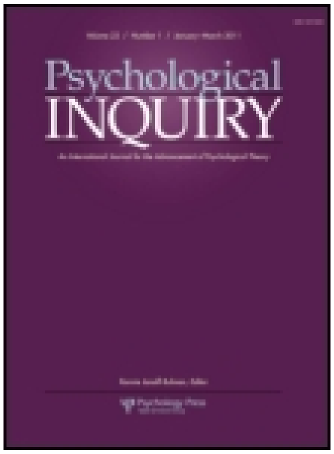

Psychological Inquiry: An International Journal for the Advancement of Psychological Theory
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hpli20

Twilight of a Dilemma: A Réplique

Joachim I. Krueger [a] , David Freestone [a] & Theresa E. DiDonato [b]

[a] Department of Cognitive, Linguistic and Psychological Sciences , Brown University , Providence , Rhode Island

[b] Department of Psychology , Loyola University of Maryland , Baltimore , Maryland

Published online: 19 Mar 2012.

To cite this article: Joachim I. Krueger , David Freestone & Theresa E. DiDonato (2012) Twilight of a Dilemma: A Réplique, Psychological Inquiry: An International Journal for the Advancement of Psychological Theory, 23:1, 85-100, DOI: 10.1080/1047840X.2012.666930

To link to this article: http://dx.doi.org/10.1080/1047840X.2012.666930

Psychological Inquiry, 23: 85–100, 2012
Copyright © Taylor & Francis Group, LLC
ISSN: 1047-840X print / 1532-7965 online
DOI: 10.1080/1047840X.2012.666930

Psychology Press
Taylor & Francis Group

REPLY

Twilight of a Dilemma: A Réplique

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

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

In light of the nine open peer commentaries, we further explicate the social projection model and highlight its boundaries. We acknowledge that our model—like any extant model—cannot account for all the available empirical data. Yet the model is strong because it explains cooperation in a variety of social dilemmas and experimental games while deploying only one psychological construct as a parameter. Compared with its competitors, the social projection model more openly recognizes the struggle with uncertainty a person caught in a social dilemma must confront. In a series of simulations (and brute math), we show that cooperation can survive when players (humans) stay the Bayesian course.

Not to be absolutely certain is, I think, one of the essential things in rationality.—Bertrand Russell

"Social projection can solve social dilemmas." This is the title of our target article in this issue. The article is ambitious, and most of our commentators acknowledge that we are onto something. They engage, as we hoped they would, constructively in the conversation, elaborating the properties of their own models and calling us to task where we overreached. We hope that as a packet, the target article with the nine responses and this rejoinder will make a useful contribution to one of the great puzzles in the study of human (and nonhuman) behavior: Why do so many individuals choose to cooperate with others when their self-interested rationality seems to demand defection, no matter what others do. Our solution is that cooperation can, in fact, be compatible with self-interest. To see why, we have extended the rationale of Bayesian induction from the postdecisional behavioral stage to the predecisional deliberation stage. It is here that strategic reasoning takes place. Projection and induction after the dice are cast is also interesting, but it might be a mere rationalization of behavior that has already occurred (Krueger & DiDonato, 2010).

By suggesting that cooperation can be compatible with self-interested rationality, our model is more than just an addition to the growing suite of theories attempting to explain anomalous behavior. Our model overcomes the dilemma between individual interest and collective need, whereas most other models seek to find some way to integrate the two, either by finding some weighted utility function (Kerr, this issue; Van Lange & Van Doesum, this issue) or by debasing individual rationality until collective interest can be satisfied epiphenomenally (Chater & Vlaev, this issue). Our model can lay claim to descriptive, normative, and adaptive validity. The model is descriptively valid because it can predict rates of cooperation in a variety of games with strategic uncertainty and across variations in the particular payoff arrangements. The model is normatively valid because under the conditions specified, the Bayesian inferences about others' behavior are coherent. The model is adaptively valid because it can explain how cooperators survive in repeated encounters (more on this later).

Last, but not least, our model is more parsimonious than its competitors. Some (Kerr, this issue; Van Lange & Van Doesum, this issue) argue that the problem of cooperation is irreducibly complex and that multiple models are needed or at least models with multiple parameters. Others (Chater & Vlaev, this issue; Colman & Pulford, this issue; Yamagishi, this issue) rely on poorly specified extraneous variables to make their models work. To some, parsimony is a false goddess, but to us she is appealing. She arbitrates disputes where the evidence or the normative rules of science are not enough to do the job (Pitt & Myung, 2002). One normative rule is extolled in textbooks and ignored in research practice. Popper (1962) urged that theories that do not do well be abandoned, but in our Lakatosian (1978) era they tend to survive. The price we pay for keeping an abundance of theories in play is that many phenomena are so multidetermined that any specific explanation does not mean much. It seems that somehow and somewhere, any particular hypothesis can make a contribution. Parsimony protects against theory inflation. If we do not have sufficient grounds to reject a particular hypothesis, we might be able to rank hypotheses according to standard criteria (e.g., explanatory and generative power) as well as parsimony.

The target article plus commentaries plus rejoinder format of the conversation is progress over the standard practice of waiting (and hoping) for a publication to have impact. Authors and audiences benefit from the open exchange of ideas. Yet the format falls short of a counterfactual ideal, in which claims, critiques, and rebuttals are exchanged in a round-robin fashion. Even in this rejoinder, we cannot give each commentary equal attention (although we have a strong social preference for equality). We focus on the critical commentaries, believing that this tactic has some Gricean appeal. It allows us to say something new.

Other Induction Models

Our model does not stand alone. Pothos and Busemeyer (2009) and Fischer (2009) have independently developed theories of inductive reasoning, which show how perceptions of self-other similarity can trigger cooperative action. Like our model, these two models make the radical assumption that individual decision makers can rationally entertain different expectations about what others will do and that these expectations are correlated with their own action tendencies. From the point of view of orthodox probability theory, this idea is heresy, and its proponents should be burned at the stake. So it is gratifying to see that support for this break with orthodoxy is broadening.

With regard to the particular psychological processes, Pothos and Busemeyer originally theorized along the lines of cognitive dissonance, but now they seem to prefer the concept of social projection (Busemeyer & Pothos, this issue). Fischer (this issue) suggests that the social projection model is a special case of his Subjective Expected Relative Similarity (SERS) model. In a series of elegant experiments, he shows that once self-other similarity is manipulated, cooperation follows with corresponding strength. Incidentally, we pursued the same strategy in our initial experiments (Acevedo & Krueger, 2005). Yet this strategy alone is not enough to show where perceptions of similarity come from when they are not manipulated in the laboratory. Here the social projection model makes a unique contribution.

Toma and Woltin (this issue) make significant progress toward an integration of predecisional with postdecisional projection. Whereas our work is focused on the former, Toma and Woltin show in their experiments how the two types of projection can reinforce each other. Moreover, they argue that sometimes projection does not reduce to the cold calculus of induction but that instead people have a motivated interest in cooperation, and that this interest can stimulate further inductive reasoning. In short, the predecisional social projection is embedded in a dynamic loop (see also Parks, this issue). Some commentators note that we say little (actually, nothing) about the origins of individual differences in social projection. We now think that individual differences in motivated cognition can play a role to answer this question. Parks (this issue) notes that many people dislike uncertainty, and some abhor it (cf. Hogg, 2007). We suspect that this dislike also contributes to individual differences in social projection.

Committing Errors Without Being Stupid

Perhaps some people cooperate in a social dilemma because “they just don’t get it,” but random error does not do particularly well as an explanatory construct. There is another meaning of error, as in “error management” (Haselton et al., 2009), which is more promising. People know that they may err, and the social projection model implies as much. Unless people project perfectly, which very few do, they must anticipate that they could be wrong. Cooperators must consider the possibility of regret that will come if defectors sucker them, and defectors must consider the possibility that cooperators might guilt-trip them (Krueger & DiDonato, 2010). The error-managing decision rule is this: Cooperate if guilt is more aversive than regret, otherwise defect. This is a high hurdle, because cooperators meeting defection look particularly dumb (Krueger & Acevedo, 2007), which makes the predictive success of the projection model all the more remarkable.

As this example shows, error may have method. This possibility is part of an explanation that is seldom seen in discussions of social dilemmas, namely, foraging. People may veer from time to time into cooperation even if they recognize that defection is the dominating strategy. Mutual defection is, as it were, a depleted patch of food (Hardin, 1968), and mutual cooperation is more nourishing. Of course, leaving a depleted, though not entirely barren, patch in hopes of finding a more productive one is risky business; one might stumble into a sucker’s patch, which bears no fruit at all.

Animal models show that foraging with the mixed strategy is not a bad idea (Krueger, 2012a), even though it looks like irrational probability matching (Tversky & Edwards, 1966). Exploration requires a departure from stereotyped behavior of exploitation even if that behavior maximizes output for the moment. In an environment that is subject to change (and which environment is not?), the willingness to keep one's options open is advantageous in the long run (Cohen, McClure, & Yu, 2007). Kavanau (1967) reported that "the habit of deviating fairly frequently from stereotyped 'correct' responses, together with a high level of spontaneous activity, underlie the remarkable facility with which white-footed mice can be taught to cope with complex contingencies" (p. 1628). What's good for the mouse is good for the man. Even if people are not looking for greener patches, they may simply value their freedom to choose. If the strategy of defection were all but a foregone conclusion, there would be no choice in it—and no fun. Perhaps cooperation is not so much a cognitive error but a principled act of defiance (Brehm, 1966) or the expression of "an evolved tendency toward occasional spontaneity" (Waller, 2011, p. 56).

Stumbling into Simpson's paradox is yet another kind of error, which is central to Chater and Vlaev's (this issue) account of cooperation. Our concern about their model is that it leaves the door open to circularity. The model assumes that people tend to cooperate inasmuch as they have experienced the rewards of mutual cooperation and the penalties of mutual defection in the past, when there was a lot of cooperation in nice games and little cooperation in nasty games. The trouble is that the nice-versus-nasty differential in cooperation is both the phenomenon to be described and the principle that explains it.

We agree with Chater and Vlaev (this issue), however, that learning models should be carefully considered so that we can see what they can tell us about how people navigate social dilemmas over time. Once a differential reward structure is in place, it would be surprising if it were ignored; and, as Chater and Vlaev note, people cannot always be counted on to mentally correct for confounding variables (Meiser & Hewstone, 2004; Schaller & Maass, 1989). Although Chater and Vlaev redescribe their model, they do not present an alternative view on how the process can be started if not by social projection. The allusion to "some factor, independent of the players' deliberations, which tends to influence players moves in the same way" (p. 37) has limited utility until we learn more what those factors might be. Certainly, it is not enough to appeal to common sense. If it is "obvious" that one would cooperate in an easy game and defect in a nasty game, it would trivially follow that one would cooperate with an intermediate probability in a middling game. There would be nothing left to explain. The history of having been rewarded for cooperation in nice games but not in nasty games is no longer necessary as a causal process. To overcome these difficulties, we explore in the last section of this article how learning models can be integrated with the social projection model.

At this point, we note that our "last-minute-intrigue" experiments point to a specific shortcoming of the Thorndikean learning model. The law of effect entails what game theorists have misnomered the PAVLOV strategy of win–stay/lose–shift. Individuals following this strategy cooperate after mutual cooperation and shift to cooperation after mutual defection, but they defect after unilateral defection and shift to defection after unilateral cooperation. Our results show that expectations regarding the other's choice are critical. Most important, and in contrast to PAVLOV-Thorndike, people shift to cooperation only if they, after mutual defection, believe the other would do the same.

Social Projection Does Not Collapse

When we wrote the target article, we expected spirited defenses of the defection-by-the-sure-thing-principle argument, but there are few. No commentator protests our heretical admission of multistable probability estimates of others' cooperation. They do not provide arguments to dispute our analysis of Newcomb's problem or the claim that the Bayesian defense of one-box strategy is logically equivalent to the strategy of cooperation in a social dilemma. Instead, Colman and Pulford (this issue) baldly state that

> [t]here is no debate among game theorists about whether or not it is rational to defect in the Prisoner's Dilemma game . . . [, that] it must be rational to choose that action [of defection] even if [we] do not know the circumstances . . . [and that] there is nothing of substance to debate on that score. (p. 40)

All that is correct, of course, if one asks only classic game theorists. The present discussion proves otherwise when social scientists of different stripes are included (cf. Campbell & Sowden, 1985; Nozick, 1969).

Chater and Vlaev (this issue) take a different tack by trying to show that evidential reasoning (i.e., social projection) "collapses" once it is recognized that many different reasoning processes can lead to the same decision. It may well be true that many psychological processes can support the same choice, but this does not affect inductive reasoning about the commonness of that choice. The statistical logic that one's own behavior, whatever it may be, is more likely to be the behavior of the majority than the behavior of the minority is silent on the sources of that behavior. Chater and Vlaev's argument would seem to suggest that the expectation that you are trustworthy because you are a Quaker, whereas I am trustworthy because I am a Calvinist, undermines my ability to project trustworthiness to you. That is not so. It undermines only my ability to project Calvinism to you.

In their trivia test example, Chater and Vlaev (this issue) suggest that you would no longer expect me to think that the capital of Germany is Bonn if you found yourself mentally stuck on this obviously wrong answer. Evidential reasoning does not collapse here, however, but instead shows its flexibility. What is happening in the example is that the distribution of prior probabilities has changed. Apparently, I already have reason to believe that Bonn is the wrong answer, aside from the fact that it has come to mind. The distribution of priors is critical to the Bayesian projection model. The assumption of uniform priors we made in the target article is a convenient and defensible simplification to represent the psychological state of ignorance (Dawes, 1989; Laplace, 1783/1953).

There is another way in which Bayesian induction can handle the Bonn–Berlin scenario without undermining the social projection hypothesis. Suppose the assumed probability of producing the correct answer is .8 and that there are only two hypotheses, Bonn and Berlin. Further suppose that you have no reason to believe either is more likely to be correct; each is correct with a prior probability of .5. If Bonn comes to mind, the probability of that happening if Bonn were the correct answer is .8. Bayes's Theorem says that the probability of Bonn being the correct answer given that this answer came to mind is the prior probability of Bonn being correct times the probability of Bonn coming to mind if it is correct divided by the overall probability of Bonn coming to mind, where the latter is the prior probability of Bonn being correct times the probability of Bonn coming to mind if it is correct plus the prior probability of Berlin being correct times the probability of Bonn coming to mind if Berlin is correct. Formally, this mouthful presents itself as

$$\mathrm{p}\frac{(\text{Bonn correct})}{(\text{Bonn in mind})} = \frac{\mathrm{p}(\text{Bonn}) \times \mathrm{p}(\text{Bonn in mind}|\text{Bonn correct})}{\mathrm{p}(\text{Bonn}) \times \mathrm{p}(\text{Bonn in mind}|\text{Bonn correct}) + \mathrm{p}(\text{Berlin}) \times \mathrm{p}(\text{Bonn in mind}|\text{Berlin correct})}.$$

or $\frac{.5\times.8}{5\times.8+.5\times.2} = .8$. The epistemic doubt that one might have gotten it wrong translates into a higher p(Bonn in mind|Berlin correct), which reduces p(Bonn correct|Bonn in mind) because it increases the base rate probability of Bonn coming to mind (i.e., the denominator of the likelihood ratio).

Chater and Vlaev (this issue) further suggest that evidential (projective) reasoning collapses as the expected probability of similarity moves down toward .5. This is not so; attenuation of projection only reduces the probability of cooperation, and this an integral part of the model. As our model makes explicit (see target article), much depends on the variability of projection scores around the population mean.

Citing Gauthier (1986) and Howard (1988), Colman and Pulford (this issue) endorse the hypothesis that projection to one's own *Doppelgänger*[1] would be rational if such a mirror-image person existed. They do not seem to realize that the social projection hypothesis is the very same argument generalized over the probability space. Yet they claim that it is irrational to act on the basis of expected reciprocity when $p_r < 1$, "because the players choose their strategies independently" (p. 45). If it is rational to project if the other is identical to the self, $p_r = 1$, why is it irrational to do so if the other is just very similar to the self, $p_r = .99$ (with as many 9s added as you wish)? In our target article, we explain that although individuals choose independently from each other, the choices of both depend on the same base rate. That base rate is the common cause that drives the statistical association between choices.

Citing Weber (1904–1905), Colman and Pulford (this issue) make a final pitch for the presumed collapse of projection. To them, the Calvinist calculus is transparently irrational. Classic Calvinists believe that mundane success cannot cause salvation but that success is correlated with salvation due to the common cause of God's plan. Hence, Calvinists work hard to signal the attainment of the state of grace to themselves. We argue that this is coherent in the Bayesian sense (Krueger & Acevedo, 2008) but that it is fallacious if the premise is false. If God does not exist or if she is mischievous, hard work only increases the probability of earthly success, but it does not reveal a place in heaven. Calvinists, in other words, tend to get wealthy in spite of themselves.

One-Shot Cooperation as a Derivate of Long-Term Interests

It is often noted that cooperation may be useful in repeated interactions but that it is strictly irrational in a one-shot anonymous game. Colman and Pulford (this issue) distinguish between games with a finite number of rounds and infinitely repeated games. For the former, they argue with Luce and Raiffa (1957) that backward induction provides compelling proof for the dominance of defection in *any* round. Yet, they suggest that if the game is "repeated an indefinite number of times, then the optimal strategy for the rational player is far from clear, and there are valid reasons to cooperate, the most obvious being reciprocity" (p. 41).

The term "indefinite" is critical here. If it means "infinite," we need not worry because, thanks to mortality, the prospect of endless play is itself merely a playful notion. If we are, however, referring to a situation in which we simply do not know how many

[1] A literary device denoting a (usually sinister) double of a living person.

rounds will be played, the Luce and Raiffa argument holds. Suppose that, after working through the logic of backward induction, you defect if you know there are eight rounds and you defect if you know there are six rounds. Now the sure-thing principle dictates that you must also defect if you do not know if you are in an eight-round game or in a six-round game (or in *any* finitely repeated game for that matter). We return to the issue of repeated play in the final section of this article and show that cooperation can be rationally sustained regardless of the number of rounds and whether players know this number.

The decision to cooperate must be made against the pull of defection in the one-shot game and in multishot games alike. But how does it work, if not by social projection? Colman and Pulford (this issue) suggest strong reciprocity as an alternative. Reciprocity is strong if it comes with a taste for vengeance. Strong reciprocators not only reward cooperators with cooperation but also punish defectors at own cost. This strategy tends to bring free riders to heel (Fehr, Fischbacher, & Gächter, 2002), but it creates its own epistemic puzzle. In groups of players, the decision to punish is itself a volunteer's dilemma (Diekmann, 1985). Solving one dilemma with another amounts to dilemma replacement, and it raises the specter of infinite regress. Incidentally, social projection can also contribute to the solution of the volunteer's dilemma, but that's a story for another day.

Perhaps inclusive fitness is the answer. Colman and Pulford (this issue) review the evidence for this powerful concept, but they cannot apply it to one-shot anonymous games without resorting to the notion of error. The question remains of why people would overgeneralize a sensitivity to inclusive fitness to the one-shot anonymous situation. Perhaps it is a matter of error management, as we previously noted, or "there may be an instinctive pattern of behavior in certain circumstances, and one that is likely to have evolved by natural selection" (Colman & Pulford, this issue, p. 42). This conclusion smacks a bit of McDougal's circular psychology, which Kerr (this issue) aptly criticizes.

Collectivism

Colman and Pulford (this issue) consider other-regarding preferences and team reasoning as promising accounts of cooperation. Their defense of other-regarding preference is somewhat surprising in light of Colman's (2003) critique that such preferences fall flat in coordination games (Kerr, this issue, agrees). Like Colman and Pulford, we see that benevolence (regard for the other) can turn cooperation into the dominating strategy. The problem is that these preferences have to be quite strong before they can turn defection into cooperation. Van Lange (1999) found that benevolent values can be significant(ly above zero), but they tend to be weaker than self-interested values, even among prosocials. Van Lange and Van Doesum are also skeptical with regard to benevolence; they regard the preference for fairness as more important.

With regard to team reasoning, Colman and Pulford (this issue) think that our theoretical sketch is a grotesque misrepresentation ("travesty") of the real thing. They assert in the "technical literature" (Bacharach, 1999) individuals must expect others to team reason before they themselves can adopt this way of thinking. We could not agree more. Team reasoning devolves into either social preferences by assuming very strong benevolence or social projection by requiring the expectation that others will do likewise. Where are the unique predictions? In earlier work, Colman, Pulford, and Rose (2008) suggested the Ball Gown Game as a setting for unique team reasoning, but we find that social projection can also handle this situation (see Target article). Colman and Pulford (this issue) do not dispute our reframing of the game.

In their concluding remarks, Colman and Pulford (this issue) seek to disarm one of our central claims, namely the idea that social projection can account for the nice-versus-nasty effect. They deem this effect trivial, as "the correlation of K with cooperation therefore seems to follow directly from common sense, without the need for any special theory" (p. 46).[2] Mind you, the sure-thing principle, which demands defection, is also common sense once you know how to subtract ($T - R$ and $P - S > 0$). So which common sense is the right one? One cannot say that the need for defection is self-evident and in the same breath say that cooperation, obviously, becomes easier as Rapoport's K index gets larger.

Game On!

Kerr (this issue) favors a Social-Norm/Tempered by Self-Interest model to explain dilemma behavior. We refer to this approach here as the ST model. Its basic idea is that people are motivated and influenced by a host of factors. Van Lange and Van Doesum (this issue) strike a similar note when proposing a theoretical "package" approach. In the thin air of epistemic reasoning, we would rather go with parsimony (see earlier) and the idea that it is easier to explain complex phenomena with simple mechanisms than it is to explain simple phenomena with complex mechanisms (see target article). Down at the terranean level of theory comparison, we accept Kerr's invitation to the ballpark. Let's play!

Kerr (this issue) finds that both, social projection and the ST model, can account for behavior in the prisoner's dilemma, the nice-versus-nasty effect, and

[2] Colman has also suggested that coordination games can be solved by common sense and team reasoning (but not by social preference). Once common sense is invoked, how can team reasoning distinguish itself?

last-minute intrigue. They get 1 point for each.[3] Social projection can account for behavior in coordination games, thus gaining another point, whereas the ST does not. The score is now 4:2. If we stopped here, social projection would carry the day, but that would be misleading, writes Kerr, because we also have to see how the two theories handle other games. We agree. Kerr then considers the dictator game, DG, the trust game, TG, from the perspective of the trustor and from the perspective of the trustee, and the ultimatum game, UG, from the perspective of the proposer and from the perspective of the responder. After nine innings, his final count is 2:8 in favor of the ST.[4]

Here comes the rematch. Although we agree with Kerr's scoring for the first three innings (prisoner's dilemma, nice–nasty, intrigue), we think he is being too generous to the ST in the coordination game. Recalling Colman's (2003) point that no symmetrical payoff transformation can solve a coordination game, we subtract a point. Next, Kerr asserts that the ST can explain choices in the DG (+1), while social projection cannot (–1). We too believed the first assertion until we did the math. It turns out that no combination of weights for self-regard, other-regard (benevolence), and fairness can explain that which is most common in the DG, namely, transfers from the dictator that are larger than 0 and smaller than 50%. When we take the derivative of van Lange's (1999) utility function, we find that players always do best by either giving nothing or 50% (Krueger, Massey, & DiDonato, 2008). What is needed is a different kind of syncretist model, such as the one we *posthocked* by combining social preferences with a strategic (i.e., not entirely sincere) respect for social norms (Krueger et al., 2008). Born of desperation, this model is not parsimonious. It is an exercise in model fitting, which will have to do until a stronger a priori model comes along. In the case of the DG, social projection is not that type of model. It does not even apply to this game because there is no interdependence that could breed uncertainty. We thus subtract a point from the score of the ST. We neither add nor subtract a point with regard to social projection, because this model is designed to deal with situations of strategic uncertainty, which the DG does not provide.

As for the TG, we begin with the trustee. Again, Kerr (this issue) scores a point for the ST and subtracts one from social projection. This is no surprise, because the trustee plays a DG. Hence, our analysis of the DG applies here, too, and we subtract a point from the score of the ST while doing nothing to the projection score. The role of the trustor is most difficult because she needs to predict what the trustee will do. If the trustee were expected to be rational in van Lange's sense, the trustor could expect a return of all or nothing, but trustors seem to know that trustees like partial transfer. Trustors also make partial investments, and Kerr thinks that social projection has some success accounting for it (+.5), whereas the ST has great success (+1). We agree that social projection is of some value here. Trustors' expectations of reciprocity are positively correlated with the size of their own investments—as they should be. Trustors may attempt to solve the dilemma by mental simulation (Krueger, 2012b), asking themselves what they would do if they were in the role of the trustee, and then act accordingly. We disagree, however, with the suggestion that the ST "easily accounts for this result" (p. 78). We think that (a) *post hoc* fitting of partial investments is just as impossible as is the *post hoc* fitting of partial returns and that (b) any attempt to do so is further marred by the fact that the ST ignores that which is of paramount importance to any sane trustor, namely, the probability of reciprocity. Hence, we subtract a point.

This leaves the UG. Its structural simplicity notwithstanding, this game has led to much head scratching in game-theoretic circles. Why do proposers offer more than the minimal amount and why do responders reject nonminimal offers? Kerr (this issue) suggests that social projection may be of some benefit, and we agree. We all give +.5 points on the assumption the people are able and willing to simulate the minds of others who perform different roles (Bazinger & Kühberger, 2012; van Boven & Loewenstein, 2005). Kerr also suggests that the ST does very well, but we disagree again. If dictators cannot apply the social preference calculus to motivate partial transfers, proposers, who also have to be mindful of the responder's values and expectations, cannot do it either, *a fortiori*. We must subtract another point from the ST score.

The responder in the UG has it easier. There is no uncertainty, only the need to choose between acceptance and rejection. Kerr (this issue) subtracts a point from social projection, but we think it is fair to say that our model does not apply here, for the same reason that it does not apply to other games that do not feature strategic uncertainty. Kerr adds a point for the ST, but we again think that he is flirting with model fitting. One is left concluding that responders who reject a generous but less than 50/50 offer must have a strong preference for equality and low self-regard. In our view, no points should be added to or subtracted from the ST score.

Having played another nine innings with a new umpire, we finish with a score of 5:–1.5 favoring the social projection model. The score differential might be even more extreme. As we noted, and Kerr does not seem to object, social projection can also account for

[3] Before moving through the innings, we already suspect that the ST will beat social projection because the latter is included in the former. If social projection is excluded, the ST looks very similar to van Lange's social value model, which already includes self-interest.

[4] Kerr adds and subtracts full and half points along the way, so we do the same, although it is a practice that baseballers might frown upon.

cooperation in the game of chicken and in the assurance game (a.k.a., the stag hunt). The ST can deal with the latter fairly well but has trouble with the former. The game of chicken is particularly tricky because the successful player figures out what the opponent will do, and then does the opposite. The social projection model can help here (see target article).

We tried to play this ball game using our best approximation to what we think Kerr means when referring of the ST. He clearly favors a version of van Lange's (1999) integrative social value model. Whichever version of this model one prefers, it says little about psychological processes beyond the idea that people translate objective payoffs into subjective ones that reflect their values. The model is expressed as a weighted utility function. There is, however, another way to think about social preferences. One could embed them in a sequential process of decision making. Suppose people care not only the most about fairness but about fairness *first*. Now one can explain cooperation by saying that they first ask under which conditions payoffs are equal. They find that this is the case under mutual cooperation and mutual defection. They then ask under which condition they personally fare better, and they find that this is the case under cooperation. Hence, they cooperate. Notice that this works well as a *post hoc* fit, but its scientific worth will depend on evidence that corroborates the psychological process. For the TG, we have begun to explore the usefulness of sequential decision-making models (Evans & Krueger, 2011).

Examining the properties of competing models with the baseball metaphor is fun and enlightening. It helps clarify theoretical positions and underlying assumptions. The unbiased reader (and the biased one, too, for that matter) may now have a better understanding of how future research might be designed to obtain useful empirical or mathematically simulated results that help us refine (or reject) the models.

Unexplained (But Not Inexplicable) Findings

A theory is strong inasmuch as it can account for the same type of behavior in a variety of contexts. By our count, the social projection model fares better than the hybrid ST model. A strong theory is also able to account for different types of phenomena within the same situation. Yamagishi (this issue) suggests that by this measure his theory of moderated mutual aid and reciprocity fares better than the social projection model.

Yamagishi's theory posits that thanks to cultural and personal learning, humans have come to expect that individuals who belong to the same social group will help one another and reciprocate favors. If helping one another is the key ingredient, Yamagishi's theory may be expressed in terms of a social preference for benevolence; if reciprocation is the key, then his theory is a model of conditional reciprocity or it reduces to the social projection model (where people expect reciprocity before they themselves have acted). Either way, it is difficult to ascertain the unique contribution of this perspective.

Yamagishi is clearer about what he perceives to be the shortfalls of the social projection model. He asserts that social projection cannot explain Kelley and Stahelski's (1970) finding that defectors project more strongly than cooperators do. In response, we note that our model is designed to account for prechoice projection under conditions of high uncertainty, not postchoice projection under conditions of repeated learning. Incidentally, the postchoice projection differential has disappeared in many studies conducted after Kelley and Stahelski's work (Dawes, McTavish, & Shacklee, 1977; Krueger & Acevedo, 2007; Messé & Sviacek, 1979).

Yamagishi reports that his studies document that people expect ingroup members to help one another. Again, if this finding refers to a differential in the expectation of reciprocity, p_r, we are back to projection, and indeed, the moderating role of social categorization is well documented. People project strongly to ingroup members and barely to outgroup members (Robbins & Krueger, 2005). If, however, the attitude attributed to the ingroup is one of unconditional helpfulness (i.e., a high p_c), we are begging the question of why p_c would be higher for the ingroup than for the outgoup, which takes us back to differential projection.

To break this impasse, Yamagishi (this issue) reports the results of an experiment to show that social projection cannot explain mutual aid (cf. Kropotkin, 1902). Participants were paired with other ingroup members who presumably either did or did not know that the focal participant belonged to the same group. The critical finding was that the "participant did not expect a high level of cooperation from an in-group member who did not know that the participant was a member of the same group" (p. 70) and hence did not cooperate very much. In other words, "in-groups favored cooperation only when the group membership was common knowledge" (p. 70). This, however, is exactly what the projection hypothesis predicts. Having one participant with the relevant social categorization knowledge and another one without it breaks the symmetry necessary for projection. In a situation of strategic interdependence, projection to another ingroup member makes sense only if that person is in the same strategic situation, that is, if that person also knows that she is playing with another ingroup member, who knows that she is playing with another ingroup member, and so on. Common knowledge is essential (Nozick, 2001).

Kerr makes a similar suggestion when noting that shared-fate manipulations increase cooperation within a group, and he assumes that an increased sense of ingroup identity is the mediating mechanism. We think that such manipulations eliminate asymmetries in the perspectives among ingroup members and thereby facilitate projection. A stronger ingroup identity and ingroup favoritism are then outcomes rather than mediating mechanisms (DiDonato, Ullrich, & Krueger, 2011; see also Gaertner & Insko, 2000; Rabbie & Horwitz, 1969).

Yamagishi then suggests that his model, but not the social projection model, can account for the high correlation between trust/distrust in the TG and cooperation/defection in the prisoner's dilemma. He reasons that if trust is based on the expectation of reciprocity, so is cooperation. Indeed, the social projection model applies to both games, although its application is less straightforward in the TG than in the prisoner's dilemma (as discussed in Inning 6).[5]

Kerr (this issue) points out that social projection cannot explain why some individuals continue to cooperate "in the face of evidence that most others defect" (p. 82). This is true. We even find that some cooperate with a computer they know is programmed to defect (Acevedo & Krueger, 2005). Must we rhetorically ask whether these individuals feel benevolent toward the hardware? Perhaps there is a layer of pathological altruism that is not reducible to thoughts or feelings that make sense (Krueger, 2011). Kerr counts studies that show an increase in cooperation after the priming of social values as unique evidence for morality theories. This is fair enough. We are not claiming that other theories can explain nothing.

Other Alleged and Real Limitations of the Social Projection Model

An overriding concern of Kerr's is that the success of the social projection model is somewhat chimerical because it is purchased with a cheap strategy of knocking down strawmen. Yamagishi feels similarly. We do not agree with the suggestion that we erected strawman versions of the other theories. We tried to distill from each theory its essential and unique claims regarding cooperation in the one-shot anonymous prisoner's dilemma. Without this reduction, some of the competing theories would be too flexible to be tested. The ST, for example, has an unstated number of free parameters. We took some parameters, such as van Lange's weight for benevolence, and asked how large their values would have to be before cooperation dominates defection. We found that it would have to be very large indeed.

A related concern—a chicken coming home to roost—is the suggestion that the social projection model is the one that is guilty of capitalizing on a free parameter. Indeed, we do not have a good fix on the provenance of the individual differences in social projection. We know that projection scores vary widely, but we know little about why this variation is as large as it is and where it is coming from (Krueger & Stanke, 2001). Currently, the task of explaining these differences lies beyond the model. We can even ask if a search for an explanation is necessary. Research on social preferences and values typically settles for the assessment of these preferences, without asking how they arise. Yamagishi concludes that our model would be stronger if everyone projected equally, ideally with the optimal Bayesian value of $p_r = .67$. We find this hypothetical situation bizarre. Why should there not be individual differences in social projection when there are individual differences in everything else? Incidentally, if there were no individual differences, the model's fit with the empirical data would be worse, not better. Recall that our model can explain that some people cooperate even if the average projection score suggests defection.

Questions also arose with regard to our last-minute-intrigue paradigm. We created this design to test a set of predictions that only the social projection model could make. Necessarily, there was more overlap among the predictions of other theories. If there is discontent with this approach, we invite proponents of other theories to design a paradigm in which the predictions of their models are unique relative to all others.

Alas, the intrigue design has some features that, at first glance, seem undesirable. Yamagishi (this issue) points out that there is no direct evidence for social projection in our data. That is exactly right, and we wanted it to be so. We already have evidence for the hypothesis that measured individual differences in projection predict cooperation/defection (Acevedo & Krueger, 2004; Krueger & Acevedo, 2007, 2008; Quattrone & Tversky, 1984).[6] The objective of the intrigue experiments (at least the first one) was to control the strength of social projection, not to measure it. For the same reason, we opted to put participants in the role of advisor. In that role, they could contemplate the situation of another person choosing between cooperation and defection; they did not need to bring their own precommitments to favor one or the other strategy into this contemplation.

From Rationality to Adaptiveness

Yamagishi's (this issue) main point of critique is that high projectors will cooperate too much and will

[5] We should also note that a significant correlation by itself is weak evidence for a common cause. The common cause may be so remote that it is of little interest as a mechanism. For example, the median age of people in the developed world has been correlated with a rise in average global temperatures and the price of milk. What is the common mechanism? In contrast, the social projection model specifies proximal mechanism.

[6] This is another example for the desirability of individual differences. Without them, these correlations would be undefined.

therefore disappear from the evolutionary scene. Contrary to this suggestion, high projectors and cooperators seem to be all around us. Nonetheless, Yamagishi's argument deserves examination. The thought experiment he presents shows that the cooperative high projectors will be exploited by defecting low projectors over repeated rounds of the game. If payoffs translate into units of fitness, the high projectors will end up starving and not reproducing. This is true, but this dismal picture is realized only if the high projectors refuse to learn from experience. As the social projection model is essentially Bayesian, it assumes that individuals learn. Certainly, repeated defection will be noticed and it will become part of the evidentiary calculus, and will thus lead to a higher probability of self-protective defection. Although the specific claims of our model refer to Round 1, we now explore a few scenarios of repeated play by way of analysis and simulation.[7]

Scenario 1: Evolution Without Reproduction

In this first simulation, we study the rate at which cooperators and defectors die. To simplify, we assume that players who have decided to cooperate, given the nature of the game (K) and their own level of projection, will do so consistently, and likewise for defectors. We also assume that in a game of evolution, the probability of survival is proportional to the payoff in the game. These assumptions alone are sufficient to give defectors a survival advantage; but with math we can be more precise. In our simulation, the payoff T (for unilateral defection) guarantees survival, and the payoff S (unilateral cooperation), being 0, spells death. The probabilities of survival after receiving the payoffs R (mutual cooperation) and P (mutual defection) respectively are R/T and P/T.

In our simulation, surviving players are paired randomly with other survivors. The survival rate for cooperators is $p_c \frac{R}{T} + (1 - p_c)\frac{S}{T} = p_c \frac{R}{T}$, and for defectors it is $p_c + (1 - p_c)\frac{P}{T}$. The value of p_c depends on the payoffs of the game and the level of social projection. Regardless of the value of p_c, defectors survive longer because $T > R$ and $P > S$. As p_c grows, the discrepancy between the survival rates grows, with defectors dying ever more slowly than cooperators. Figure 1 shows this result for a nasty game (top panel; $K = .2$) and a nice game (bottom panel; $K = .8$; $\mu_r = .6\bar{6}$ and $\sigma_r = .2$ for both cases). A more detailed description of the mathematics is presented in Appendix A.

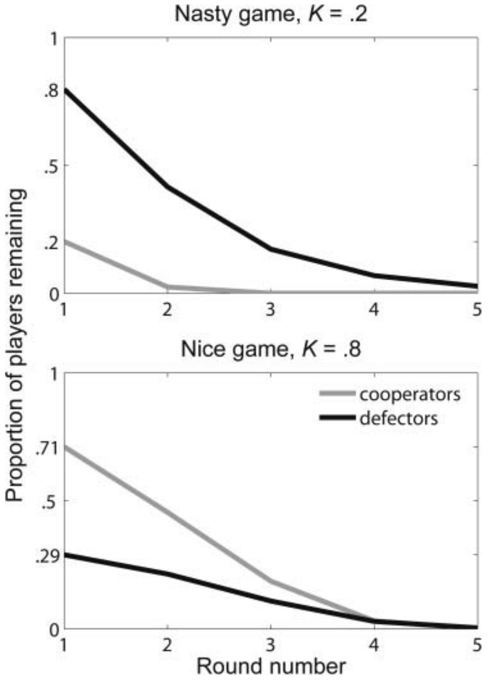

Figure 1. Evolution without reproduction. Cooperators (gray) and defectors (black) survival decline in the nasty game (top panel) and nice game (bottom panel).

Scenario 2: Evolution With Reproduction

This scenario is similar to the first one ($\mu_r = .6\bar{6}$ and $\sigma_r = .2$), but not only do some individuals die, others are also born so that the size of the population remains constant. As the defectors' survival rate is higher than the cooperators' rate, the question is how fast the population will change. After every round of the game, we replenish the population with cooperators at the rate of p_c and defectors at the rate of $1 - p_c$ from the last round. Figure 2 shows how much more slowly cooperators lose ground in nice as compared with nasty game. A more detailed description of the mathematics is presented in Appendix B.

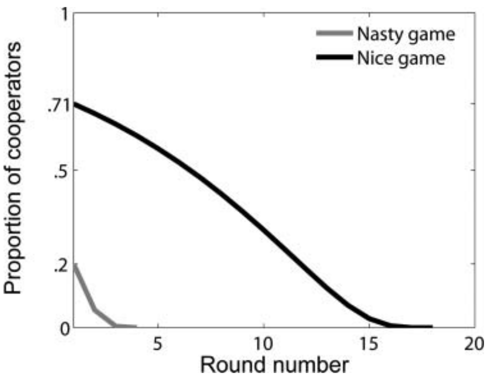

Figure 2. Evolution with reproduction. The proportion of cooperators in the nice (gray) and nasty (black) games.

[7] High projectors do well if group selection is allowed. If intergroup dilemmas are superimposed on interpersonal dilemmas, as in war, the group that mobilizes the most cooperators will prevail over groups whose members defect from one another (Krueger, 2007).

Together, the first two simulations corroborate Yamagishi's assertion that if cooperation is based on social projection, cooperators—and thus projectors—will lose evolutionary ground. But why would they not learn? The following scenarios are more realistic because they allow learning, either by a change of strategy after Round 1 (Scenario 3) or by continued Bayesian updating contingent on reinforcement (Scenarios 4 and 5).

Scenario 3: Social Projection Plus Tit-For-Tat or PAVLOV

An individual playing tit-for-tat does whatever the other player did in the previous round. The proportion of cooperators is initially given by p_c. Regardless of what Player A chooses, the probability with which she is paired with a cooperator is p_c. In the next round, she will therefore cooperate with that probability. Because all pairings and rounds are independent, the probability of cooperation remains p_c in all successive rounds. The value of social projection is clear. With a reciprocating strategy like tit-for-tat, once there is evidence of other players' choice, strong social projection sets a higher plateau of cooperation than weak projection.

An individual playing PAVLOV stays with her own choice after winning (i.e., after receiving one of the high payoffs, T or R) and shifts to the other strategy after losing (i.e., after receiving one of the low payoffs, P or S). Stated differently, a player cooperates after R or P and defects after T or S. In the long run, cooperation and defection have the same probability of .5. Perhaps surprisingly, initial settings of projection, game difficulty, and hence initial rates of cooperation only have short-term effects. In this sense, PAVLOV is a great equalizer. A more detailed description of the mathematics is presented in Appendix C.

Scenario 4: A Single Bayesian Learner

Like Fischer (2009, this issue), we contend that the probability that two players make the same choice can and should be learned. After this value—his p_s and our p_r —is learned, our models make the same predictions. The main difference lies in what they say about cooperation in the one-shot prisoner's dilemma or *on the first round* of a repeated game. In the SERS model, the player brings a value of p_s to the game, but the model is silent about the value's origins. In contrast, our model is explicitly Bayesian, and it uses an idealized state of ignorance (i.e., a uniform distribution of the prior probabilities) as its starting point. With the psychological mechanism of social projection, Bayesian belief revision yields probability estimates that are greater than .5.

Chater, Vlaev, and Grinberg (2008; see also Chater & Vlaev, this issue) also invoke learning as a critical process enabling cooperation in social dilemmas. In their reinforcement model, players learn the values of cooperation and defection, which arise from combinations of both payoffs and the probabilities of obtaining those payoffs (see target article for details).

In light of our agreement with Fischer and Chater regarding the relevance of learning, we now sketch a reinforcement learning model that is compatible with Bayesian learning in repeated dilemmas. We note at the outset that the model is deterministic; it can be amended with a random error term to acknowledge imperfections in learning performance.

The model. Like Fischer, but unlike Chater et al., we assume that the probability of reciprocity (making the same choice) is the object of learning. The payoffs are available from the description of the game. The most influential reinforcement model to date is the one proposed by Rescorla and Wagner (1972); it describes how organisms learn associations between their choice alternatives and outcomes. However, to show how individuals learn the probability of reciprocity, or making the same strategic choice, we turn to an earlier model introduced by Bush and Mosteller (1955). Here, the probability of choosing an alternative is updated in light of the outcomes of the trial (round number in the prisoner's dilemma game):

$$p_r^{(n)} = p_r^{(n-1)} + \alpha \left(\lambda - p_r^{(n-1)}\right) \tag{1}$$

The player's estimate of the probability of reciprocity on round n (the superscript) is a function of the estimate on the previous round plus the new evidence. The new evidence, λ, is the outcome on a given round of the game. In our case, λ equals 1 if the players agreed and 0 if the players did not agree. Over rounds of the game, the mean value of λ is the true probability of reciprocity, the value that the player must learn.[8] The value of α represents the speed with which the player learns the true value of reciprocity: the mean of λ. Equation 1, with the expression in parentheses, which denotes the prediction error, lies at the heart of nearly all reinforcement models today (Dayan, 2001; Sutton & Barto, 1998).

The parameters of the model. There are two parameters in the model. The first is the initial estimate of the agreement probability, $p_r^{(0)}$.This parameter determines how the player will play the first round of the game. This is where our model departs from the SERS model, because in the latter, this value is truly a free parameter—it could come from many different sources, and those sources are not specified. In our

[8] If two players always agree, we want the true value of $p_r = 1$, which means the mean value, $\bar{\lambda} = 1$. This is only accomplished if every value of $\lambda = 1$. This is why an agreement is coded as a 1 and a disagreement is coded as a 0.

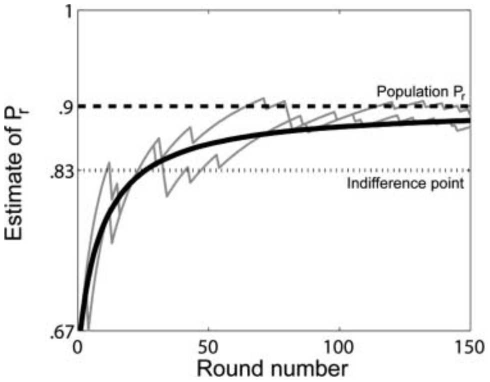

Figure 3. Reinforcement learning with Bayesian updating. The two gray lines show two example players, and the solid line shows the average player. For this example, the players start out with an estimate of the agreement probability, $p_r^{(0)} = .6\bar{6}$ (the values for the a and b are 6 and 3, respectively) and then learn the true probability of reciprocity of the population, $\bar{\lambda} = .9$ (in this simulation). This value was arbitrarily chosen. On each round of the game, the player updates its estimate of p_r according to equation 1, with α defined in equation 2. A player will cooperate when they learn a value for p_r that is greater than the indifference point. The indifference point for the nice game is shown.

model the value is specified: If no other information is given, the initial estimate comes from social projection.

The second parameter of the model is α. In most contexts, this too is a free parameter. This is where describing our model in Bayesian terms helps (see target article for details). Bayes's theorem gives the exact value for α, as long as the priors are specified. For simplicity, we assume a beta prior[9] over values of p_r, which results in a specific level for α (see Appendix D for details, and Kruschke, 2011, for a similar equation). Now,

$$\alpha = \frac{1}{a + b + n} \tag{2}$$

where $a/(a + b)$ is the initial estimate of the probability of agreement ($p_r^{(0)}$), and n is the round number. Because α depends on n, as more rounds of the game are played, the less the player adjusts the estimate based on the prediction error (i.e., as n grows, α shrinks). Two example Bayesian learners, and the average learner, are shown in Figure 3. The learners will cooperate when they learn a value for p_r that is greater than the indifference point (shown for the nice game).

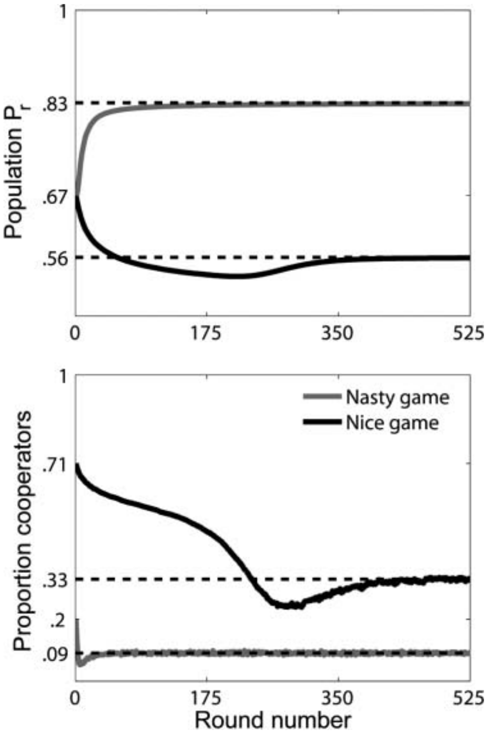

Figure 4. A population of Bayesian learners. Top panel: The average Bayesian learner in a nasty game (gray line) and nice game (black line). The population learns the indifference point as the limiting value of the probability of reciprocity $p_r^{(n \to \infty)} = p_r^0$. Bottom panel: The average probability of cooperation over games for the nasty game (gray line) and nice game (black line) are shown. The values are less than $\frac{1}{2}$ because the true probability of reciprocity is never fully learned, only continually approximated with greater precision, most of the players are consistently under the true value.

Scenario 5: A Population of Bayesian Learners

Last, we ask how a population of Bayesian learners would interact. Every player in the group is randomly paired with another player, and all update their individual estimates of p_r according to Equation 1. The speed of learning is given by Equation 2. Starting with a prior of $p_r^{(0)} \approx 2/3$ (normally distributed in the population with $\mu_r = 2/3$ and $\sigma_r = .2$), the players learn the indifference point, $p_r^{(n \to \infty)} = p_r^0$. However, they approach the indifference point from below even if they started above, so that $p_c \leq \frac{1}{2}$. A consequence of the prediction error (expression in parentheses in Equation 1) is that the players never reach the indifference point, and thus the values of p_c remain below .5 (see Figure 4).

This last simulation takes us beyond our original intent, as expressed in the target article. We now know that a group of Bayesian learners will continue to cooperate at a moderate and predictable rate, and that they remain sensitive to the difficulty of the game (K). The extended Bayesian solution also has parsimony and coherence on its side. With this solution, there

[9] A uniform distribution is a special case of the beta distribution (with $a = 1$ and $b = 1$), so Laplacian ignorance is preserved. Being the conjugate prior for the binomial, the beta prior makes analytical solutions for Equation 2 possible (see Appendix D for details).

is no need to treat one-shot and repeated games as fundamentally different, and there is no need to develop different explanatory models. With this solution, players and other humans can be allowed to be rational; they learn what they need to know to behave adaptively.

Conclusions

We have suggested that social projection can explain cooperation in social dilemmas and related situations. Compared with its competitors, our theory (and closely related theories) provides a more detailed, concrete, and plausible account of the psychological processes that support choice in a social dilemma. If, however, the demand is that all data must be explained—as some commentators seem to have hinted—then our theory has overreached. Then again, would this uncompromising standard not have to be applied to all theories? Which alternative theory can fit the empirical data without taking a bath in a pool of free parameters? We acknowledge that the present state of affairs cannot be the final one (if there ever is one). By necessity, our objective is to put the social projection model on the map and provide enough detail for others to engage in further "exploitation and exploration." To set this process in motion, we have to pursue a strategy of seeking and pointing out the theory's unique strong points.[10] We invite the proponents of the other theories to do likewise and to then compare notes again.

We close with a thought experiment. The typical social dilemma dramatizes the individual's conflict between cooperation and defection, between the personal and the collective good. The perspective of the ego is paramount in this frame. Now, suppose an attempt were made to frame the situation from an allocentric (i.e., not egocentric) perspective. Would we be as torn between wanting the other player in a prisoner's dilemma to cooperate and wanting her to defect? This is difficult to imagine. We would, presumably, want her to cooperate (with us!) no matter what. Cooperation now dominates and the sure-thing strategy is *not* in conflict with moral concerns. Only the egocentric perspective gives rise to the dilemma, and that is why it should be resolved from point of view.

Acknowledgments

We thank Andra Geana for suggestions on the simulations and mathematical appendices. She took us to a higher place so that we could see farther.

Note

Address correspondence to Joachim I. Krueger, Department of Cognitive, Linguistic and Psychological Sciences, Brown University, Box 1821, Providence, RI 02912. E-mail: Joachim@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.

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

Bazinger, C., & Kühberger, A. (2012). Is social projection based on simulation or theory? Why new methods are needed for differentiating. *New Ideas in Psychology*. Advance online publication. doi:10.1016/j.newideapsych.2012.01.002

Brehm, J. W. (1966). *Response to loss of freedom: A theory of psychological reactance*. New York, NY: Academic Press.

Bush, R. R, & Mosteller, F. (1955). *Stochastic models for learning*. New York, NY: Wiley.

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

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical transactions of the Royal Society: B, 362*, 933–942. doi:10.1098.rstb.2007.2098

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences, 26*, 139–153. doi:10.1017/S0140525×03000050

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

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., 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

Dayan, P. (2001). Reinforcement learning. In C. R. Gallistel (Eds.), *Steven's handbook of experimental psychology* (pp. 103–192). New York, NY: Wiley.

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.

Evans, A. M., & Krueger, J. I. (2011). Elements of trust: Risk taking and expectation of reciprocity. *Journal of Experimental*

[10] We follow Russell (see epigraph) in our conception of rationality, and we apply the same criterion to our research participants. Individuals engaging in evidential reasoning respect the nature of uncertainty; individuals thinking along moral lines tend not to.

Social Psychology, 47, 171–177. doi:10.1016/j.jesp.2010.08.007

Fehr, E., Fischbacher, U., & Gächter, S. (2000). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature, 13*, 1–25. doi:10.1007/s12110-002-1012-7

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

Gaertner, L., & Insko, C. A. (2000). Intergroup discrimination in the minimal group paradigm: Categorization, reciprocation, or fear? *Journal of Personality and Social Psychology, 79*, 77–94. doi:10.1037/0022-3514.79.1.77

Gauthier, D. (1986). *Morals by agreement*. Oxford, UK: Oxford University Press.

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

Haselton, M. G., Bryant, G. A., Wilke, A., Frederick, D. A., Galperin, A., Frankenhuis, W. E., & Moore, T. (2009). Adaptive rationality: An evolutionary perspective on cognitive bias. *Social Cognition, 27*, 733–763. doi:10.1521/soco.2009.27.5.733

Hogg, M. A. (2007). Uncertainty–identity theory. *Advances in Experimental Social Psychology, 39*, 69–126. doi:10.1016/S0065-2601(06)39002-8

Howard, J. V. (1988). Cooperation in the Prisoner's Dilemma. *Theory and Decision, 24*, 203–213. doi:10.1007/BF00148954

Kavanau, L. J. (1967). Behavior of captive white-footed mice. *Science, 155*, 1623–1639. doi:10.1126/science.155.3770.1623

Kelley, H. H., & Stahelski, A. J. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology, 16*, 66–91.doi:10.1037/h0029849

Kropotkin, P. A. (1902). *Mutual aid. A factor of evolution*. New York, NY: McClure, Philips.

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

Krueger, J. I. (2011). Altruism gone mad. In B. Oakley, A. Knafo, G. Madhavan, & D. S. Wilson (Eds.), *Pathological altruism* (pp. 392–402). New York, NY: Oxford University Press.

Krueger, J. I. (2012a). The (ir)rationality project in social psychology: A review and assessment. In J. I. Krueger (Ed.): *Social judgment and decision-making* (pp. 59–75). New York, NY: Psychology Press.

Krueger, J. I. (2012b). Social projection between theory and simulation. *New Ideas in Psychology*. doi:10.1016/j.newideapsych.2012.01.003

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. 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., 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., & Stanke, D. (2001). The role of self-referent and other-referent knowledge in perceptions of group characteristics. *Personality and Social Psychology Bulletin, 27*, 878–888. doi:10.1177/0146167201277010

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York, NY: Academic Press.

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers Volume 1*. Cambridge, MA: Cambridge University Press.

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

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

Meiser, T., & Hewstone, M. (2004). Cognitive processes in stereotype formation: The role of correct contingency learning for biased group judgments. *Journal of Personality and Social Psychology, 87*, 599–614. doi:10.1037/0022-3514.87.5.599

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

Norris, J. R. (1999). *Markov chains*. Cambridge, UK: Cambridge University Press.

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

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

Pitt, M. A., & Myung, J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences, 6*, 421–425. doi:10.1016/S1364-6613(02)01964-2

Popper, K. R. (1962). *Conjectures and refutations*. New York, NY: Basic Books.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of "rational" decision theory. *Proceedings of the Royal Society B, 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

Rabbie, J. M., & Horwitz, M. (1969). Arousal of ingroup-outgroup bias by a chance win or loss. *Journal of Personality and Social Psychology, 13*, 269–277. doi:10.1037/h0028284

Rescorla, R. A., & Wagner, A. R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York, NY: Appleton-Century Crofts.

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

Schaller, M., & Maass, A. (1989). Illusory correlation and social categorization: Toward an integration of motivational and cognitive factors in stereotype formation. *Journal of Personality and Social Psychology, 56*, 709–721. doi:10.1037/0022-3514.56.5.709

Sutton, R. S, & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choice. *Journal of Experimental Psychology, 71*, 680–683.

Van Boven, L., & Loewenstein, G. (2005). Cross-situational projection. In M. D. Alicke, D. Dunning, & J. I. Krueger (2005). *The self in social judgment* (pp. 43–64). New York, NY: Psychology Press.

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

Waller, B. N. (2011). *Against moral responsibility*. Cambridge, MA: MIT Press.

Weber, M. (1904–1905). Die protestantische Ethik und der "Geist" des Kapitalismus [The Protestant ethic and the spirit of capitalism]. *Archiv für Sozialwissenschaft und Sozialpolitik, 20*, 1–54, *21*, 1–110.

Appendix A

The probability of survival is proportional to the payoff obtained.

For cooperators

$$s_c = p_c \left(\frac{R}{T}\right) + (1 - p_c)\left(\frac{S}{T}\right) \tag{A.1}$$

and for defectors

$$s_d = p_c \left(\frac{T}{T}\right) + (1 - p_c)\left(\frac{P}{T}\right) \tag{A.2}$$

The overall probability that a player survives, regardless of own choice in the game is

$$s_a = p_c s_c + (1 - p_c)\, s_d \tag{A.3}$$

The proportion of cooperators on round 1 is $p_c^{(1)}$ and depends on the population of p_r (see target article). The probability that a player is both a cooperator *and* survives Round 1 is $p_c^{(1)} s_c^{(1)}$.

It is with this probability that a cooperator enters Round 2. Because the proportion of cooperators changes, the survival probability for cooperators also changes. There is a *recursive* relationship between the proportion of cooperators on round n based on the previous round

$$p_c^{(n+1)} = p_c^{(n)} s_c^{(n)} \tag{A.4}$$

For defection, the relationship is

$$p_d^{(n+1)} = p_d^{(n)} s_d^{(n)} \tag{A.5}$$

On Round 1, $p_c^{(1)} + p_d^{(1)} = 1$. The total number of players shrinks over the course of the game, so that this relationship generally does not hold for $n > 1$. Instead,

$$\frac{p_c^{(n)} s_c^{(n)}}{s_a^{(n)}} + \frac{\left(1 - p_c^{(n)}\right) s_d^{(n)}}{s_a^{(n)}} = 1 \tag{A.6}$$

Appendix B

The probability of survival is proportional to the payoff obtained. As in the text, these probabilities are weighted by their probability that the other player will cooperate (p_c) or defect ($1 - p_c$). The probability of survival differs for cooperators

$$s_c = p_c \left(\frac{R}{T}\right) + (1 - p_c)\left(\frac{S}{T}\right) \tag{B.1}$$

and for defectors

$$s_d = p_c \left(\frac{T}{T}\right) + (1 - p_c)\left(\frac{P}{T}\right) \tag{B.2}$$

The overall probability that a player survives, regardless of own choice in the game is

$$s_a = p_c s_c + (1 - p_c)\, s_d \tag{B.3}$$

The proportion of cooperators on Round 1 is $p_c^{(1)}$ and depends on the population of p_r (see target article). At this point, one of three things can happen. A cooperator can survive. The probability that a player is both a cooperator *and* survives is

$$P\,(\text{C survives}) = p_c^{(n)} s_c^{(n)} \tag{B.4}$$

A cooperator can die and be replaced by another cooperator. The probability that a player is a cooperator *and* dies *and* is replaced by a cooperator is

$$P\,(\text{C replaces C}) = p_c^{(n)} \left(1 - s_c^{(n)}\right) \left(\frac{p_c^{(n)} s_c^{(n)}}{s_a^{(n)}}\right) \tag{B.5}$$

Last, a defector can die, and be replaced by a cooperator. The probability that a player is a defector *and* dies *and* is replaced by a cooperator is

$$P\,(\text{C replaces D}) = \left(1 - p_c^{(n)}\right) \left(1 - s_d^{(n)}\right) \left(\frac{p_c^{(n)} s_c^{(n)}}{s_a^{(n)}}\right) \tag{B.6}$$

Any one of these outcomes leads a cooperator on the next round. We can write the recursive relationship as the sum of the probabilities these three outcomes

$$\begin{aligned} p_c^{(n+1)} &= P(\text{C survives}) + P(\text{C replaces C}) \\ &\quad + P(\text{C replaces D}) \end{aligned} \tag{B.7}$$

Because the number of players is constant, the probability of a defector on any round of the game is $1 - p_c^{(n)}$.

Appendix C

Initial State Vector

The initial probability that a player cooperates on Round 1 can be represented as a probability vector

$$U = \begin{bmatrix} p_c & 1 - p_c \end{bmatrix} \tag{C.1}$$

Transition Matrix

A player will either make the same choice on the next round, or will change to the other choice on the next round. The probability with which a player stays or switches depends on the other player's choice (see text for commentary on this), and is given by the transition matrix

$$P = \begin{bmatrix} p_c & 1 - p_c \\ 1 - p_c & p_c \end{bmatrix} \qquad \text{(C.2)}$$

where $P_{i,j}$ is the probability of transitioning from choice i to choice j ($i, j = 1$for cooperation and 2 for defection). For example, the probability that a player transitions from defection ($i = 2$) to cooperation ($j = 1$) is $P_{2,1} = 1 - p_c$, the lower left value in the transition matrix.

On round n, a player will be in a state depending on where it started (initial state vector), and how it transitioned from one round to the next (the transition matrix)

$$u^{(n)} = U P^n \qquad \text{(C.3)}$$

Limiting Behavior

After n rounds, the transition matrix has been applied n times, and it is P^n. For a two state transition matrix like the one above (both its rows and columns sum to 1), the limiting behavior, as $n \to \infty$, is known: the probabilities in the matrix spread out evenly over the possible transitions (Norris, 1999).

$$\lim_{n \to \infty} P^n = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \qquad \text{(C.4)}$$

That is, after n rounds of the game, regardless of where the players started, the same number of players stay as switch.

Appendix D

Bayesian Updating

We assume each player has a beta prior $P\,(H_x)$ over the possible values of p_r

$$P\,(p_r = x) = \frac{x^{(a-1)}\,(1 - x)^{(b-1)}}{\mathrm{B}\,(a, b),} \qquad \text{(D.1)}$$

where B (a, b) is the Beta function, defined by

$$\mathrm{B}\,(a, b) = \int_0^1 x^{(a-1)}\,(1 - x)^{(b-1)}\,dx. \qquad \text{(D.2)}$$

The data, D, is a series of agreements or disagreements over rounds of the game

$$D^{(n)} = \begin{cases} 1 & if agreement \\ 0 & otherwise. \end{cases} \qquad \text{(D.3)}$$

That is, each round of the game is a Bernoulli trial with probability of agreement $\bar{\lambda}$. The likelihood function is given by the binomial distribution conditioned on the current hypothesis H_x

$$\mathrm{P}\,(D|H_x = p_r) = \binom{n}{k} p_r^k\,(1 - p_r)^{n-k}\,, \qquad \text{(D.4)}$$

where k is the number of agreements in n rounds.

Bayes's Theorem gives the posterior distribution, the likelihood of the hypotheses given the data

$$\mathrm{P}\,(H_x|D) = \frac{\mathrm{P}\,(D|H_x = p_r)\,P\,(H_x)}{\int_0^1 \mathrm{P}\,(D|H_x = p_r)\,P\,(H_x)\,dx.} \qquad \text{(D.5)}$$

The beta distribution is the conjugate prior of the binomial distribution. A conjugate prior is a prior distribution which results in a posterior distribution of the same form. The posterior distribution, then, also takes on a beta distribution

$$\mathrm{P}\,(H_x|D) = \frac{x^{(a+k-1)}\,(1 - x)^{(n-k+b-1)}}{\mathrm{B}\,(a + k, n - k + b).} \qquad \text{(D.6)}$$

But the parameters $a,$ and b, have changed to $a + k$ and $n - k + b$.

Bayesian Point Estimation

We assume that players act as if they are Bayesian point estimators, each player acts according to the mean of the posterior distribution. The mean of the beta distribution with parameters a and b is $\frac{a}{a+b}$, so the prior estimate before playing the first round of the game is

$$p_r^{(0)} = \frac{a}{a + b.} \qquad \text{(D.7)}$$

On any round of the game, the new estimate is

$$p_r^{(n)} = \frac{a + k^{(n)}}{a + b + n}, \qquad \text{(D.8)}$$

where $k^{(n)}$ is the number of successes up to the n^{th} round. The final task is to write $p_r^{(n)}$ as a function of $p_r^{(n-1)}$. Recall that λ is the binary outcome so that $k^{(n)} = k^{(n-1)} + \lambda$

$$p_r^{(n)} = \left(a + k^{(n-1)}\right) \frac{1}{a + b + n} + \lambda \frac{1}{a + b + n.} \qquad \text{(D.9)}$$

We can rewrite $a + k^{(n-1)}$ as a function of $p_r^{(n-1)}$

$$a + k^{(n-1)} = (a + b + n - 1)\, p_r^{(n-1),} \tag{D.10}$$

so that

$$p_r^{(n)} = p_r^{(n-1)} \frac{a + b + n - 1}{a + b + n} + \lambda \frac{1}{a + b + n.} \tag{D.11}$$

And notice that

$$\frac{a + b + n - 1}{a + b + n} = 1 - \left[\frac{1}{a + b + n} \right]. \tag{D.12}$$

We can let $\alpha = \frac{1}{a+b+n}$. This is the learning rate for our Bayesian learner. And we substitute this into equation D.11

$$p_r^{(n)} = p_r^{(n-1)} (1 - \alpha) + \lambda\alpha. \tag{D.13}$$

This can be rewritten into the Bush-Mosteller equation (Equation 1 in the text)

$$p_r^{(n)} = p_r^{(n-1)} + \alpha \left(\lambda - p_r^{(n-1)} \right). \tag{D.14}$$

An Extra Note

In the target article, we simplify things by assuming that a player's p_r comes from a normal distribution with a population mean and variance. This is what allows us to obtain the probability of cooperation, p_c. This is a simplification because our Bayesian model is that a player has a full prior distribution $P(H_x)$. But we assume that players act according to the mean of their posterior distribution. When we say that a player's p_r comes from a normal distribution with a population mean and variance, we mean that the means of the priors for the players in the population are normally distributed with some variance.