2

Social Projection and the Psychology of Choice

JOACHIM I. KRUEGER
Brown University
MELISSA ACEVEDO
Valencia Community College

We certainly use our knowledge of ourselves in order to frame hypotheses about some other people, or about all people.
—Karl R. Popper, 1957, *The Poverty of Historicism*, p. 138

We humans enjoy some awareness of our temporary states and enduring properties. We feel that we are in a particular mood, we know that we have certain preferences and traits, and we intend to behave in certain ways. Outside observers often validate such introspective knowledge, but sometimes there are discrepancies. These discrepancies can be particularly distressing if they involve the results of professional psychological assessment. After in-depth interviews, for example, clinical psychologists may conclude that a client is depressed, although the client denies being in this state (Shedler, Mayman, & Manis, 1993). Likewise, social psychologists may claim on the basis of an implicit attitude test that a self-described liberal research participant is prejudiced against a certain group (Greenwald, McGhee, & Schwartz, 1998).

When professional assessment and subjective experience diverge, many psychologists assume that self-reports are in error (Wilson & Dunn, 2004). One prevalent perspective is that preconscious neural activity is a sufficient cause of behavior (Bargh & Ferguson, 2000), and that self-perception arises only as a set of fallible inferences and constructions (Wohlschläger, Haggard, Gesierich, & Prinz, 2003). With such doubts about the value of introspection, one important goal of psychological assessment is to do without it. For three reasons, however, self-reports have withstood attempts to eliminate them. One reason is that objective measures (e.g., human observers or sophisticated apparatus) often lack the desired reliability.

Another reason is the classic argument of privileged access, according to which there are many mental events for which there are no adequate objective measures.

Last but not least, self-reports remain attractive because of their economy. When mental events lie close to the surface of consciousness, interested researchers need only ask what they are. Self-reports can then be compared to represent individual differences. Attitudes regarding a certain proposition, for example, may range from strong support to stiff resistance. To the average person, such variation may not be evident. Indeed, the idea of privileged access implies that social knowledge is more fragile than knowledge about the self. Nonetheless, knowledge of others is vital for accurate self-perception (see Alicke & Govorun or Mussweiler, this volume) and effective social interaction. Lack of dependable social knowledge hampers efforts to understand one's place in the social world.

How do people get around the relative inaccessibility of social knowledge? This chapter suggests that social projection is a judgmental heuristic that leads people to expect that others will behave as they themselves do. The first part of this chapter is a review of how this heuristic operates when the self is seen as a fixed structure. Noting that social projection is a type of inductive reasoning, we show that expecting others to be similar to the self improves the accuracy of social predictions (see Van Boven & Loewenstein, this volume, for some of the risks involved in projective reasoning). We then extend this analysis to show that when the self changes, social predictions change too. The second part of this chapter offers an analysis of strategic behavior in social games. On the assumption that social projection enhances the accuracy of predictions, we suggest that projection serves a person's self-interest by facilitating adaptive behavior that also promotes the common good.

A BAYESIAN FRAMEWORK FOR SOCIAL PREDICTION

Self as Entity

Floyd Allport (1924) introduced the idea of social projection; other prominent social psychologists (e.g., Asch, 1952; Heider, 1958) as well as psychometricians (Cronbach, 1955) transformed and elaborated upon it. Allport theorized that by using information about the self to generate social predictions, people come to assume that others are much like them. As the introductory quote from Popper shows, even a philosophy of science concerned with the problem of inductive inference acknowledges the pivotal role of self-knowledge as a source of social hypotheses. If people have only one bit of readily accessible information, why should they not use it to make predictions about others? Sometimes, even a sample consisting of one observation can make a difference. A microbe discovered on Mars refutes the idea that only Earth bears life. At other times, such samples change current views very little. A photo showing nothing but rocks makes it *more likely* that Mars is barren, but does *not prove* it. One can make a sport of thinking up novel activities,

such as making an omelet without cracking an egg or completing the first nude ascent of Mt. Everest. Once executed, these activities refute the idea that they are impossible, and the question becomes how easily and how often they will be replicated. However strange an activity might seem, social projection will make it quite doubtful that one would be the first or the last to do it.

To Popper, the goal of empirical observation was to weed out poor hypotheses. People would learn the most if they found evidence that others are different instead of similar to them. According to the Bayesian approach to induction, however, outright falsification is rare. Instead, most observations gradually alter the credibility of certain hypotheses or beliefs (Howson & Urbach, 1989). When evidence becomes available, all hypotheses consistent with it become more probable, and all inconsistent hypotheses become less probable. Suppose there are two hypotheses regarding the prevalence of a certain attitude in a particular social group. According to one hypothesis, 70% of group members believe that, say, brown eyes are more attractive than blue eyes, but according to the other hypothesis only 30% hold that view. If there are no grounds to favor either of these hypotheses *a priori*, a state of indifference prevails, in which each hypothesis is equally likely to be true (i.e., p(H1) = p(H2) = .5).[1]

When a person is selected at random from a group, the probability that this person has the attitude in question is either $p(A|H_1) = .7$ or $p(A|H_2) = .3$. Because the two hypotheses are deemed equally likely to be true, the overall probability that the person has the attitude, $p(H_1)$, is .5. Now suppose that this random person actually has the attitude. Bayes's Theorem gives the revised probability of each hypothesis as the product of the probability of the attitude under that hypothesis and the ratio of prior probability of the hypothesis over the overall probability of the attitude, namely

$$p(H_i \mid A) = p(A \mid H_i) \cdot \frac{p(H_i)}{p_1(A)}.$$

Because in the present case the ratio is 1, $p(H_1|A) = .7$ and $p(H_2|A) = .3$. The probability that the next randomly selected person will hold the attitude, $p_2(A)$, can be computed by multiplying the revised probability of each hypothesis with the conditional probability of the attitude under that hypothesis and by summing the products. Thus,

$$p_2(A) = p(H_1 \mid A) \cdot p(A \mid H_1) + p(H_2 \mid A) \cdot p(A \mid H_2) = .58$$

and the difference between p_2 and p_1 captures the effect of past evidence on future expectations.[2] Here, observing one instance of the attitude increases its estimated prevalence by 8 percentage points.

The degree to which observations change beliefs depends on the hypotheses being considered and their respective prior probabilities. Consider a scenario in which the attitude is thought to be either extremely rare (i.e., $p(A|H_1) = .1$) or

extremely common (i.e., $p(A|H_2) = .9$). In another scenario, the attitude is assumed to be either moderately rare (i.e., $p(A|H_1) = .3$) or moderately common (i.e., $p(A|H_2) = .7$). In both scenarios, the initial probability of the attitude is $p_1(A) = .5$. Now suppose that the prior probability of the hypothesis that the attitude is rare, $p(H_1)$, ranges from .1 to .9 (with $p(H_2) = 1 - p(H_1)$). Figure 2.1 shows belief revision, $p_2(A) - p_1(A)$, across levels of $p(H_1)$. The steep line represents the scenario of $p(A|H_1) = .1$, and the shallow line represents the scenario of $p(A|H_1) = .3$. The difference in elevation between the two lines shows that the degree of belief change corresponds to the extremity of the available hypotheses. The prior probabilities of the hypotheses also matter. Beliefs change the most when the hypothesis which suggests that the attitude is rare has a high prior probability.[3]

Now let's return to the question of social projection. A person may wonder: "What does my having this attitude tell me about how others feel?" If there is no information about how others feel, the probability of finding the attitude in a random other person, $p(A|H_i)$, may be anywhere between 0 and 1. The person is in a state of indifference if there is no reason to assign different prior probabilities to these hypotheses. In this idealized state of affairs, the revised probability of the attitude can be shown to be $(k + 1)/(n + 2)$, where k is the number of positive instances (e.g., people with attitude A) and n is the total size of the sample (see Howson & Urbach, 1989, pp. 42–45). Following this logic, a person with the

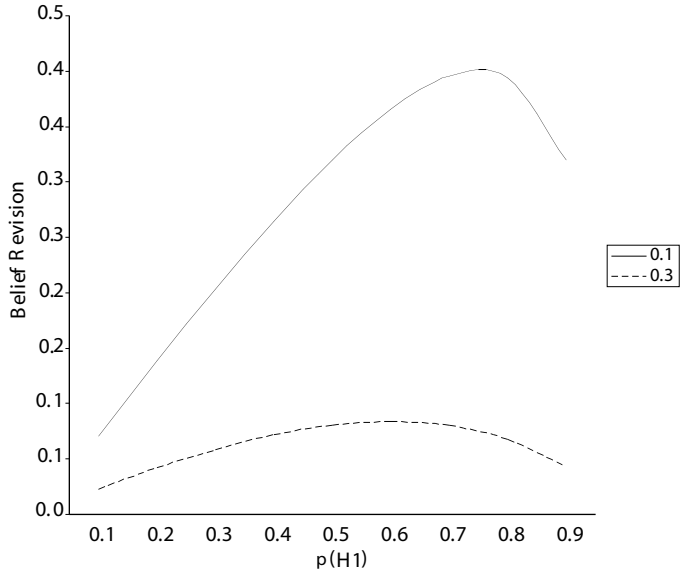


FIGURE 2.1

attitude will estimate its probability to be .67, whereas a person without it will estimate its probability to be .33.

When two people make different predictions, they need not be mistaken in their thinking. Both could have made the best estimate given the information they had (Dawes, 1989). Even scientists often draw different conclusions from different observations in the laboratory. When information is shared, however, predictions should converge. Only when people continue to depend primarily on what they know about themselves, can it be said that their social predictions are egocentric (Krueger, 2000). By the same token, scientific disagreements should diminish when data are integrated as they are, for example, in meta-analyses. But here too, a risk of egocentric prediction remains. Some researchers question whether disparate findings can be aggregated, claim special status of their own findings, or simply ignore the results of others (Tetlock, 2002).

The idea that people base social predictions on their own responses assumes that this sample information is random. But is it possible for people to regard their own responses as random samples of behavior? Compare the ordinary person's perspective with that of a research scientist. Researchers seek to sample randomly from a domain so that, on average, their observations are unbiased. Their statistical inferences depend on this assumption. Ordinary people cannot develop a comparable sense of how random their own responses are with regard to the group. Nothing in their phenomenal experience represents the idea of randomness. It does not seem to make sense to say that a single behavior or attitude is random. Indeed, statistical assessments of randomness require knowledge of the process by which a sample is drawn. If the process is free of bias (as in a drawing of the winning lottery ticket), a single observation may be considered random. But it is this insight into the sampling process that individuals do not have when they reflect on their own attitudes or behaviors.

To summarize, the statistics of belief revision can serve as a model for the psychological heuristic of social projection. According to this model, people hold prior beliefs about the social world, they consult their own attributes or behaviors, and they revise their social beliefs accordingly. The statistical properties of this model say nothing, however, about the underlying mental processes people use to access, weigh, and integrate their prior beliefs with new observations. The detection of such processes has been the task of experiments (Krueger, 1998). One interesting finding concerns the type of cause people identify to explain their own behavior. When people conclude that some aspect of the social situation controls their own behavior, they tend to believe that the effect on others will be similar. When, however, they see their behavior as a reflection of their personality, they project less (Gilovich, Jennings, & Jennings, 1983).

This difference resembles the pattern in Figure 2.1. The top line represents the prior belief that behavior will be relatively uniform. In a scripted social situation, people will do one thing or another, although one may not know which beforehand. Once a behavior is observed, it strongly affects further expectations.

The bottom line represents the belief that people are variable (as, by definition, in their personality differences), and thus new information has little effect. This example illustrates how the statistical modeling of social projection can be reconciled with experiments searching for its psychological sources. At minimum, experimental findings yield estimates of the prior assumptions from which people generated their predictions.

WHEN THE SELF OR THE GROUP CHANGES

Many researchers working in social cognition regard the individual self-concept as a rather stable structure (Gaertner, Sedikides, & Graetz, 1999; Markus, 1977; but see Onorato & Turner, 2004, for a contrary view). In particular, evidence suggesting that judgments about the self are more stable than judgments about groups is critical for the social projection hypothesis (Krueger, Acevedo, & Robbins, 2005). If group judgments were stable and self-judgments were malleable, any perceived similarities between self and group would indicate that people self-stereotype.

Nonetheless, self-judgments are hardly carved in stone. Consider a person whose political orientation shifts toward more conservative views over time. Perhaps this change simply reflects the jadedness of middle age or improved financial circumstances. More importantly, we suspect that a change of social projection will go along with a change of attitude. If both liberals and conservatives see themselves as a majority, a person changing sides may believe that the new attitude is indeed the more common one.

Longitudinal changes are difficult to track and research does not attempt it very often. It is far easier to introduce new attributes to the self-concept and to change them. When personal feedback comes from a credible source (e.g., when it is ostensibly based on the person's scores on a psychometric test), most people gladly accept it (Forer, 1949). When the feedback changes, they come to believe that they now have a particular attribute that they lacked before, or vice versa. Consistent with the social projection hypothesis, social predictions change in correspondence with the changing feedback (Clement & Krueger, 2002). Similarly, when transient drive states, such as hunger or thirst, are induced, people project these states to others, even to those whom they know to be in a different situational context (Van Boven & Loewenstein, 2003).

CONSTRAINTS ON THE SPREAD OF PROJECTION

Until now, we have assumed that the self is a sample from a particular group. Like other statistical models, the Bayesian induction model assumes that inferences about the properties of a population should rest on samples that were drawn from this same population. This raises the question of whether people use their

own attributes as sample information to make inferences about groups to which they do *not* belong.

Generalization across category boundaries is a general problem. Research psychologists face it every time they ask whether their results obtained with undergraduate students at a particular university generalize to students at other institutions or even to people of different ages or cultures (Sears, 1986). Searching for answers in the discussion sections of research articles usually reveals little about this matter. The issue of generalizability is typically ignored except by authors of handbook chapters, who urge investigators not to overestimate the external validity of their findings.

One popular remedy is to perform replication studies using different participant populations. The current interest in cross-cultural research is an important attempt to find a broader basis for generalization. Here, the effects observed in participating groups are viewed as samples from the most inclusive social category, the world's population. When the variability of these effects is not greater than what one would expect from chance, investigators can claim they have discovered a human universal. Otherwise, cultural differences are the story to be told (Nisbett & Norenzayan, 2002).

The scientific criteria for the random sampling of individuals or groups are idealizations that scientists strive to meet but often cannot. Still, some opportunities to generalize remain even when the data are not fully random. Findings from social-psychological studies, for example, are commonly generalized beyond the population of college students from which the research participants are sampled. Even before the classic studies were replicated elsewhere, it was not unreasonable to think that cognitive dissonance (Festinger & Carlsmith, 1959) and destructive obedience (Milgram, 1963) are phenomena that occur only in California and Connecticut, respectively.

Like scientists, ordinary perceivers need to figure out just how far beyond their own groups they may project their own attributes or behaviors. Suppose you were a participant in the classic projection study by Ross, Greene, and House (1977). The experimenters asked you to assist in a study on mass communication. If you agree, you need to walk around the Stanford campus with a sandwich board reading "Eat at Joe's," or, more ominously, "Repent!" Once you have made a decision, the experimenters ask you to estimate the percentage of Stanford students who agree to participate. As a good Bayesian, you think that about two thirds of the students decide as you did, whatever that may have been (which is what Ross et al. found).

Now suppose you were asked to estimate the percentage of compliance among students at Berkeley or among students at the University of Tobago. What to do? One option is to look for the smallest category that subsumes both groups. Berkeley students can be grouped with Stanford students as students in California, and Tobago students can be subsumed in the category of, well, students. Inasmuch as larger groups tend to be more heterogeneous, inferences from a sample should

lead to smaller changes in belief (see Figure 2.1 or Krueger & Clement, 1996; Rehder & Hastie, 1997, for empirical findings). Stanford students might be less inclined to project to students in general than to students in California, and thus, they may project less to Tobago than to Berkeley students.

Alternatively, one might suspend projection altogether to any group that does not include the self. This does not seem practical, however, because many outgroup members also belong to ingroups according to other schemes of social categorization (Mullen, Hewstone, & Migdal, 2001). Research shows, for example, that women project their own attitudes to other women, but not to men; men project their own attitudes to other men, but not to women (Krueger & Zeiger, 1993). But suppose the same effect occurs when people are categorized by sexual orientation. Now straights only project to other straights, and gays project only to other gays. Next, suppose the effect occurs for categorizations of age, then of race, and so on. The weak projections to outgroups (meta-analytic r between .1 and .15; Robbins & Krueger, 2005) suggest that people overlook alternative ingroup categorizations. This raises the interesting possibility that people may perceive the same individual other as being similar or different from themselves depending on which social category they apply to that person.

The selective application of projection to a salient ingroup matters when the perceiver's own group membership changes. In one study, participants learned that they belonged to a hitherto unfamiliar social group made up of people of a particular psychological type. These participants assumed that others of the same type (but not others of a different type) shared most of their attitudes. When some of these participants were later informed that they belonged, after all, to what they thought to be the outgroup, their pattern of projection reversed itself. Now, they projected to the new ingroup, but not to the new outgroup (Clement & Krueger, 2002). A social-science equivalent of this result is that of a researcher generalizing findings only to freshmen students when thinking that the participants were recruited from this pool, and of generalizing only to juniors when informed by a research assistant that the study participants were, in fact, juniors.

FROM PROJECTION TO CHOICE

The heuristic of social projection is easy to use and it makes social judgments more accurate. If, however, "thinking is for doing," in William James's famous words, one must also wonder how social projection affects behavior. And if projection influences behavior, what are the consequences for the person's well-being and social adaptation?

Everyday predictions are often made under circumstances that are more complex than the sanitized ecology of the research laboratory. Often, a person's own behavior depends on what others do, or on what one thinks they will do. As long as the behaviors of others remain unknown, these behaviors need to be simulated in the perceiver's mind. The question is no longer "What will others do given

that I have done X?" but "What will others do *if* I do X, and what will they do *if* I do Y?" Questions like these lie at the heart of game theories of social behavior. "Players" in social games evaluate an array of outcomes that can result from their own choices in conjunction with the choices of others (Colman, 2003).

The Prisoner's Dilemma

The most famous game is the prisoner's dilemma (PD), which has baffled scientists since it was first proposed (Flood, 1952). The canonical story involves two suspected criminals whom the prosecutor can get convicted for a minor offense. To get them convicted on the major crime, however, she needs a confession. The suspects are held separately and they cannot communicate with each other. The prosecutor visits both and makes the following proposal: "If you confess and your accomplice does not, you will go free and he will be in jail for 12 years. If you confess and your accomplice does too, you will both go to jail for 8 years. If neither of you confesses, you will both be convicted on the lesser charge and sentenced to 4 years."

The sharp suspect, who is motivated to serve as little time as possible, knows that the prosecutor hopes to elicit two confessions allowing her to put both criminals away for a total of 16 years. To avoid jail, he needs to confess while hoping that the other will keep quiet. But because the other receives the same offer, he too hopes to go free by confessing. There is a chance that both suspects confess hoping that the other will not, and thereby end up giving the prosecutor what she wants, namely two sentences of 8 years. Would it not be better for both to refuse to talk? The outcome would be more desirable to the suspects and rather frustrating to the prosecutor. But then again, if the other's silence were somehow ensured, or even merely assumed, the suspect would be tempted to confess in order to go free (Shafir & Tversky, 1992).

Social scientists have not been able to reach consensus on how a choice ought to be made. Their recommendations come from two schools of thought. One is concerned with the way in which inductive inferences inform choice. This view is related to the ideas discussed earlier in this chapter, and we will elaborate on it shortly. First, we consider the alternative approach, which suggests that a player in the PD (or any other experimental game of this sort) select the *dominating* option. A dominating option is one that yields the best result regardless of what the other player does. In the PD, confession yields the best outcome if the other confesses (i.e., 8 years instead of 12), and it also yields the best outcome if the other does not confess (i.e., freedom instead of 4 years). If by confessing the player is better off regardless of what the other player decides to do, confession is a "sure thing" (Savage, 1954), and choosing the sure thing is the rational way to go for a person interested in his own welfare (Dawes & Messick, 2000).

This approach to rational choice brings out the dilemma. If both players choose rationally, both will be worse off than if both refuse to confess. The PD not only pits individual rationality against the collective good, it also leaves the

individual player with the sinking feeling that he would have been better off if he and the other had refused to choose rationally. Yet, both know that they cannot act unilaterally. Realizing that individual rationality tells both to make a confession, the individual player cannot change his mind and decide not to confess. If he did (while the other presumably would not), he would multiply his own years in prison, while letting the other go free. In other words, he would reap the "sucker's outcome." Thus, this theory of rational choice predicts that everyone will confess. Bilateral confession is an equilibrium state because no player can improve his outcome by a unilateral switch.

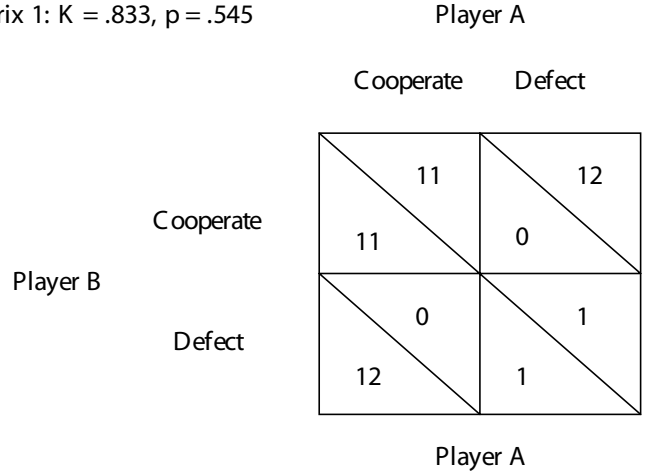
This is a grim picture, and one wants to applaud the prosecutor for coming up with such a shrewd proposal (or rather Professor Albert Tucker for imagining such a prosecutor; see Poundstone, 1992). From the perspective of dominance reasoning, the PD is a dilemma because players can always wonder how they could have done better if neither one of them had confessed. But then again, they should not wonder too much, because both did exactly what had to be done. More problematic for dominance reasoning is the empirical finding that many experimental players do not seem to care about it. Nearly 50% of players cooperate (i.e., select the dominated option; Komorita & Parks, 1995; Sally, 1995). Because players choose independently, 25% of games result in mutual cooperation, 25% in mutual defection, and half in a split outcome.

Consider the consequences of cooperation for the players (and the experimenter's budget). Figure 2.2 shows three payoff matrices. All payoffs may be thought of as dollar amounts. Each player chooses between "cooperation," which is analogous to keeping quiet in the original game, and "defection," which is analogous to confessing. The four possible payoffs can be termed as follows: T is the "Temptation" payoff, representing what happens when the defector successfully exploits a cooperator. R stands for "Reward," which is the payoff for mutual cooperation. P stands of "Punishment," which is the payoff for mutual defection. Finally, S stands for the "Sucker" payoff, which is left to the unilateral cooperator. The defining characteristic of the PD is the inequality $T > R > P > S$.

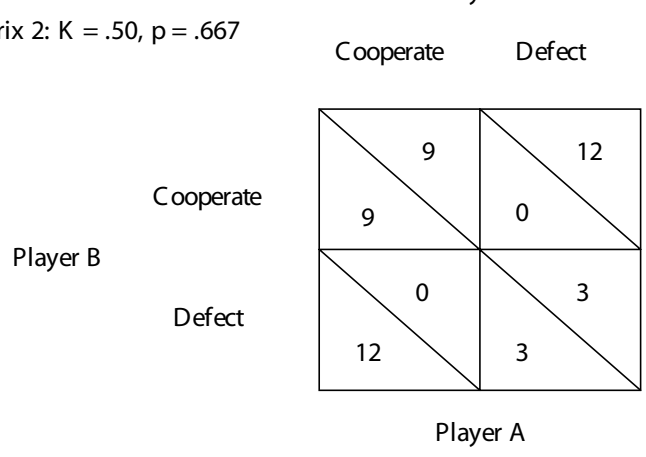
In the three matrices displayed in Figure 2.2, the values for T and S are the same (12 and 0, respectively), whereas the difference between R and P becomes smaller from the top to the bottom matrix. In the top matrix, unilateral defection yields only a small improvement over mutual cooperation, whereas in the bottom matrix, the difference is considerable. This suggests that defection is more compelling in the bottom than in the top matrix. One way to quantify this difference is to divide the difference between R and P by the difference between T and S. According to dominance reasoning, differences in this *K* statistic do not matter as long as the definitional inequality among the four payoffs holds (Rapoport, 1967). All players should defect, and the individual payoffs would be \$1, \$3, or \$5 for the top, middle, and bottom matrix, respectively.

The grand mean of 50% cooperation is somewhat misleading because of the considerable variations from study to study. Figure 2.3 shows the sums of the obtained payoffs for probabilities of cooperation ranging from .1 to .9. If there is

Matrix 1: K = .833, p = .545



Matrix 2: K = .50, p = .667



Matrix 3: K = .167, p = .857

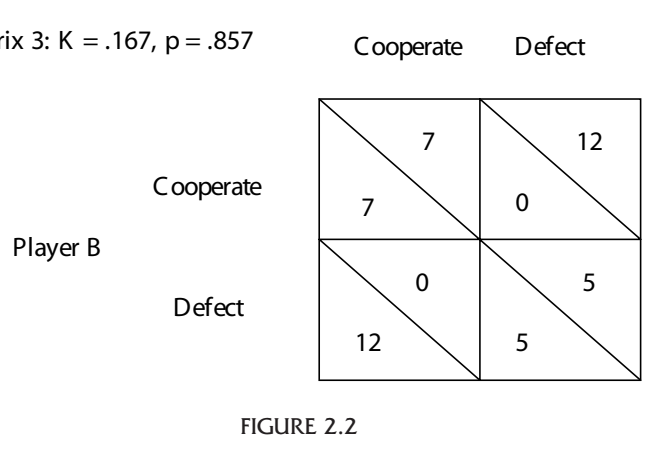


FIGURE 2.2

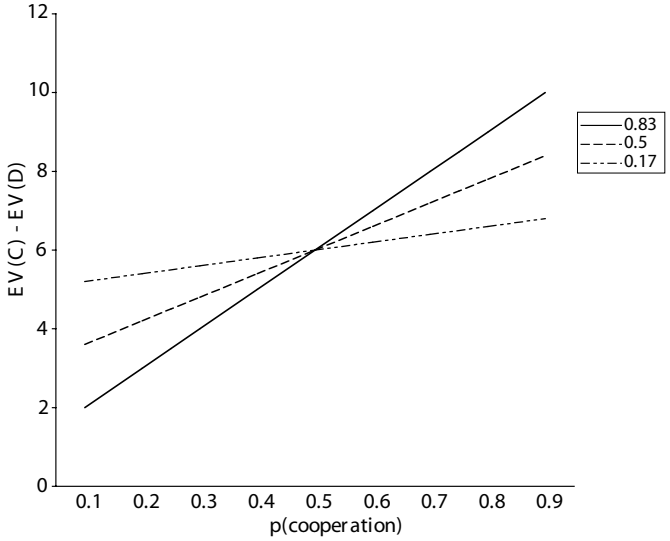


FIGURE 2.3

any cooperation at all, the total success of the players, as defined by their summed payoffs, is always greater than it is when all defect. This is so even if there is only a single cooperator among a million defectors. For the cooperator, the difference between P and S is smaller than the difference between P and T for the defector. The cooperator loses less than the defector gains. As the percentage of cooperators increases, so does the sum of all payoffs, and these gains increase more sharply as *K* becomes larger.

With the moderate to high levels of cooperation routinely observed, dominance reasoning fails as a descriptive model of human choice. This failure is troublesome in two ways. Not only does it leave a gap between normative choice and actual behavior, it also fails to explain why people reap greater benefits when they act irrationally. Dominance reasoning is thus at odds with the many aspects of culture and social policy that are geared toward encouraging cooperative behavior.

When a normative theory of human behavior fails as a descriptive model, there are three options. First, one can insist that people are simply irrational and proceed with research to uncover the psychological sources of this irrationality. Second, one can "repair" the model by introducing auxiliary assumptions. Third, one can look for an alternative model to bring normative rationality into alignment with observed behavior (Gigerenzer & McElreath, 2003). An example of the first type of response is work on the "disjunction effect" (Shafir, 1994). This work suggests that people miss a sure thing because they fail to think through all possible combinations of choices and outcomes. As such, the disjunction effect results from limited attention, effort, or intelligence. An example of the second type of

response focuses on individual differences in how people transform monetary payoffs into personal values. Inasmuch as some people place a greater value on mutual cooperation than others do, they are more inclined to cooperate themselves (De Cremer & van Lange, 2001). This approach is limited to the extent that it needs to postulate preexisting tendencies toward cooperation or defection, which are then realized once a person enters a game. In other words, cooperation is explained by a cooperative disposition (see Rachlin, 2002, who explains altruism with prosocial habits, or Parfit, 1984, who explains cooperation with values attached to benefits reaped by others). To elaborate on an example of the third type of response, we now return to the question of how inductive reasoning might help.

Cooperation after Projection

Now recall the lessons of the Bayesian analysis of social projection: A person may infer that others are similar to him or her even if (or rather, especially if) the sample used to make this inference is a single, self-generated event. We have also seen that social predictions change when the nature of the person's own information changes. In a one-shot PD, one player does not know what the other will do. Both are anonymous, they cannot communicate, and the information they have (i.e., the payoff matrix) is as sparse as the experimenter can make it. Each player only knows that the other is in exactly the same situation, that he faces the same skeletal information, and that he is as ignorant about this player's strategy as this player is about the opponent's strategy. The knowledge of their interchangeability is their common psychological ground (Lewis, 1969; Nozick, 2001).

Up to this point in the analysis, the PD resembles an ordinary prediction situation in which nothing is known about the behavior of others. The difference is that the player has not chosen yet between cooperation and defection. The logic of induction only says that whatever the player will ultimately choose, is—by definition—more likely to be the choice of the majority than the choice of the minority. In other words, it is more likely that the other player will match rather than mismatch his choice.

Now the question is whether a player's choice may be affected by the knowledge that the choice is more likely to be matched than mismatched. Inductive reasoning suggests that a player might as well choose to cooperate when the expected value of cooperation is greater than the expected value of defection. The argument against this idea is that such a choice amounts to the magical belief that one's own cooperation can induce the other to cooperate too. When, however, the absence of any such causal effect is guaranteed, as it is in the standard PD set-up, a return to defection seems obligatory, just as dominance reasoning demands.

To have merit, the inductive approach must respond to this critique. The answer we present has two parts. First, we show that people are sensitive to the probability that others will respond as they themselves do. That is, they generate projective expectations of reciprocity that remain fluid until a final decision, which maximizes the expected value of the outcome, is made. Second, we suggest that

projection-induced choices are valid because they can unfold without implying beliefs in magical causation.

Expectations of Reciprocity

Thinking inductively, a player expects that it is more likely that his opponent will match rather than mismatch his own choice. If the player chooses cooperation, he will expect the other to cooperate; if he chooses defection, he will expect the other to defect. These expectations are purely statistical. The perceived probability of a matching choice, p(M), depends on the hypotheses the player brings to the task. As in the context of a stable self discussed earlier, a player might (but is not obliged to) take a stance of indifference by regarding all hypotheses to be equally probable *a priori*. Then, as we have seen, p(M) = .67.

Whether this expectation leads to cooperation depends on the *K* value of the payoff matrix. For the matrix at the top of Figure 2.2, *K* = .83. The expected value of cooperation, EV[c], is \$7.33 (i.e., and the expected value of defection, \$11.67 + .\$) (1 –.67)), EV[d], is \$4.67 (i.e., \$12 (1 – .67) + .\$1.67). The difference, EV[c] – EV[d], is \$2.67. It pays to cooperate. As *K* diminishes, EV[c] becomes smaller and EV[d] becomes larger. For the center matrix, the two expected values are the same, and the player is therefore indifferent. It can be shown that this point is reached when p(M) = 1/(1+*K*) (Acevedo & Krueger, 2005). Finally, a player will choose defection for the bottom matrix because the difference between the two expected values is \$2.67.

The induction model does not prescribe choice. It only asks players to integrate their own expectations of reciprocity with the given payoff values. The result may be cooperation or defection depending on the player's level of projection and the incentives set by the experimenter. In contrast, the dominance model uses only the relative differences in payoffs. Although the use of less information may seem desirable because it is parsimonious, the fact remains that the data from experimental games are far more consistent with the induction model. Rates of cooperation tend to be intermediate as one would expect from intermediate levels of projection, and rates of cooperation increase with *K* (e.g., Komorita, Sweeney, & Kravitz, 1980).[4]

According to the induction model, people who project more also cooperate more. This hypothesis was supported when the probability of reciprocity was manipulated in an experiment. Participants played several rounds of a PD against a computer (Acevedo & Krueger, 2005). Before each round, they were told the probability with which the computer would match whichever choice they had made. When there was no expectation of reciprocity, cooperative choices were rare (23%). In contrast, cooperation was common when reciprocity was ensured (93% for p(M) = 1). Here, the Temptation payoff and the Sucker payoff were no longer available. The players faced a choice between the payoff for mutual cooperation and the payoff for mutual defection. The most interesting result emerged for the

intermediate level of expected reciprocity. When p(M) = .75, most choices were cooperative, and the rate of cooperation rose with the *K* value of the matrix (54%, 65%, and 80% respectively for *K* = .17, .50, and .83).[5]

In this study, social projection was independently manipulated in terms of the predetermined probability with which a player's choice would be reciprocated. Players found themselves in a position in which they could make choices *as if* they were projecting at a particular rate (see also Baker & Rachlin, 2001). A separate study measured individual differences in social projection independently of the PD situation and to test whether greater projection was associated with a greater willingness to cooperate. Participants rated themselves on a series of personality-descriptive trait adjectives and they estimated the percentages of others who would endorse each trait. The correlation between self-ratings and percentage estimates represented each person's strength of projection. When presented with the PD, those who projected more were also those who more likely to cooperate (r = .18, Krueger & Acevedo, unpublished).

In large-scale social dilemmas, social projection can also be beneficial. Quattrone and Tversky (1984) found that participants in a simulated election expected their own political party to fare better in a national election if they themselves voted (i.e., cooperated) rather than abstained (i.e., defected). Moreover, the strength of this expectation was associated with participants' willingness to vote. They seemed to reason that "If I vote, more supporters of my party will vote than if I abstain. Therefore, I should vote." Note that this reasoning can make a voter hopeful of victory only if intentions are projected selectively to supporters of one's own party, but not to supporters of the opposition. As we saw earlier, projection to ingroups is stronger than projection to outgroups, which makes this possible.

Reasoning Inductively Without Causing Anything

If people were to cooperate in hopes that they could make others cooperate, their thinking would indeed be more magical than normative. Morris, Sim, and Girotto (1998) detected such thinking among players who were more willing to cooperate in a PD when they made their move before the opponent did. Consistent with the induction model, however, expectations of reciprocity predict cooperative choices even when the behavior of others has already occurred (Acevedo & Krueger, 2004). It is sufficient that players assume that their own behavior is *diagnostic* of the behavior of others (see Dawes, 1991, for further distinctions between diagnostic and causal reasoning).

How can inductive reasoning enable cooperation without simultaneously fostering false hopes of exerting a causal influence? To find an answer, we explore two ideas. The first idea is that players can generate different probabilities regarding opponent cooperation depending on whether they themselves are currently contemplating cooperation or defection. The second idea is that players may then choose that behavior which offers the best value.

Changing Predictions

Inductive thinking suggests that one's own choice will be matched with a probability greater than .5 (with p = .67 under the principle of indifference). When players make such estimates after they have committed themselves to their own choice, they are not faulted for projecting. A cooperator's expectation that the other player cooperated is considered as optimal as the defector's expectation that the other player defected (Dawes, McTavish, & Shaklee, 1977; Messé & Sivacek, 1979). But the cooperator and the defector need not be different people. A single individual can anticipate the predictions he would generate if he were a cooperator or a defector. Before settling on a final decision, the player can ask "What is the probability of receiving cooperation if I cooperate?" and "What is the probability of receiving defection if I defect?" In either case the answer is the same (e.g., .67). The expectations generated by one player at two times are as optimal as the expectations generated by two players at the same time. There are no separate statistical rules for the predecisional and the postdecisional phase.[6]

To refute this idea, one would have to deny the equivalence of predecisional and postdecisional induction. One would have to show that there is a separate logic of induction for contemplated behaviors and for enacted behaviors. Alternatively, one would have to deny the validity of induction altogether. This can be a lot of fun as Hume showed, but it makes it difficult to get up in the morning to greet the sun.

Choosing By Expected Value

If the expected value of cooperation is greater than the expected value of defection, a player who is motivated by self-interest will cooperate. The logic of induction is the same regardless of *how* a player generates an expectancy. All that matters is whether one's own cooperation is introduced as evidence. Bayes's Theorem works the same way for a player who cooperates because it makes him optimistic, a player who cooperates to placate a guilty conscience, and a player who cooperates because he does not grasp the dilemma.

In contrast, these distinctions among a player's possible mental states are critical to a defense of dominance reasoning. Hurley (1991), for example, argued that if one "finds oneself [cooperating], that is good news, because of the statistical correlation of such symptomatic acts with the desired symptomatic outcome, but it would be irrational to [cooperate] for the "news value" of that fact that one has [cooperated]" (p. 174). This argument condemns a player for choosing cooperation because of its statistical implications. The player is supposed to have chosen differently. In contrast, a player who *finds himself* cooperating is considered lucky because even devotees of dominance expect him to be a winner.

When a distinction is drawn between the mindful and the unwitting cooperator, a problem arises that is much like the one that is often discussed with regard to the induction model. Recall that a criticism of the induction model is that by cooperating, players cannot generate a statistical similarity between their own behavior and that of others. They can only experience it. With regard to dominance

reasoning, the equivalent charge is that by defecting, players cannot eliminate that same statistical similarity between themselves and others.

Choice in a Deterministic World

Perhaps the focus on a player's mental states does more to obscure than to clarify. The logic of induction unfolds the same way from an observer's perspective. An observer knows that the two players in the PD are interchangeable. With respect to the game, they are both randomly selected specimens. If their choices are revealed one at a time, the first is diagnostic of the second. If the first choice is cooperation, the probability that the second one is also cooperation is .67 (assuming the principle of indifference). In the language of decision theory, the prediction of cooperation is a hit, H, if the second player's choice is indeed cooperation. If the second player defects, the prediction of cooperation is a false positive, FP. Likewise, if the first player defects, the second player is expected to defect with a probability of .67. If the second player defects, the outcome is a correct rejection, CR; if he cooperates, the outcome is a miss, M.

The power of one player's choice to predict the choice of the other can be expressed as an odds ratio, namely the product of the probabilities of correct predictions divided by the product of the probabilities of false ones, or $(H \cdot CR)/(FP \cdot M)$. When the probability of reciprocal choice is .67, as presently assumed, projective predictions are four times as likely to be correct than incorrect. Just like an observer can predict the second player's choice from the first player's choice, each player can predict his opponent's choice from his own. What is more, each player can assume that the opponent's choice predicts his own. The affair is symmetrical, which is easily understood by an observer, but a player is constrained by having to witness his own decision first.

We noted before that the induction model is not concerned with how players arrive at a decision. We now need to qualify this point because, clearly, players (and inmates in Professor Tucker's penitentiary narrative) experience having—and making—a choice between cooperation and defection. They feel the pull of greed (i.e., of being able to reap the Temptation payoff for unilateral defection) and the push of hope (i.e., of achieving mutual cooperation). They may even (falsely) believe that they can make an opponent cooperate by cooperating themselves. These and other mental activities are critical for induction to work because they ensure that players do not make choices at random. If they said, "Since I cannot find a good reason for either option and since I cannot control what the other will do, I might as well flip a coin," the probability of a matching choice would be .5. Only by thinking about the game can players achieve a majority response of some kind. When their choices are strategically nonrandom, a majority will end up favoring one alternative, and an individual player's choice will be diagnostic of it.

An observer who knows this may conclude that the optimistic prediction made by a cooperative player is invalid because of all the cogitation and agitation this player has experienced. Again, this argument invokes dominance reasoning,

which warns that "Thou shalt not make predictions based on thy own behavior if thou chooseth this behavior in order to make that prediction." Yet, this argument overlooks the brute fact that regardless of their individual hopes and fears, most people end up choosing like most others. A champion of dominance reasoning would have to find a way to help a player beat the logic of induction.

How might this be done? Suppose a contemplative player thinks through the changing predictions while considering cooperation and defection. When considering cooperation, opponent cooperation seems likely; when considering defection, opponent defection seems likely (Kay & Ross, 2003, show how imagination can be primed to make it so). The dominance champion now advises the player to make a final switch from considering cooperation to actually defecting. This maneuver must be swift and unilateral. That is, the strategic player must believe that he is faster than the opponent, thereby replacing the belief in interpersonal similarity with the belief in own superiority (Alicke & Govorun, this volume).

Acting more decisively, the dominance champion could offer her mentee, but not the opponent, an opportunity to reconsider his choice after the game is ostensibly done. However enticing it may be to the individual player, this arrangement also destroys the game by violating the premise of common ground. Perhaps the most damaging argument against any such attempt to outflank other players is the impossibility to extend these kinds of special offers to many players without making the unattractive payoff for mutual defection the disappointing norm. Try as one might, induction cannot be outrun. Whatever a player's choice inclination is at a given time, that is what ought to be seen as the most common one.[7]

Projection has the appealing property of acting as a brake on undesirable behavior in social dilemmas. More generally, strong projectors should find it difficult to cheat in a variety of social situations. For would-be cheaters with a conscience, intentions to cheat trigger an unpleasant state of arousal, which then, by virtue of projection, they fear to be obvious to others (Gilovich, Savitsky, & Medvec, 1998). Hence, these intentions are less likely to become actions. For cheaters who act on their designs, projection spoils the fun because now suspicions of others cheating run high (Katz & Allport, 1931; Sagarin, Rhoads, & Cialdini, 1998).

Although a brake on cheating is arguably a good thing, the same logic applies to some desirable behaviors. Creative thinkers and artists, for example, can venture to go where no one has gone before only if they manage to keep projection at bay. If they can't, no idea will seem novel enough to be worth working for. If a creative project is completed nonetheless, projection can spoil it by invoking the "curse of knowledge," which makes the hard-won fruits of imagination and labor obvious in hindsight (Camerer, Loewenstein, & Weber, 1989).

Newcomb's Problem Reconsidered

The clash of inductive reasoning and dominance reasoning highlights the paradox of human choice in a deterministic world. Most of the relevant philosophical arguments have been made with regard to a mind-bending scenario known

as Newcomb's Problem (Campbell & Sowden, 1985). This problem features a person, or player as it were, who is presented with two boxes labeled A and B. Box A is known to contain $1,000. A demon with awesome predictive powers placed $1,000,000 in box B if she predicted that the player would take only that box. If the demon predicted that the player would take both boxes, she left box B empty. What is the player, who is assumed to prefer getting more rather than less money, to do?

The payoffs in Newcomb's Problem show the inequalities familiar from the prisoner's dilemma (Lewis, 1979; Nozick, 1969; 1993). The top panel of Figure 2.4 shows the payoffs for the canonical scenario. Dominance reasoning mandates taking both boxes because the player will be better off by $1,000 regardless of the demon's prediction. In contrast, inductive reasoning suggests taking only one box

Matrix 1: K = .998, p = .5005

Player's Choice	Demon's Prediction: 1 box	Demon's Prediction: 2 boxes
1 box	$1,000,000	$0
2 boxes	$1,001,000	$1,000

Matrix 2: K = .5, p = .6667

Player's Choice	Demon's Prediction: 1 box	Demon's Prediction: 2 boxes
1 box	$750,750	$0
2 boxes	$1,001,000	$250,250

FIGURE 2.4

because of the demon's impressive record of making accurate predictions. Because the *K* ratio is a whopping .998, any p(M) over .5005 can induce a player to forego the second box. The bottom panel shows payoffs for $K = .5$. Here, a greater level of accuracy would be demanded of the demon to justify taking only one box (i.e., $p(M) > .6667$). Just as inductive reasoning does not demand cooperation in the prisoner's dilemma, it does not require a player to settle for one box in Newcomb's problem. The choice between taking only one box or both boxes continues to depend on the expected values of the two alternatives.[8]

Taking only one box in Newcomb's Problem signals the belief that the omniscient demon stocked it, but it does not imply an illusion of influence. From the perspective of induction, the presumed timing of the demon's prediction (i.e., before, after, or concurrent with the player's choice) is irrelevant. The paradox is that a player who cannot claim control over the demon's decision cannot simultaneously believe both that the demon most likely made a correct prediction and that there is still freedom of choice. A player who accepts the statistical association between his own choice and the demon's prediction, and who knows that he cannot influence the demon, must also admit that he cannot influence his own choice. Having to refer to a common cause underlying both the demon's prediction and one's own choice may be a jarring realization for a player who has a strong sense of being in charge of his own decision (Eells, 1985).[9]

A truly free choice is unconstrained by the past; it remains undetermined in the sense that it can still favor either one of the available options. If, when the choice is made, it turns out that the demon was again correct, as she was so many times in the past, the player's experience of free choice can only be an illusion (Wegner, 2002). In dilemmas such as these, people need to act *as if* they had choice, and to do so, they need to think hard before deciding. By accepting the responsibility of choice in the face of determinism, people can discover what they were meant to do.[10] By taking this Taoist path, they can let go of the dilemma, come to understand that they are not unique, and reap the rewards. Cooperating *en masse*, most individuals do well for themselves, while doing good as a collective. Indeed, they do better than they would if they "rationally" sought a behavior that dominates.

CONCLUSION

We began this chapter by noting that there exists an inductive rationale for social projection. Most people are, by definition, members of social majorities. When they know nothing about other group members, their own responses are valid cues to what the majority does. The use of these cues is a "fast and frugal heuristic" (Gigerenzer & Selten, 2001) that leads to more accurate social predictions than a strategy of random guessing. We extended this analysis by noting that the same projective inferences can be made when the properties or preferences of the self change, and to some extent, when the social group does not include the self. Our

main objective was to show that social projection may affect one's choices when no preexisting preferences exist. The one-shot Prisoner's Dilemma served as the paradigm for this discussion. In it, players are said to face a conflict between self-interest and collective interest, and researchers are faced with choosing between irreconcilable theories of rationality.

Our analysis addressed both conflicts. From the player's point of view, social projection offers an opportunity to cooperate out of self-interest. This is so because cooperation has the highest expected value when projection is strong. The fact that the opponent also benefits is of no consequence. Of course, this view does not imply that altruistic motives, commitments to do one's duty, or the limitations of mindful thinking never play any role in social interaction. It simply asserts that many choices in the PD can be predicted without recourse to any of these psychological variables. From the researcher's point of view, it should be reassuring that the logic of induction in general, and the psychological phenomenon of social projection in particular, apply to both selves as entities and selves in flux.

Perhaps most importantly, the framework of induction offers an explicit way to think about how people make choices in a deterministic world. But a difficulty remains: How can people accept determinism as a scientific doctrine, and continue to act as if they had freedom of choice? The pragmatist William James declared that his first act of free will was his decision to believe in it. Edward Lorenz, the founder of modern chaos theory, offered a less paradoxical strategy. "We must then wholeheartedly believe in free will. If free will is a reality, we shall have made the correct choice. If it is not, we shall still not have made an incorrect choice, because we shall not have made any choice at all, not having a free will to do so" (Lorenz, 1993, p. 160). This is good advice to inductive thinkers. It allows them to cooperate in social dilemmas and to bet on a single box in Newcomb's problem without having to worry about being accused of magical thinking.

ACKNOWLEDGMENTS

We are indebted to Theresa DiDonato and Alexandra Freund's for their insightful comments on a draft version of this chapter.

NOTES

1. Laplace (1814) suggested that all hypotheses be regarded as equally probable before evidence is gathered. This (controversial) idea is variously known as the principle of insufficient reason or the principle of indifference (Keynes, 1921; see Howson & Urbach, 1989, for review and discussion).
2. Alternatively, the degree of belief revision can be expressed by a ratio of p_1 over p_2, but this choice in metric has little effect on the present analysis because

$$\frac{p_1}{p_2} = 1 + \frac{p_1 - p_2}{p_2}.$$

3. It can be shown that belief revision is at its maximum when the prior probability of the hypothesis, $p(H_1)$ is equal to

$$\frac{p(A \mid H_2) - \sqrt{p(A \mid H_1) \cdot p(A \mid H_2)}}{p(A \mid H_1) - p(A \mid H_2)}.$$

4. Extreme sets of payoffs readily illustrate this effect. If T, R, P, and S were, respectively, 100, 99, 1, and 0, cooperation would come more easily than if the payoffs were 100, 51, 49, and 0. According to the dominance model, the differences between these two sets should not matter because both satisfy the inequalities that define the PD.
5. The induction model assumes that people compute expected values for cooperation and defection, as well as the difference between the two. When the difference is positive, they cooperate. If these computations were error free and p(M) = .75, everyone would cooperate if $K > .33$. Because estimates cannot be completely reliable, cooperation drifts toward 50% as p approaches $1/(1+K)$.
6. When social projection is recognized as a mental process that affects choices in the PD, postgame predictions of the opponent's choice appear in a different light. For cooperators, the inductive model suggests that projection led them to cooperate, whereas defectors may be attempting to justify their choice after the fact by claiming that others would do the same (see Arndt, Greenberg, Solomon, Pyszcynski, & Schimel, 1999, for research on defensive projection).
7. Many motorists try to outrun induction (and traffic) by deftly and frequently switching lanes. If they projected more, they would be less surprised when ending up in the most clogged lane more than 1/k of the time (where k is the number of lanes). Unnecessary frustrations in the grocery check-out line stem from the same source (Surowiecki, 2004).
8. Also analogous to the prisoner's dilemma is the fact that the odds of making a choice that matches the prediction are the same as the odds of making a prediction that matches the choice. In Newcomb's Problem, p(demon 1-box|player 1-box) differs from p(player 1-box|demon 1-box) if p(demon 1-box) ≠ .5 (Levi, 1975). In the PD, the principle of indifference ensures that these two conditional probabilities are the same.
9. Although the PD and Newcomb's Problem pose a similar prediction paradox, there is a difference. Newcomb's Problem does not make the assumption of common ground. The demon and the player are different creatures, with the demon being the one who knows more. Therefore, the unilateral advice for switching (here, to take both boxes) will not work as well. The very definition of Newcomb's Problem entails that the demon foresees all factors influencing the player's decision, thereby including the advice of those who believe that the demon can be tricked.

   In an informal survey, we found that participants ($N = 84$) were more comfortable taking only one box in Newcomb's Problem (68%) than they were cooperating in the prisoner's dilemma (37%). This difference may be explained by the huge K ratio (i.e., .998) of Newcomb's payoffs, and the fact that the demon's powers were touted as great. The same difference may explain the lack of a correlation between choices in the two contexts (= –.05).
10. As Schopenhauer (1985) advised, "From what we do we know what we are" (p. 98).

REFERENCES

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

Acevedo, M., & Krueger, J. I. (2005). Evidential reasoning in the prisoner's dilemma game. *American Journal of Psychology, 118*(3).

Allport, F. H. (1924). *Social psychology*. New York: Houghton Mifflin.

Arndt, J., Greenberg, J., Solomon, S., Pyszcynski, T., & Schimel, J. (1999). Creativity and terror management: Evidence that creative activity increases guilt and social projection following mortality salience. *Journal of Personality and Social Psychology, 77,* 19–32.

Asch, S. E. (1952). *Social psychology*. Oxford: Prentice-Hall.

Baker, F., & Rachlin, H. (2001). Probability of reciprocation in repeated prisoner's dilemma games. *Journal of Behavioral Decision Making, 14,* 51–67.

Bargh, J. A., & Ferguson, M. J. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin, 126*, 925–945.

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy, 97,* 1232–1254.

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

Clement, R. W., & Krueger, J. (2002). Social categorization moderates social projection. *Journal of Experimental Social Psychology, 38,* 219–231.

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

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

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

Dawes, R. M. (1991). Probabilistic versus causal reasoning. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology, Volume 1: Matters of public interest* (pp. 235–264). Minneapolis: University of Minnesota Press.

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

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

De Cremer, D., & Van Lange, P. A. M. (2001). Why prosocials exhibit greater cooperation than proselfs: The role of social responsibility and reciprocity. *European Journal of Personality, 15,* S5–S18.

Eells, E. (1985). Causality, decision, and Newcomb's Problem. In R. Campbell & L. Sowden (Eds.). *Paradoxes of rationality and cooperation: Prisoner's dilemma and Newcomb's problem* (pp. 183–213). Vancouver: University of British Columbia Press.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology, 58,* 203–210.

Flood, M. M. (1952). *Some experimental games. Research Memorandum RM-789.* Santa Monica, CA: RAND Corporation.

Forer, B. R. (1949). The fallacy of personal validation: a classroom demonstration of gullibility. *Journal of Abnormal and Social Psychology, 44,* 118–123.

Gaertner, L., Sedikides, C., & Graetz, K. (1999). In search of self-definition: Motivational primacy of the individual self, motivational primacy of the collective self, or contextual primacy? *Journal of Personality and Social Psychology, 76,* 5–18.

Gigerenzer, G., & McElreath, R. (2003). Social intelligence in games. *Journal of Institutional and Theoretical Economics, 159,* 188–194.

Gigerenzer, G., & Selten, R. (2001). *Bounded rationality: The adaptive toolbox.* Cambridge, MA: The MIT Press.

Gilovich, T., Jennings, D. L., & Jennings, S. (1983). Causal focus and estimates of consensus: An examination of the false-consensus effect. *Journal of Personality and Social Psychology, 45,* 550–559.

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology, 75,* 332–346.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–1480.

Heider, F. (1958). *The psychology of interpersonal relations*. Hillsdale, NJ: Erlbaum.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. Chicago: Open Court Publishing.

Hurley, S. L. (1991). Newcomb's Problem, Prisoners' dilemma, and collective action. *Synthese, 86,* 173–196.

Katz, D., & Allport, F. L. (1931). *Students' attitudes*. Syracuse, NY: Craftsman Press.

Kay, A. C., & Ross, L. (2003). The perceptual push: The interplay between implicit cues and explicit situational construals on behavioral intentions in the Prisoner's Dilemma. *Journal of Experimental Social Psychology, 39,* 634–643.

Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan.

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

Komorita, S. S., Sweeney, J., & Kravitz, D. A. (1980). Cooperative choice in the n-person dilemma situation. *Journal of Personality and Social Psychology, 38,* 504–516.

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.

Krueger, J. (2000). The projective perception of the social world: A building block of social comparison processes. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 323–351). New York: Plenum/Kluwer.

Krueger, J. I., & Acevedo, M. (unpublished). *Person perception in the Prisoner's Dilemma*. Brown University.

Krueger, J. I., Acevedo, M., & Robbins, J. M. (2005. Self as sample. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition.* New York: Cambridge University Press.

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

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

Laplace, P. S. (1814). *Essai philosophique sure les probabilities.* Paris: Courcier.

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

Lewis, D. K. (1969). *Convention: A philosophical study.* Cambridge, MA: Harvard University Press.

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

Lorenz, E. N. (1993). *The essence of chaos.* Seattle: University of Washington Press.

Markus, H. R. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology, 35,* 63–78.

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.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.

Morris, M. W., Sim, D. L., & Girotto, V. (1998). Distinguishing sources of cooperation in the one-round prisoner's dilemma: evidence for cooperative decisions based on the illusion of control. *Journal of Experimental Social Psychology, 34,* 494–512.

Mullen, B., Migdal, M. J., & Hewstone, M. (2001). Crossed categorization versus simple categorization and intergroup evaluations: A meta-analysis. *European Journal of Social Psychology, 31,* 721–736.

Nisbett, R. E., & Norenzayan, A. (2002). Culture and cognition. In H. Pashler & D. L. Medin (Eds.), *Stevens' Handbook of Experimental Psychology: Cognition* (3rd ed., vol. 2, pp. 561–597). New York: Wiley.

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, Holland: Reidel.

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

Nozick, R. (2001). *Invariances: The structure of the objective world.* Cambridge, MA: Harvard University Press.

Onorato, R. S., & Turner, J. C. (2004). Fluidity of the self-concept: The shift from personal to social identity. *European Journal of Social Psychology, 34,* 257–278.

Parfit, D. (1984). *Reasons and persons*. Oxford: Clarendon Press.

Popper, K. R. (1957). *The poverty of historicism*. New York: Harper & Row.

Poundstone, W. (1992). *Prisoner's dilemma*. New York: Doubleday.
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.
Rachlin, H. (2002). Altruism and selfishness. *Behavioral and Brain Sciences, 25,* 239–296.
Rapoport, A. (1967). A note on the index of cooperation for prisoner's dilemma. *Journal of Conflict Resolution, 11,* 101–103.
Rehder, B., & Hastie, R. (1996). The moderating influence of variability on belief revision. *Psychonomic Bulletin and Review, 3,* 499–503.
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.
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.
Sagarin, B. J., Rhoads, K. L., & Cialdini, R. B. (1998). Deceiver's distrust: Denigration as a consequence of undiscovered deception. *Personality and Social Psychology Bulletin, 24,* 1167–1176.
Sally, D. (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958–1992. *Rationality and Society*, 7, 58–92.
Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.
Schopenhauer, A. (1999). *On the freedom of the will*. New York: Blackwell.
Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data set on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51,* 515–530.
Shafir, E. (1994). Uncertainty and the difficulty of thinking through disjunctions. *Cognition*, 50, 403–430.
Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. *Cognitive Psychology, 24,* 449–474.
Shedler, J., Mayman, M., & Manis, M. (1993). The illusion of mental health. *American Psychologist, 48,* 1117–1131.
Surowiecki, J. (2004). *The wisdom of crowds*. New York: Doublday.
Tetlock, P. E. (2002). Theory-driven reasoning about plausible pasts and probable futures in world politics. In T. Gilovich, D. Griffin & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 749–762). New York: Cambridge University Press.
Van Boven, L., & Loewenstein, G. (2003). Social projection of transient drive states. *Personality and Social Psychology Bulletin, 29,* 1159–1168.
Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: MIT Press.
Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value and potential for improvement. *Annual Review of Psychology, 55,* 493–518.
Wohlschläger, A., Haggard, P., Gesierich, B., & Prinz, W. (2003). The perceived onset time of self- and other-generated actions. *Psychological Science, 14,* 586–591.