dilemmas. *Rationality and Society, 7*, 58–92. doi:10.1177/1043463195007001004

Savage, L. J. (1954). *The foundations of statistics*. New York, NY: Wiley and Sons.

Schmid, H. B. (2003). Rationality in relations. *American Journal of Economics and Sociology, 62*, 67–101. doi: 10.1111/1536-7150.t01-1-00003

Segre, S. (2008). Durkheim on rationality. *Journal of Classical Sociology, 8*, 109–144. doi:10.1177/1468795×07084697

Shafir, E., & Tversky, A. (2004). Thinking through uncertainty: Nonconsequential reasoning and choice. In E. Shafir (Ed.), *Preferences, belief, and similarity* (pp. 703–727). Cambridge, MA: MIT Press.

Simpson, E. H. (1951). The interpretation of interaction and contingency tables. *Journal of the Royal Statistical Society, 13*, 238–241.

Simpson, J. (2007). Psychological foundations of trust. *Current Directions in Psychological Science, 16*, 264–268. doi:10.1111/j.1467-8721.2007.00517.x

Singer, P. (1993). *Practical ethics*. New York, NY: Cambridge University Press.

Smith, V. (2003). Constructivist and ecological rationality in economics. *American Economic Review, 93*, 465–508. doi:10.1257/000282803322156954

Steele, M. W., & Tedeschi, J. T. (1967). Matrix indices and strategy choices in mixed-motive games. *Journal of Conflict Resolution, 11*, 198–205. doi:10.1177/002200276701100207

Sugden, R. (2000). Team preferences. *Economics and Philosophy, 16*, 175–204. doi:10.1017/S0266267100000213

Tanis, M., & Postmes, T. (2005). A social identity approach to trust: Interpersonal perception,group membership and trusting behavior. *European Journal of Social Psychology, 35*,413–424.doi:10.1002/ejsp.256

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review* (Monograph Supplement), *2*(8).

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

van Boven, L., Dunning, S., & Loewenstein, G. (2000). Egocentric empathy gaps between owners and buyers: Misperceptions of the endowment effect. *Journal of Personality and Social Psychology, 79*, 66–76. doi:10.1037/0022-3514.79.1.66

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. doi:10.1037/0022-3514.77.2.337

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will. *Psychological Science, 19*, 49–54. doi:10.1111/j.1467-9280.2008.02045.x

Von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*, 645–647. doi:10.1111/j.1467-9280.2008.02136.x

White, M. D. (2006). Multiple utilities and weakness of will: A Kantian perspective. *Review of Social Economy, 64*, 1–20. doi:10.1080/00346760500529914

Yamagishi, T., & Kiyonari, T. (2000). The group as the container of generalized reciprocity. *Social Psychology Quarterly, 63*, 116–132. doi:10.2307/2695887

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the United States and Japan. *Motivation and Emotion, 18*, 129–166. doi:10.1007/BF02249397