Skeptical Reciprocity and Principled Defection:
On Social Roots of Non-Cooperative Behavior

Johannes Müller-Trede and Yuval Rottenstreich

Rady School of Management
University of California, San Diego
Abstract

Cooperating with others often entails some sacrifice of one’s own material interests. Many theories of reciprocity and cooperation thus assert a social-material tradeoff. In these theories, people reciprocate cooperation when they prioritize how they treat others and how others treat them. They do not reciprocate cooperation when they prioritize their own material well-being. We challenge this perspective by presenting and testing a game-theoretic model in which people frequently do not reciprocate cooperation precisely for social reasons. The model is attribution-based and taps people’s skepticism about each other’s motives. To illustrate, consider first-move cooperation in a sequential, one-shot prisoners’ dilemma. This behavior could be attributed to kindness; the first-mover caring about the second-mover. But it could also be attributed to tactical self-interest; the first-mover attempting to elicit the second-mover’s reciprocal cooperation because it is materially profitable. Individuals we term “skeptical reciprocators” do not reciprocate cooperation they attribute to tactics rather than kindness. Even when they prioritize social concerns, good treatment alone does not prompt their positive reciprocity. They ask why they received good treatment, and absent a strong signal of kindness, do not reciprocate. If skeptical reciprocity is prevalent, the fundamental impediment to mutual cooperation in sequential prisoners’ dilemmas is not that non-cooperation is materially payoff dominant. It is that first-movers cannot prove their kindness. More generally, much non-cooperation may arise not because people are self-interested, but because they worry others are. Two experiments highlight an upside of skeptical reciprocity: when first-movers can convincingly signal kindness, reciprocity rates are extremely high.

Keywords: Reciprocity, Cooperation, Attribution, Prisoners’ Dilemma, Self-Interest, Social Preferences
Cooperating with others often entails foregoing the maximization of one’s own material interests. Nevertheless, cooperation abounds (Rapaport & Chammah, 1965; Nowak, 2006). To explain this fundamental aspect of social life, researchers dating back to Kelley and Thibaut (1978) have proposed theories of psychological payoffs. In these theories, people act cooperatively whenever the psychological payoff of doing so compensates for the material reward they are forsaking.

Reciprocity is an important source of psychological payoffs (Pruitt, 1968; Campbell, 1975; Komorita, Parks, & Hulbert, 1992; Kiyonari, Tanida, & Yamagishi, 2000; Keysar, Converse, Wang, & Epley, 2008; Gray, Ward, & Norton, 2014; Malmendier, te Velde, & Weber, 2014; Yamagishi, Li, Takagishi, Matsumoto, & Kiyonari, 2014). According to the influential theory of “strong reciprocity,” people simply like to respond to kindness with kindness and to unkindness with unkindness (Fehr, Fischbacher, & Gächter, 2002; Gintis 2000). For example, in the one-shot, sequential prisoners’ dilemma of Figure 1, psychological payoffs may tempt the second-mover to cooperate if the first-mover has cooperated and to defect if the first-mover has defected. In turn, a first-mover who anticipates reciprocity may cooperate in the first place. A mutually cooperative outcome will then obtain if the second-mover’s psychological payoff from reciprocating cooperation outweighs the material reward he or she forfeits by not defecting.

In this article, we are interested in the determinants of reciprocity. But our point of departure is a common example of non-reciprocating behavior: a second-mover’s defection in response to a first-mover’s cooperation. In many studies of one-shot, sequential games, between a third and two-thirds of second-movers defect on cooperation (Clark & Sefton, 2001; Hayashi, Ostrom, Walker, & Yamagishi, 1999; Kiyonari, Tanida, & Yamagishi, 2000; Tversky & Shafir, 1992).

Theories of strong reciprocity explain defection on cooperation straightforwardly. They suggest that the second-movers in question place pronounced weight on their own material rewards and little weight on the psychological payoff of reciprocation. That is, theories of strong reciprocity assume a material-social tradeoff and paint the defectors as relatively greedy, as focusing on material profit over social considerations (Fehr & Gintis, 2007).
We challenge this perspective. We suggest that second-movers who defect on a cooperative first-mover often do so precisely for social reasons. We call defection that is socially-driven “principled defection” in order to distinguish it from defection driven by mere self-interest. The intuition we rely on is that a second-mover who faces cooperation will think about why the first-mover cooperated. The first-mover may have cooperated out of genuine kindness, that is, because he cares for the interests of the second-mover. On the other hand, if he anticipated possible reciprocity, the first-mover may have cooperated out of tactical self-interest, that is, out of what is often termed a “weak reciprocity” motive (see, e.g., Guala, 2012). After all, defecting almost certainly consigns a first-mover to a material outcome of $2, but by cooperating he has some chance at $4 (at the risk of $0). Second-movers whom we term “skeptical reciprocators” opt not to reciprocate cooperation that might be attributable to ulterior motives.

Whereas strong reciprocity theories do not distinguish between different motivations for good treatment, skeptical reciprocators ask why they received good treatment. When the good treatment they receive may be a by-product of another player’s calculated self-interest, they do not feel that it merits reciprocity. Only when the good treatment they receive can be attributed to another player’s good-heartedness do they feel that it merits reciprocity. To skeptical reciprocators, a first-mover taking a calculated chance at maximizing his own material payoff is different from a first-mover acting out of true niceness. They feel a bond or social obligation only with the latter. As a result, their defection can be principled: A skeptical reciprocator may be extremely other-regarding, in the sense of caring a great deal about how others treat her and how she treats others, yet defect on cooperation that she believes may not be genuine.

If skeptical reciprocity is prevalent, the primary impediment to mutual cooperation in sequential prisoner’s dilemmas is arguably not that defection is materially payoff dominant (cf. Tyler 2013; Tyler & Blader, 2001, 2003). It is that a first mover cannot send a credible signal of genuineness. More generally, across many situations much non-cooperative behavior may arise not because people are self-regarding, but because they worry others are, and they believe they lack the evidence to conclude otherwise (see also Miller, 1999; Heath, 1999; Fetchenhauer & Dunning, 2010; Markle, 2011; Saito, 2015; Vohs,
SKEPTICAL RECIPROCITY AND PRINCIPLED DEFECTION

Baumeister, & Chin 2007). In other words, much non-cooperative behavior may be driven by social considerations rather than the pursuit of material reward.

We next outline a game-theoretic model of skeptical reciprocity that includes strong reciprocity as a special case. Then, using the model, we devise two variants of the prisoners’ dilemma that allow us to test for manifestations of skeptical reciprocity. In each variant, first-move cooperators can effectively signal genuine kindness. Relative to a standard sequential game, these variants should therefore yield an increased rate of reciprocal cooperation.

Model

In extant game-theoretic models of strong reciprocity (Charness & Rabin, 2004; Dufwenberg and Kirchsteiger, 2005; Falk and Fischbacher, 1998; Levine, 1998; Rabin, 1993; see Sobel, 2005 for a comprehensive review), a player judges a counterpart’s behavior by considering the material payoff she receives due to the counterpart’s actions. When the material payoff she receives exceeds a relevant standard of fairness, she perceives her counterpart as kind. When it falls short of the standard, she perceives her counterpart as unkind. The models differ in their details, but all assume that the fairness standard falls between the payoffs to cooperation and defection. This framework can accommodate a great deal of behavior (Falk, Fehr, & Fischbacher, 2008; Murphy & Ackermann, 2014). However, it implies that cooperation is necessarily kind and defection is necessarily unkind. It cannot accommodate the skeptical assessment that a player’s cooperation is not kind because it is not genuine.

Accommodating skepticism requires a richer attributional structure. We suggest that it is natural for a player to consider both the material reward she accrues from a counterpart’s actions and the material reward the counterpart himself accrues. For instance, as a first-mover’s cooperation provides him with greater material gain relative to his other available actions, it may be attributed less to genuine kindness and more to ulterior motives. Essentially, the potential kindness of a counterpart’s cooperation (or the potential unkindness of a counterpart’s defection) can be discounted by an assessment of how materially profitable the action is for him or her. Such discounting allows for meaningful attributions of motives
and can help capture the notion of skepticism (cf. Kelley & Stahelski, 1970a, 1970b; Inbar, Pizarro, & Cushman, 2012; Stanca, 2010).

The specific model of attributions we propose can be applied to any two-player game with finite action sets. It builds on a framework introduced by Segal and Sobel (2007). We formally state our model in the Appendix. We next explicate the model’s workings, using the sequential prisoners’ dilemma as an example.

**Motivations and Motivation Scores**

Suppose the players in the game of Figure 1 anticipate that the first-mover will cooperate and that the second-mover will reciprocate both first-move cooperation and first-move defection. How will they assess the motivation underlying the second-mover’s cooperation?

Taking the anticipated strategy profile as given, the model generates an assessment in three steps. First, it identifies all alternative pure strategies that the second-mover could take and that would change at least one player’s payoff. In the game of Figure 1, there is one set of such alternatives: the second-mover could defect. Second, the model compares the payoffs from the selected and alternative strategies, and on that basis classifies the second-mover’s strategy into one of the four cells of Table 1. Given the anticipated play from Player 1, the second-mover’s cooperation belongs in the upper-right cell: Relative to the alternative, it materially “helps” the other player, the first-mover, because the first-mover garners $4 and would garner less, $0, if the second-mover defected. Furthermore, relative to the alternative, the second-mover’s strategy materially “hurts” the second-mover herself. She receives $4 but would receive $6 by defecting. Third, on the basis of the strategy’s classification, the model assigns a “motivation score” ranging from +1, which corresponds to unambiguous kindness, to −1, which corresponds to unambiguous unkindness. In the situation at hand, because the second-mover has given up material payoff of her own to increase the first-mover’s material payoff, her strategy is deemed a +1. She has self-sacrificed to improve the lot of the first mover, and her genuineness is thus undoubted; she could not possibly have been motivated by self-interest.
What about an assessment of the first-mover’s cooperation? We have already argued that this strategy can be viewed skeptically. The model makes the argument as follows. First, the only alternative is for the first-mover to defect. Second, cooperation by the first-mover yields the players $4 each. However, given the anticipated strategy profile, if the first-mover instead defected, the players would receive $2 each. Thus, the first-mover’s cooperation materially “helps” the second-mover, but it also materially “helps” the first-mover himself. It is thus classified in the upper-left cell. Third, on that basis, it receives a motivation score of $1 - \theta_i$ from player $i$. The parameter $\theta$, which satisfies $0 \leq \theta \leq 1$, is a measure of skepticism. As a player is increasingly skeptical, her assessment of the first-mover’s kindness is increasingly discounted by the possibility that the first-mover cooperated out of calculated self-interest rather than genuine other-regard.

The remainder of the table is similarly derived. If a player sacrifices her own material payoff to hurt the other’s material payoff, her strategy is classified into the lower-right and scored $-1$ (cf. Fehr & Gächter, 2002). Because she effectively paid to hurt the other player, the strategy is unambiguously mean. On the other hand, a player who lowers her counterpart’s material payoff in the course of pursuing her own material self-interest receives a motivation score of $-1 + \theta_i$. Because the strategy might be driven by self-interest, it is not seen as necessarily inherently mean. A second-mover’s defection provides one example. It materially hurts the first-mover (relative to the alternative of cooperating). But it also maximizes the second-mover’s own material payoff. So the second-mover’s intentions may not be purely mean. Herein lies a key to the model: $\theta$ dilutes attributions of both kindness and unkindness.

In the above example involving the first-mover, the player had just one relevant alternative to his anticipated strategy. In general, a player may have many alternative strategies, and his motivation score will be the average score derived from comparisons with every alternative pure strategy. This aspect of the model is important in generating predictions for our second experiment, and we will discuss it in detail in that context. Relatedly, and as we detail fully in the Appendix, motivation scores in strategy profiles involving mixed strategies are probability-weighted averages of the motivation scores accrued
under each resulting pure strategy profile. Conceptually, this approach follows Dufwenberg & Kirchsteiger (2004) in viewing mixed strategies as reflections of incomplete information about population behavior and not as an individual’s conscious decision to randomize (see also Segal & Sobel, 2007).

The total payoff of player $i$ is given by $\pi_i = u_i + \lambda_i u_j M_i M_{ij}$. Here, $u_i$ is the player’s utility for his material outcome. His psychological payoff includes his counterpart’s utility for her material outcome, $u_j$, weighted by several parameters. First, $\lambda_i \geq 0$ indexes the player’s degree of other-regard; the greater is $\lambda_i$, the more the player cares about reciprocating kindness with kindness and unkindness with unkindness. Second, $M_i$ and $M_{ij}$ are player $i$’s assessments of her and the other player’s motives. Combining motivation scores multiplicatively allows for reciprocity. If a counterpart is being kind, then a player’s total payoff rises as she is increasingly kind in return. Likewise, if a counterpart is being unkind, then a player’s total payoff rises as she is increasingly unkind in return.

The total payoff functions we specify satisfy a set of conditions identified by Segal and Sobel’s (2007) that guarantee the existence of at least one “skeptical equilibrium” in every two-player game. That is, given the payoffs and motivation scores we have outlined, every two-player game includes at least one strategy profile from which neither player can unilaterally deviate to increase his or her total payoff. Many games will, of course, permit multiple equilibria. For instance, if the second-mover’s $\lambda$ is sufficiently large and her $\theta$ is sufficiently small, then in a sequential prisoners’ dilemma one equilibrium is the strategy profile on which we have focused: cooperation by the first-mover along with reciprocity of both cooperation and defection by the second-mover. Another equilibrium can obtain given any values of $\lambda$ and $\theta$: defection by the first-mover along with defection by the second-mover in response to any first move.

Note that conditioning players’ judgments on an entire anticipated strategy profile is sensible, because anticipated actions off the path of play must factor into the assessment of motivations. First-move cooperation, for instance, can only be viewed skeptically if the second-mover is expected to reciprocate defection as well as cooperation. Both expectations are necessary for the first-mover to
surmise that cooperating rather than defecting is materially beneficial for him, which in turn engenders second-mover skepticism.

Finally, there are surely additional motivations, beyond the pursuit of material self-interest and a desire for reciprocity, that often play important roles in social interactions. We later explore two such motivations that likely arise in Experiment 1, altruism and a form of impression management. Our model leverages the simple observation that people are often skeptical of each other’s genuineness because they worry about tactical behavior. It amounts to a ceteris paribus analysis. Holding other factors constant, it examines the implications of ambiguity about kindness or unkindness versus material self-interest.

Comparisons with Strong Reciprocity

Extant models of strong reciprocity correspond to the special case of $\theta = 0$ for both players. In that case, the material profit a player himself accrues from some action does not factor into an assessment of his motivations. Instead, his motivations are assessed strictly on the basis of their material implications for his counterpart. As a result, cooperation, which on a relative basis always materially helps the other, can be identified with kindness. Likewise, defection, which on a relative basis always materially hurts the other, can be identified with unkindness.

With $\theta$ set to zero, a second-mover’s $\lambda$ is the sole determinant of whether she reciprocates cooperation. In line with the notion of a material-social tradeoff, second-movers with small $\lambda$ are relatively self-regarding and do not reciprocate cooperation. Second-movers with large $\lambda$ are relatively other-regarding and do reciprocate cooperation. In contrast, and as we have emphasized, with $\theta > 0$, there is room for principled defection. Even a player who cares a great deal about how she treats others and how others treat her (large $\lambda$) can defect on cooperation if she is sufficiently worried (large $\theta$) that her counterpart’s cooperation was tactical rather than genuine.

We next present two variants of the sequential prisoners’ dilemma that allow us to experimentally test for manifestations of skeptical reciprocity and principled defection. The first variant is simpler, but as
we will detail, may provide a rather conservative test of our theoretical predictions. The second variant is a more complex game but may provide a more even-handed test.

**Variant 1: De-Coupled Prisoners’ Dilemmas**

Consider the following framing of the game in Figure 1. To begin, the first-mover faces a choice between $2 for himself or $4 for the second-mover. Then, the second-mover faces an analogous choice, between $2 for herself or $4 for the first-mover (cf. Pruitt, 1970; Vinacke, 1969). In a standard, sequential prisoners’ dilemma, the players have common knowledge of this sequence of choices from the outset of the game. Our “de-coupled” variant differs in only a single feature: The players are not initially informed that the second player will eventually have a choice of her own. At the outset of the game, they know only (a) that the first player faces his choice and (b) that the second player will be notified of the action the first player selects. They are not told anything else. Only later, after the first player has selected an action and the second player has been notified, are they informed that the second player indeed faces her own choice. (The first-mover’s decision in a de-coupled prisoners’ dilemma is similar to the active player’s decision in the well-known dictator game. For a review of dictator game experiments, see Camerer, 2003).

Importantly, de-coupled game first-movers thus select an action without realizing that their selection could subsequently influence the second-mover’s choice. As a result, first-move cooperation in the de-coupled game cannot be a tactic for eliciting reciprocal cooperation. By our model, it must be driven by genuine kindness rather than self-interest. Put differently, de-coupling removes ambiguity about a first-move cooperator’s motives. Indeed, in equilibrium, a de-coupled game first-move cooperator believes he is “hurting himself” (the first-mover is taking $0 rather than $2) to “help the other player” (who will receive $4 rather than $0). He therefore accrues a motivation score of +1, reflecting no skepticism from the second-mover. The score of +1 stands in contrast to the score of $+1 – \theta$ accrued by a sequential game first-move cooperator who anticipates reciprocal cooperation and whose motives may thus be questioned.
On the other hand, de-coupling does not remove ambiguity about a first-move defector’s motives. In equilibrium, a sequential game first-move defector anticipates that the second-mover will defect on cooperation rather than reciprocate it. Given this anticipated strategy profile, a sequential game first-mover’s defection “hurts the other player” (who will receive $2 rather than $6) but also “helps himself” (by taking $2 rather than $0). In a de-coupled setting, a first-move defector similarly “hurts the other player” (who will receive $0 rather than $4) but “helps himself” (by taking $2 rather than $0). First-move defection therefore accrues a motivation score of $-1 + \theta$ in both games, reflecting the possibility that it is driven by either meanness or self-interest. Note, incidentally, that contingent on a negative motivation score for the first-mover, only motivations beyond our model, like altruism and impression management, could induce the second-mover to cooperate with a first-move defector. Within our model, a second-mover gains both materially and psychologically by defecting rather than cooperating in response to first-move defection.

In sum, in equilibrium, de-coupling removes ambiguity about the motives for first-move cooperation but not first-move defection. Our model thus makes the following dual prediction about de-coupled game second-movers: relative to a standard, sequential game, they will be more likely to reciprocate cooperation but not defection.

Experiment 1

Method

Participants. A total of N = 288 undergraduates (50% female, mean age = 21.3 years) at UCSD’s Rady School of Management took part in our experiment, which was the first of a series of unrelated studies lasting about an hour. They received course credit for participation. Their choices were also incentivized with the monetary payouts shown in Figure 1. Power calculations based on a pilot study suggested that during the period in which the experiment was conducted, approximately one week’s worth of participants would generate a sufficient sample size; we thus ran during one Monday through Friday period.
Procedure. We conducted sessions of between five and twenty participants in a large computer lab with thirty-two private workstations. Participants were seated in an arrangement that maximized physical separation. Instructions were displayed on each participant’s computer screen. They included several training questions that participants were required to answer correctly before continuing. The experimental software was programmed in z-Tree (Fischbacher, 2007).

Each session was randomly assigned to either the sequential or de-coupled condition. Participants played anonymously with others in the same session. Because our central hypotheses concerned second-movers, we matched each first-mover with up to three second-movers. This feature of the matching procedure was not explicitly mentioned in the instructions, which read “You have been matched with another participant here today.” The first-mover was paid on the basis of the outcome generated by him and a second-mover randomly chosen from those with whom he had been matched. Each second-mover was paid on the basis of the outcome generated by her and her first-mover.

Materials. The specific framing we implemented for both the sequential and de-coupled games did not employ the terms defection and cooperation. At the outset of the sequential game, we credited each player with $2. We then had the first-mover choose between keeping his $2 (i.e., defecting) versus giving the money to his counterpart (i.e., cooperating); we also told participants that if the first-mover chose to give his money, we as the experimenters would add $2 to the amount conveyed to the second-mover, so that the second-mover would receive a total of $4. Thus, the first-mover was in effect asked to choose either $2 for himself or $4 for his counterpart. Furthermore, we told the sequential game players that after the first-mover selected an action, the second-mover would be informed of that action and be given the same choice. That is, the second-mover would choose either to keep her $2 (i.e., defect) or give her money to the first-mover (i.e., cooperate), with the same proviso about us as the experimenters adding $2 to the amount conveyed, so that the first-mover would receive a total of $4. These instructions yield the one-shot, sequential prisoners’ dilemma depicted in Figure 1. They are included in the Supplemental Materials.
At the outset of the de-coupled game, we credited only the first-mover with $2. We then had the first-mover make his choice, and we told the players that after the first-mover selected an action, the second-mover would be informed of that action. But we did not inform the players that the second-mover would eventually face the same choice. Instead, that choice was later introduced by informing the players that the second-mover was also being credited with $2 and had a “surprise” decision to make. De-coupled instructions are also included in the Supplemental Materials.

Finally, we collected rating scale data concerning perceptions of motives. In both conditions, after they had been notified of the first-mover’s action, second-movers in a randomized order (a) made their own choice and (b) assessed their counterpart’s motives. In the assessment task, second-movers were asked “to what extent do you think the first-mover was being genuinely nice in giving his or her $2 to you (keeping his or her $2)?” and “to what extent do you think the first-mover was calculating his or her self-interest in giving his or her $2 to you (keeping his or her $2)?” The two questions were displayed simultaneously on the same screen, with the question concerning genuine niceness always first. Participants responded on 11-point scales with endpoints labeled “Not at all” and “A lot.”

We expected that first-move cooperators would be perceived as more genuinely nice in the de-coupled than sequential game. In addition, we thought the ratings could help to differentiate between two potential drivers of \( \theta \), perceptions of others’ motives and reaction to others’ motives. To explicate, note that second-movers could vary in their tolerance of perceived self-interest. Some second-movers could be comfortable reciprocating cooperation even if they perceive a substantial degree of self-interest by the first-mover. Others might not reciprocate cooperation even if they perceive only minimal self-interest by the first-mover. That is, two individuals who agree in their perceptions of a first-mover, and who therefore provide the same ratings, may nevertheless arrive at different decisions about whether to reciprocate. If perceptions alone are determinative, our data should show that second-movers’ ratings mediate their actions. If ensuing reactions are important, mediation is less likely. In either case, however, our theory predicts that the genuine niceness of first-move cooperators should on average be rated higher under de-coupling than in the standard, sequential game.
Results

Our data corroborate the predicted manifestations of skeptical reciprocity. To begin, consider first-movers. As Table 2 shows, 76% of sequential game first-movers cooperated. This cooperation rate is in line with previous studies (Clark & Sefton, 2001; Hayashi, Ostrom, Walker, & Yamagishi, 1999; Kiyonari, Tanida, & Yamagishi, 2000). By contrast, only 51% of de-coupled game first-movers cooperated. The difference is statistically significant, $\chi^2(1, 84) = 20.9, p < .01$. It suggests that tactical cooperation is common in the sequential setting. Indeed, if first-move cooperation in that setting can reflect either genuine kindness or tactical behavior, but de-coupling precludes tactics, then about 33% of sequential game first-move cooperation could be tactical ($33\% = \frac{76\% - 51\%}{76\%}$).

Next, consider second-movers who faced first-move cooperation. In the sequential game, 72% reciprocated with cooperation of their own, which is in line with previous studies (Clark & Sefton, 2001; Hayashi, Ostrom, Walker, & Yamagishi, 1999; Kiyonari, Tanida, & Yamagishi, 2000; Tversky & Shafir, 1992). In contrast, in the de-coupled game, 85% reciprocated. The difference is marginally significant in the raw data, $\chi^2(1, 130) = 2.78, p = .096$, and significant in a logistic regression that controls for relevant participant- and session-characteristics ($b_{Game} = 1.25, SE = 0.57, z = 2.18, p = .030$).\(^1\)

The difference in positive reciprocity rates across games suggests that many failures to reciprocate cooperation constitute principled defection. If defections on cooperation in the sequential setting can reflect either self-interest or skepticism, but there is no room for skepticism in the de-coupled game, then roughly 46% of sequential game defections on cooperation can be traced to skepticism ($46\% = 100\% - \frac{[100\% - 85\%]}{[100\% - 72\%]}$). Herein lies an upside to skeptical reciprocity. Second-movers who cannot be convinced of a first-mover’s kindness in a standard, sequential game may be convincible.

\(^1\) All regression analyses that we report feature the same set of control variables. At the participant-level, we control for age and gender. At the session-level, we control for the time of the day (AM vs. PM), and the day of the week. These control variables reflect both extant research findings and local experience. Croson and Gneezy (2009) and Eckel and Grossman (2008) review research that establishes robust gender differences in reciprocity. At the lab where we conducted our experiments, PM and Friday participants are known to be relatively less attentive. Coefficient estimates for the control variables were generally of the predicted sign, and reached statistical significance in some analyses.
in other settings. Demonstrating genuineness can convert what would otherwise be principled defection into cooperation.

Turning to second-movers who faced first-move defection, de-coupling did not increase their negative reciprocity. In the sequential game, 78% reciprocated with defection of their own, which is in line with previous studies (Clark & Sefton, 2001; Hayashi, Ostrom, Walker, & Yamagishi, 1999; Kiyonari, Tanida, & Yamagishi, 2000; Murphy & Ackermann, 2014; Tversky & Shafir, 1992). In the de-coupled game, somewhat fewer, 70%, reciprocated with defection. The difference across games falls short of significance in both the raw data ($\chi^2(1, 74) = 0.50, p = .481$) and in a logistic regression ($b_{\text{Game}} = 0.31, SE = 0.69, z = 0.44, p = .657$).

Finally, we pooled all second-movers in a single logistic regression. In this analysis, the dependent variable was whether second-movers reciprocated their first-mover (i.e., whether they made the same decision as the first-mover). We regressed this measure of reciprocity on first-mover decisions and two dummy variables: A “positive reciprocity” dummy ($PR$) took the value 1 only for de-coupled game second-movers reacting to first-move cooperation and was equal to 0 otherwise. Similarly, a “negative reciprocity” dummy ($NR$) took the value 1 only for de-coupled game second-movers reacting to first-move defection and was equal to 0 otherwise. Our dual prediction that de-coupling increases positive but not negative reciprocity implies that the coefficient for the $PR$-dummy will be positive, whereas the coefficient for the $NR$-dummy will not differ from zero. This prediction found support in the data. Compared to sequential game second-movers, de-coupled second-movers were more likely to reciprocate cooperation ($b_{PR} = 1.02, SE = 0.53, z = 1.95, p = .052$) but not defection ($b_{NR} = -.22, SE = 0.62, z = -0.35, p = .726$).

The ratings provided by second-movers are also consistent with our model. Cooperative first-mover were rated more genuinely nice in the de-coupled game ($M = 7.5, SD = 1.9$) than in the sequential game ($M = 6.0, SD = 2.5$), $t(117.4) = 3.77, p < .01$. They were also rated as less self-interested (Ms = 4.4 vs. 6.3, SDs = 2.4 vs. 2.7), $t(106.1) = 4.06, p < .01$. Thus, relative to the de-coupled game, participants discounted the notion that sequential game first-move cooperation is genuinely nice, while they placed
greater credence on the notion that it is self-interested. The ratings of first-move defectors showed a more muted pattern. Relative to the sequential game, de-coupled game first-move defectors were rated slightly but not significantly less nice ($M_s = 2.1$ vs. $2.9$, $SD_s = 2.3$ vs. $2.3$), $t(54.8) = -1.35$, $p = .18$, and marginally more self-interested ($M_s = 8.7$ vs. $7.7$, $SD_s = 2.5$ vs. $1.7$), $t(39.0) = 1.73$, $p = .09$.

Second mover’s ratings did not mediate their reciprocity decisions. As mentioned above, the absence of mediation is consistent with meaningful cross-person variation in second-movers’ tolerance of self-interest and lack of genuineness. That is, individuals who agreed in their perceptions of a cooperative first-mover’s motives and provided the same ratings may nevertheless arrive at different decisions about whether to reciprocate.

We have so far assumed that second-movers’ ratings are truthful reports of their assessments of first-movers. In principle, however, second-movers could distort their ratings to cast their own actions in the best possible light. For instance, second-movers who defect on cooperation could self-justify or rationalize their behavior by depicting the first-mover as relatively less genuine and more self-interested. Our data show little evidence of such distortions. Unlike other second-movers, de-coupled game second movers who provided ratings before making their own choice could not deliberately mold their ratings; at the time they rated the first-mover, they did not yet know that they would have their own choice to make. If ratings distortion is common among the remaining second-movers, our data would thus show a critical game-by-task-order interaction. But they did not. An F-test comparing linear models of ratings of genuine niceness with or without the critical interaction clearly favored the simpler model, $F(1, 119) = .025$, $p = .95$. Analogous results obtained for ratings of self-interest, $F(1, 119) = 2.31$, $p = .13$, and indeed for reciprocal cooperation itself, $\chi^2(1, 119) = .388$, $p = .53$.

Additional Motivations

Our model contrasts kindness, meanness, and self-interest, but as we have mentioned, there are surely additional motivations that can play a role in prisoners’ dilemmas and related games (see, e.g., Andreoni & Samuelson, 2006). We next describe two motivations that bear special mention because they
SKEPTICAL RECIPROCITY AND PRINCIPLED DEFECTION

likely arise in Experiment 1. Moreover, the influence of these motivations may render Experiment 1 a conservative test of skepticism and principled defection.

Pure altruism involves self-sacrificing, generous behavior that is not conditioned on any social antecedent or expectation (Fehr & Fischbacher, 2003; Fletcher & Zwick, 2007; Simon, 1993). It is closely related to the notion of genuine kindness yet lies outside our model as well as other accounts of reciprocity, which provide positive psychological payoffs for a self-sacrificing, generous action only if that action is expected to be reciprocated or itself reciprocates prior actions. Behavior consistent with pure altruism is, nevertheless, evident in Experiment 1. Like previous researchers (see Murphy & Ackermann, 2014), we observe some second-movers cooperating in response to first-mover defection. Furthermore, first-movers in the de-coupled game frequently cooperate.

In an important paper, Dana, Cain, and Dawes (2006) highlight a second motivation. These authors demonstrate that players often feel they are supposed to be generous and are therefore concerned that the experimenter and the other player may be judging them by their generosity. For example, second-movers in prisoner’s dilemmas may base reactions to first-mover cooperation not on reading and reacting to genuine kindness versus tactical self-interest, but on a belief that they are supposed to be generous and that they are being judged on that basis. Like pure altruism, this form of impression management is consistent with second-move cooperation in response to first-move defection as well as with first-mover cooperation in the de-coupled game.

In formulating our predictions, we assumed that motivations beyond reciprocity and self-interest did not differentially impact de-coupled versus sequential game second-movers. However, the two motivations just described may do so. By framing the players’ decisions as separate events rather than a unified interaction, de-coupling may reduce the subjective relevance of the first-mover’s decision to the second-mover’s decision. Reduced subjective relevance would lead de-coupled game second-movers to be less motivated by reciprocity than sequential game second-movers, and thus more motivated by pure altruism and impression management on the one hand, and self-interest on the other (cf. Ben-Ner, Putterman, Kong, & Magan, 2004).
Taken to the extreme, this line of reasoning suggests that de-coupled second-movers will behave just like their first-mover counterparts. In our sample, 51% of de-coupled game first-movers cooperated and 49% defected. Any bias in second-movers’ behavior toward such a roughly even split would lead us to underestimate the prevalence of skepticism and skeptical reciprocity in Experiment 1. It would work against our prediction that second-movers will reciprocate cooperation more frequently in the de-coupled than in the sequential game (where 72% reciprocated cooperation). It would also work against our prediction that second-movers will reciprocate defection equally frequently in the de-coupled and sequential games (where 78% reciprocated defection).

In sum, Experiment 1 is consistent with the notion that skepticism and principled defection are common. It is limited, however, because de-coupled games present some hurdles concerning first-move relevance. Accordingly, we will later introduce a second variant of the prisoners’ dilemma that seems unlikely to impact first-move relevance below, and use it to further test the predictions of our model in Experiment 2.

Related Results

Stanca, Bruni, and Corazzini (2009) studied sequential and de-coupled gift-exchange games (for background on gift-exchange experiments, see Fehr, Kirchsteiger, & Riedl, 1993, 1998). They report data that are difficult to interpret but potentially in line with Experiment 1. In their studies, each of two players was staked twenty tokens that were later redeemable for cash. The first-mover could send a gift of any portion of his stake to the second-mover; any sent tokens would be tripled. For example, if the first-mover sent ten tokens, the second-mover would receive thirty tokens. Before being notified of what the first-mover elected to do, the second-mover indicated what portion of her stake she would send the first-mover contingent on every possible first-mover action (i.e., how many tokens she would send him if he sent her zero tokens, how many she would send him if he sent her one token, etc.). Consistent with tactical self-interest, Stanca, Bruni, and Corazzini observed greater first-mover gift-giving in the standard,
sequential game. Consistent with skeptical reciprocity, they observed a greater positive Spearman correlation between potential first-mover actions and the second-mover’s responses under de-coupling.

Examining prisoners’ dilemmas allows us to build on Stanca, Bruni, and Corazzini’s findings, because there are substantial hurdles to testing for either positive or negative reciprocity using gift-exchange games. Our model can be used to illustrate the relevant intuition. Consider a first-mover in Stanca, Bruni, and Corazzini’s de-coupled setting who gives five of his twenty tokens to the second-mover. Compared to more generous gifts, the first-mover has hurt the second-mover and helped himself. The implied motivation score is therefore $-1 + \theta$. Compared to less generous gifts, the first mover has helped the second mover and hurt himself. The implied motivation score is $+1$. As we have mentioned, to derive an overall motivation score, we compute the average across comparisons to all pure strategies. Because there are fifteen more generous gifts and five less generous gifts, the overall motivation score is then $15(-1 + \theta) + 5(+1) = -\frac{1}{2} + \frac{3}{4}\theta$. If someone is skeptical enough, the overall motivation score is positive: The individual is impressed by even small measures of generosity. However, if someone is not very skeptical, the overall motivation score is negative: The individual is disappointed in merely small measures of generosity.

More generally, whenever the first-mover has made a gift that is non-zero but not too large (in either the de-coupled or sequential game), it is not possible to authoritatively identify it as kind or unkind. It could be viewed as either. Likewise, if the second-mover responds with a non-zero but not too large gift, it is not possible to authoritatively identify that action as kind or unkind. Thus, it is difficult to classify many interactions as mutually kind, mutually unkind, or a mix of kind and unkind. As a result, it is difficult to assess the prevalence of positive and negative reciprocity. This issue does not arise under the simpler structure of the prisoners’ dilemma. It also does not arise when a player gives either none or all of her tokens in a gift-exchange game. Giving no tokens is always unkind, and giving all of one’s token is always unambiguously kind. Stanca, Bruni, and Corazzini’s data for these specific gifts, however, are limited and inconclusive.
Variant 2: Endogenous Sequencing Prisoners’ Dilemmas

We now introduce a variant of the prisoners’ dilemma that seems unlikely to impact first-move relevance. This variant also has two other desirable properties. First, it features ex ante symmetry between the players. At the start of the game, nothing distinguishes the players. Second, it provides the players with a choice among equally cooperative behaviors which nevertheless differ in how credibly they demonstrate genuine kindness.

Consider an “endogenous-sequencing” prisoners’ dilemma in which the players are not assigned the roles of first-mover and second-mover. Instead, there is a visible countdown, and at each moment of the countdown either player can decide to make the first move or not. In other words, at each moment, a player may first-move cooperate, first-move defect, or simply continue to wait and see what happens. If a player does make the first move, the countdown is stopped, the other player is informed of the move, and she is provided the opportunity to respond. If neither player moves before the countdown expires, they proceed to play a simultaneous prisoners’ dilemma. As in many real-world circumstances, deciding whether to move first is a crucial aspect of strategy in this game.

Applying the Model

Our model suggests that the more rapidly a player first-move cooperates, the more he will be perceived to be acting out of kindness rather than self-interest. The key idea is that unlike the sequential game, under endogenous sequencing, a first-move cooperator can incur costs that “hurt himself” materially even in an anticipated strategy profile that features mutual cooperation. For example, a player interested in garnering the maximal possible material payoff could wait, hope his counterpart first-move cooperates, and then defect on her. By first-move cooperating, a player forsakes the chance to take advantage of his counterpart in this way. The more rapidly he first-move cooperates, the more emphatically he forsakes doing so. In our framework, this “cost” makes first-move cooperation seem less motivated by self-interest and therefore more motivated by kindness.
The intuition is that immediately renouncing any attempt to take advantage of your counterpart makes you seem genuinely kind. Waiting to do so generates ambiguity and opens the door to skepticism. Recent social psychological research is consistent with this intuition. Critcher, Pizarro, and Inbar (2013) report that decisions that are potentially morally good, like the decision to first-move cooperate, are judged especially positively when they are made quickly. Likewise, decisions that are potentially morally bad are judged especially negatively when they are made quickly. Evidently, taking time to ponder a decision dilutes attributions of both positive and negative moral intent. Indeed, Van de Calseyde, Keren, and Zeelenberg (2014) provide evidence that the amount of time people take in reaching a decision is seen as a measure of their internal conflict about what to do.

_Deriving an Equilibrium under Endogenous Sequencing_

We now more thoroughly describe how our model can characterize fast first-move cooperation as especially genuinely kind (formal details can be found in the Appendix). To begin, specify strategies according to (a) whether a player would ever make the first move, and if so what that move would be and when it would take place, (b) how she would respond at any possible moment to the other player first-move cooperating, (c) how she would respond at any possible moment to the other player first-move defecting, and (d) how she would act in the simultaneous game that would arise if neither player ever makes a first move.

Furthermore, suppose the players anticipate that each would first-move cooperate at some point, and that each would reciprocate both first-move cooperation and first-move defection that occurred at any point during the countdown. This strategy profile may be viewed as an extension of the sequential game profile discussed earlier, in which the first-mover cooperates and the second-mover reciprocates both cooperation and defection. It arguably epitomizes reciprocity: The players are maximally willing to be a part of mutual cooperation as well as mutual defection.

Under this anticipated strategy profile, the player who becomes the first-mover has two sets of alternative strategies that would alter the players’ payoffs. He could first-move defect. Relative to this
possibility, his first-move cooperation materially helps both the other player and himself. Its motivation score is thus $+1 - \theta$, just as in the sequential game. He could also wait, have the other player first-move cooperate, and then defect on that cooperation. Relative to this latter possibility, his first-move cooperation materially helps the other player but materially hurts him. Its motivation score is thus +1. His overall motivation score must therefore lie between $+1 - \theta$ and +1, because it averages over all relevant possibilities. A first-move cooperator will seem at least as kind under this anticipated strategy profile as a first-move cooperator in the sequential game. He can seem more kind, but he cannot seem less kind.

Moreover, he will seem increasingly kind as he must act earlier to become the first-move cooperator. Suppose the second-mover’s strategy would have her first-move cooperate at time $t_2$, and consider the consequences of pushing $t_2$ earlier. The first-move cooperator will have relatively fewer distinct strategies by which he could instead first-move defect. For example, he will have fewer specific moments at which to make that first-move defection. On the other hand, the first-move cooperator will have relatively more distinct strategies by which he could instead wait out his counterpart and then defect on her. Thus, as he must act earlier to become the first-mover, a player’s overall motivation score will shift away from $+1 - \theta$ and closer to +1. Indeed, if player must act right away to first-move cooperate because $t_2$ is near the beginning of the countdown, his motivation score will approach +1.

The second-mover also has two sets of alternative strategies that would alter the players’ payoffs. Suppose the first-mover’s strategy has him first-move cooperate at time $t_1$. Then, the second-mover could pre-empt the first-mover with a first-move defection prior to $t_1$. Relative to this possibility, her anticipated strategy helps both players and thus has a motivation score of $+1 - \theta$. Or she could second-move defect. Relative to this possibility, her anticipated strategy helps the first-mover while it hurts her. Its motivation score is thus +1. Furthermore, as $t_1$ is pushed earlier, the second-mover’s overall motivation score also shifts away from $+1 - \theta$ and closer to +1.

In sum, the players’ overall motivation scores both approach one as the anticipated first-move approaches the beginning of the countdown. That is, no matter how skeptical the players are, maximal
genuine kindness and positive reciprocity are engendered when the players expect each other to cooperate as quickly as possible and to reciprocate both first-move cooperation and first-move defection. In such a circumstance, each player’s strategy makes him or her as vulnerable as possible: Each player is cooperating in a manner that provides the counterpart with the greatest possible opportunity to defect. Yet, each player is also declining the opportunity to act self-interestedly in this way. The players are epitomizing reciprocity; indeed, their anticipated strategies form an equilibrium. Furthermore, compared to all other equilibria, this equilibrium yields any sufficiently other-regarding player the highest possible total payoff, summing across material and psychological rewards.

**Empirical Predictions**

Though there are many other equilibria, the foregoing suggests two manifestations of skeptical reciprocity that may emerge under endogenous sequencing. First, many games will end quickly, and when they do it will be because of first-move cooperation. Second, and most importantly, quick, first-move cooperation will be reciprocated at a higher rate than first-move cooperation in the standard, sequential game. This latter hypothesis follows from the fact that under the anticipated strategy profile just discussed, reciprocating quick first-move cooperation engenders a product of motivation scores close to +1, whereas in the standard, sequential game the corresponding product is only $+1 - \theta$. Again, herein lies an upside to skeptical reciprocity: Skeptical second-movers who cannot be convinced of a first-mover’s kindness in a standard, sequential game may be convincible under endogenous-sequencing.

Neither material payoff maximization nor strong reciprocity makes the two-part empirical prediction just outlined. Clearly, in a one-shot game, material payoff-maximization (which corresponds to $\lambda = 0$ for every player) requires that second-movers always defect. Equally clearly, strong reciprocity allows second-movers to cooperate. Strong reciprocity is silent, however, about how first-move cooperation and reciprocation of it will vary with time elapsed. When there is no skepticism, so that $\theta$ equals zero, each player’s motivation score, which corresponds to a weighted average of $+1 - \theta$ and +1, always reduces to exactly one.
Experiment 2

Method

Participants. A total of $N = 242$ undergraduates at UCSD’s Rady School of Management took part in our experiment and subsequent, unrelated studies. Their mean age was 21.0 years and 131 of them (54.1%) were female. As in Experiment 1, they received course credit for participating and were also paid according to their game outcome. Prior experience at the Rady lab suggested that during the period when the experiment was conducted, two weeks’ worth of sessions would yield sufficiently many participants.

Procedure. We conducted sessions of between 6 and 20 participants. Each session was randomly assigned to either the sequential or the endogenous-sequencing condition. For robustness, we ran endogenous sequencing sessions with three different timer lengths: 20, 60, and 120 seconds.

We maintained the framing of Experiment 1. That is, we did not employ the terms defection and cooperation. Instead, at the outset of the game, we credited each player with $2. At any moment during the countdown, each player could continue to wait, choose to keep his $2 (i.e., defect), or choose to give his money to his counterpart (i.e., cooperate) with the proviso that we as the experimenters would add $2 to the amount conveyed, so that the counterpart would receive a total of $4.

As in Experiment 1, participants were anonymously paired with someone in the same session. Unlike Experiment 1, they were paired one-to-one; that is, each first-mover was paired with only one second-mover. In sessions with an odd number of participants, a research assistant who was blind to our hypotheses filled in. Research assistants were not remunerated for their choices, and their data were not included in our analyses.

Materials. The experiment was programmed using the Z-Tree software (Fischbacher, 2007). The sessions took place in a large computer lab with thirty-two private work stations. Participants were seated at workstations in an arrangement that maximized physical separation. Moreover, unlike Experiment 1, participants wore headphones to prevent them from hearing each other’s vocal reactions and typing.
noises. The instructions were displayed on participants’ computer screens and read aloud by the research assistant conducting the session. To further facilitate understanding of the experimental task, participants answered five training questions during the course of the instructions.

Participants correctly answered an average of 4.0 of the five comprehension questions. Participants in the endogenous-sequencing game on average answered slightly fewer questions correctly than participants in the sequential game, but the difference was not significant (p = .29). We did not exclude any participants from the analyses that follow; our results are qualitatively unchanged if only participants who correctly answered all five questions are included.

Results

Players’ behavior in the sequential game was in line with both Experiment 1 and previous research (Clark & Sefton, 2001; Hayashi, Ostrom, Walker, & Yamagishi, 1999; Kiyonari, Tanida, & Yamagishi, 2000). As Table 3 shows, 68% of first-movers cooperated, 69% of second-movers reacting to cooperation reciprocated with cooperation, and 94% of second-movers reacting to defection reciprocated with defection.

Players’ behavior under endogenous sequencing corroborated the predicted manifestations of skeptical reciprocity. Recall the initial element of our prediction: Games will often end quickly via first-move cooperation. The rightmost column of Table 3 confirms that many games indeed ended very rapidly, in a matter of seconds. Collapsing across the various timer lengths, the median time elapsed before a first-move was approximately 15% of the countdown. This aggregate statistic includes the very short 20-second games, in which a median time elapsed of a mere 6.0 seconds corresponds to 30% of the countdown. In what follows, we employ 20% of the countdown as a conservative cutoff defining fast versus slow first-moves, and we show that our conclusions are qualitatively unchanged as the cutoff is moved earlier.

Figure 2 provides initial evidence that games typically ended quickly because one of the players elected to first-move cooperate: 35 out of 40 (88%) first-moves that occurred within the initial 20% of the
countdown were cooperative (right panel, leftmost bar). As Table 4 indicates, the prevalence of cooperation remains very high as the cutoff defining fast versus slow first-moves is pushed earlier. For instance, 22 out of 25 (88%) of first-moves occurring within the initial 10% of the countdown were cooperative.

Recall the second element of our prediction: Quick first-move cooperation under endogenous sequencing should be reciprocated at a higher rate than first-move cooperation in the sequential game. Figure 3 illustrates that the data support this prediction: 31 out of 34 (91%) second-movers responding to cooperative first moves within the initial 20% of the countdown reciprocated cooperation (right panel, leftmost bar). This reciprocity rate significantly exceeds that of the sequential game (31/34 vs. 22/32, p = .031 by two-sided Fisher’s exact test). Table 4 indicates that this conclusion is robust to earlier cutoffs defining fast versus slow first moves. For instance, 21 out of 22 (95%) responses to cooperative first-moves occurring within the initial 10% of the countdown were themselves cooperative. Evidently, potent signals of genuine kindness can catalyze extremely high rates of reciprocity.

We wish to emphasize that second-movers reacting to quick first moves under endogenous sequencing are largely unaffected by selection bias. In particular, they are not necessarily unwilling to make a quick first-move; they merely happen to be paired with an individual who moved more quickly. Thus, they are directly comparable to second-movers in the sequential game. Only second-movers who faced slow first-movers form a biased sample, because they chose not to make an early first move.

Notwithstanding this caveat, a logistic regression model shows that the likelihood of second-movers responding to cooperation decreased substantially for each percentage point of the countdown elapsed prior to the cooperative first-move ($b_{\text{Elapsed}} = -5.28$, $SE = 1.77$, $z = -2.97$, $p < .01$). To put this coefficient estimate into context, the model estimates reciprocity rates of 93% for instantaneous first-move cooperation and 6% for first-move cooperation just prior to the countdown expiring.

---

2 Because of a cell count smaller than five for defection in response to cooperation under endogenous sequencing, we rely on Fisher's exact test rather than a Chi-square test to assess the statistical significance of the observed difference in proportions.
As in Experiment 1, it is useful to estimate the proportion of sequential game failures to reciprocate cooperation that may be principled. To do so, we can assume that defection on cooperation can reflect both self-interest and skepticism in the sequential game, but that skepticism is virtually absent when the first-mover cooperates quickly under endogenous sequencing. Then given the conservative cutoff defining fast versus slow first-moves, about 71% of failures to reciprocate cooperation in the sequential game are principled (71% = 100% - [[100% - 91%]/[100% - 69%]]). Recall that Experiment 1 suggested that 46% of failures to reciprocate cooperation constituted principled defection. Whether the “true” figure is 71%, or 46%, or somewhere in between, skepticism and principled defection appear to be common. The prevalence of skepticism both sets the stage for principled defection and, on the flip side, provides the room for effective demonstrations of genuine kindness to catalyze high rates of positive reciprocity.

A Two System Framework and Reaction Times

Rand, Greene, and Nowak (2012) argue that an automatic, intuitive mental system is biased towards social concerns, while a slower, deliberative mental system is biased towards material concerns (cf. Loewenstein & Small, 2007). They collected reaction time data for choices in one-shot, simultaneous-play prisoners’ dilemmas and public good games and found faster choices to be more cooperative (similar results are reported by Halali, Bereby-Meyer, & Meiran, 2014; Halevy & Chou, 2014; opposite results are reported by Kessler & Meier, 2014; Tinghög et al., 2013; a meta-analytic review of relevant evidence is provided by Rand, 2016). The validity of relevant inferences from reaction time data has come under criticism for conceptual reasons (Krajibich, Bartling, Hare, & Fehr, 2015), and the cognitive and neural processes underlying social motivations constitute an extremely active research area (see Apps, Rushworth, & Change, 2016; Gluth & Fontanesi, 2016; Hein, Morishima, Leiberg, Sul, & Fehr, 2016; Yamagishi, Takagishi, Fermin, Kanai, Li, & Matsumoto, 2016). Nevertheless, the contention that there may be an instinct for cooperation and other-regard that is disrupted by greedy, sophisticated thinking has been influential.
Some of our results lend support to this contention, others do not, and still others cannot be cleanly interpreted either way. Consider first- and second-movers in turn. An automatic system could of course contribute to rapid first-move cooperation under endogenous sequencing. On the other hand, simple linear regressions of reaction times on decisions show that in our de-coupled game first-move cooperators were actually somewhat slower than first-move defectors ($p = .06$). Furthermore, as we have emphasized, sequential game first-movers who are focused on material payoff maximization may cooperate rather than defect for tactical reasons. This precludes clean interpretations of their reaction times. Though sequential game first-move cooperators did act faster than sequential game first-move defectors in Experiment 1 ($p = .12$) and significantly so in Experiment 2 ($p = .02$), their rapid cooperative choices could in principle reflect an automatic system that is selfinterested rather than other-regarding. This confound does not arise for second-movers, because it is never materially profitable for them to cooperate. But our second-mover data are mixed: Second-move cooperators acted significantly faster than second-move defectors in both the sequential and the decoupled games of Experiment 1 ($ps = .02, .04$), but neither in the sequential game nor under endogenous sequencing in Experiment 2 ($ps = .39, .89$). Finally, note that the two-system framework does not speak to the possibility of pronounced second-mover reciprocity to rapid first-move cooperation under endogenous sequencing.

Discussion

Strong reciprocity theories do not distinguish between good treatment resulting from genuine kindness and good treatment resulting from calculated self-interest. To fill this gap, we have introduced a model of skeptical reciprocity and tested it experimentally. For skeptical reciprocators, being treated well is not enough. When the good treatment they receive may be a by-product of self-interest, they do not have a desire to reciprocate. Only when the good treatment they receive is demonstrably a product of genuine kindness do they wish to reciprocate. Our model and data suggest that skepticism rather than self-interest may drive much non-reciprocation. They also suggest that when first-movers obviate skepticism by effectively signaling genuine kindness, reciprocity abounds.
Related Research on Other Games

We have focused on one-shot interactions, but the notion of skeptical reciprocity may temper some common conclusions regarding finitely-iterated, simultaneous prisoners’ dilemmas. In such games, cooperation rates are often initially high but then “unravel” as the endgame approaches (unraveling is observed by Andreoni and Miller, 1993 and Morehous, 1966; however, greater cooperation over time is reported by Rapaport & Chammah, 1965; Erev & Roth, 2002; Chater, Vlaev, & Grinberg, 2008). Strong reciprocity is at odds with unraveling, because it implies that a history of past cooperation should engender subsequent cooperation. Indeed, unraveling is sometimes framed as damning evidence against an inherent desire to respond to kindness with kindness (Selten & Stoecker, 1986). Furthermore, unraveling is often depicted as substantial support for tactical cooperation. Players may cooperate tactically early on and in the middle portion of the game, in order to facilitate and maintain a subsequent stretch of materially profitable, reciprocal cooperation. Late in the game, however, with fewer and fewer subsequent periods, the incentive to continue this tactic wanes (Axelrod, 1981; Kreps, Milgrom, Roberts, & Wilson, 1982). Skeptical reciprocity, though it presumes an inherent desire to respond to kindness with kindness, is entirely consistent with unraveling. To a skeptical reciprocator, even a long history of cooperation may reflect ulterior motives and thus not merit late-period reciprocity.

Trust games (Berg, Dickhaut, & McCabe, 1995) generalize sequential prisoners’ dilemmas. They provide a relatively rich strategy space that does not dichotomize between defection and cooperation, thereby allowing for degrees of cooperative behavior. They are often employed to estimate parametric models of other-regarding preferences (e.g., Cox, 2004) which are in turn invoked in a wider debate about the nature of other-regard. In particular, it has been argued (Charness & Rabin, 2002) that people’s interest in positive reciprocity is dwarfed by their concern for distributional considerations like inequity aversion (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999). Whether that is the case is hotly debated (Cooper & Kagel, 2013; Choshen-Hillel & Yaniv, 2011). Our work suggests a caution relevant to this
debate: Because they do not account for skepticism, extant estimates of the desire for positive reciprocity are likely biased downwards.

Interestingly, just like people may have more interest in positive reciprocity than is commonly presumed, recent evidence suggests that they may have less interest in negative reciprocity. In an important paper, Yamagishi et al. (2012) study ultimatum games. In these games, a first-mover offers some split of a shared pot of money to a second-mover; the second-mover can accept or reject this split; and if she rejects it, neither player receives any money. Yamagishi et al. make a compelling argument that the commonly observed rejection of unkind offers should not be taken as evidence of interest in negative reciprocity. Unkind responses to unkind initial moves, they show, frequently reflect second-movers’ unwillingness to accept an inferior social status.

Finally, skeptical reciprocity may aid in understanding data from field studies of gift-exchange games, which, as we have mentioned, are another form of social dilemma that generalizes sequential prisoner’s dilemmas (Fehr, Kirchsteiger, & Riedl, 1993, 1998). Employees’ efforts after they receive “gifts” of unexpected pay raises are highly disparate from one workplace to another (Gneezy & List, 2006; Falk, 2007; Kube, Maréchal, & Puppe, 2012; Cohn, Fehr, Herrmann, & Schneider, 2014; Gilchrist, Luca, & Malhotra, 2016). We suggest that workers may be appreciative and responsive to pay raises perceived as genuine gifts but not to pay raises perceived as employer tactics for increasing productivity.

Modeling Interpretations

In explaining our data, we have adopted an interpretation of the model that allows for cross-person variation in both other-regard and skepticism. Under this interpretation, differences in reciprocal behavior can be explained by differences in either λ, θ, or both. We have emphasized differences in θ. For instance, consider two second-movers with similar, intermediate values for λ, but substantially different θ. In a de-coupled game which provides clarity about the first-mover’s motives, both may reciprocate. In a sequential game which renders the first-mover’s motives unclear, only the less skeptical second-mover—the player with a lower θ—may reciprocate.
Our data are also consistent with an alternative interpretation of the model that allows for cross-person variation in other-regard but not skepticism. Under this interpretation, every player has the same non-zero $\theta$. As a result, differences in reciprocal behavior can only be explained by differences in $\lambda$. For instance, a second-mover with intermediate $\lambda$ and a second-mover with high $\lambda$ may both reciprocate in a de-coupled game. But in a sequential game, only the more other-regarding second-mover—the player with a high $\lambda$—may reciprocate.

Allowing cross-person variation in both $\lambda$ and $\theta$ best comports with the notion of principled defection. Under this interpretation, a highly other-regarding second-mover could be less likely to reciprocate cooperation than a less other-regarding second-mover, if the highly other-regarding individual is also highly skeptical. Allowing for cross-person variation only in $\lambda$ does not as fully comport with principled defection, because it implies that highly other-regarding second-movers are always more likely to reciprocate than less other-regarding second-movers. Importantly, however, both interpretations are fundamentally consistent with principled defection. Both allow for non-zero $\theta$, thereby allowing non-reciprocation of first-move cooperation to reflect lack of clarity about motives rather than mere self-regard.

Finally, by interpreting our model in terms of cross-person variation in $\theta$, we do not mean to imply that skepticism is a stable personality trait (Ackermann, Fleiß, & Murphy 2014). Some situations likely accentuate or dampen skepticism. Person-by-situation interactions could also be important. Relatedly, future work could examine the implications of viewing skepticism as a summary of an individual’s beliefs about whether others are other-regarding. For instance, in a game of incomplete information, higher values of $\theta$ could correspond to greater confidence about a counterpart having low $\lambda$.

**Social Roots for Both Cooperation and Defection**

In our introduction, we noted that because cooperation entails foregoing the maximization of one’s material interests, many theories have strived to explain why cooperative behavior nevertheless abounds. These theories differ in their attitudes towards the notion of a material-social tradeoff,
SKEPTICAL RECIPROCITY AND PRINCIPLED DEFLECTION

according to which people behave non-cooperatively when material concerns predominate but behave cooperatively when social concerns predominate.

Some theories assert the primacy of a material-social tradeoff. Strong reciprocity is an account of this form. It highlights two sources of non-cooperative behavior: a person’s material interests and her desire for negative reciprocity toward those who thwart them. It highlights one source of cooperative behavior: the psychological payoff of positive reciprocity that can make up for lost material utility. Rand, Greene, and Nowak’s (2012) two-process framework likewise embraces a material-social tradeoff.

Other work accepts the notion of a material-social tradeoff but holds that the tradeoff is often less sharp than we might assume (see, e.g., Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak, & Greene, 2014). Kameda, Tsukasaki, Hastie, & Berg (2011) argue that in many settings, there is little conflict between the pursuit of self-interest and group welfare (see also Kameda & Nakanishi, 2002; Kameda, Takezawa, & Hastie, 2003). Baron (1997) argues that many people cooperate because they believe that even when cooperation is not materially profitable in the short-run, it must be materially profitable in the long-run. This belief reflects the view that morality and self-interest cannot truly conflict, because moral behavior is rewarded karmically.

Skeptical reciprocity does not fully agree with the notion of a material-social tradeoff. It accords with research which contends that like cooperation, defection may be socially-driven. Research on group engagement (see Tyler, 2013; Tyler & Blader, 2001, 2003), social identity (Tajfel & Turner, 1979; Hogg & Abrams, 1988; Ellemers, Spears, & Doosje, 2002; see also Cropanzano, Byrne, Bobocel, & Rupp, 2001), and social relations (Clark & Mills, 1979; Fiske, 1992) suggests that people often care about material outcomes primarily because of the social messages they convey. By this view, much non-cooperative behavior is inherently socially-driven. Miller (1999; Miller & Ratner, 1998; Ratner & Miller, 2001; Heath, 1999) stresses the “norm of self-interest.” People may behave non-cooperatively not because they are truly self-interested, but because they believe they are supposed to be. Skeptical reciprocity indicates that people may often act non-cooperatively not because they are self-interested, but
because they worry that even those who treat them well are fundamentally self-interested rather than good-hearted.

Should skeptical reciprocators be portrayed as hardened cynics who fail to appreciate a counterpart’s reaching out (Fetchenhauer, & Dunning, 2010; Kerr, 1983; Kruger & Gilovich, 1999; Kruglanski & Webster, 1996; Vohs, Baumeister, & Chin, 2007)? Or as responsible adults who do not needlessly give away something that has not been earned? We believe that both portrayals contain an element of truth. They also point to the same prescription for interacting with skeptical reciprocators: Whether cynics or adults, skeptical reciprocators will eagerly cooperate with someone who effectively demonstrates genuineness.
References


Krajbich, I., Bartling, B., Hare, T., & Fehr, E. (2015). Rethinking fast and slow based on a critique of reaction-time reverse inference. *Nature Communications, 6*:7455. DOI: 10.1038/ncomms8455


Appendix

Model

We begin with the following standard constituents of a finite, two-player game: Associate with each player \( i = 1, 2 \) a space of material outcomes \( X_i \) and a finite collection of strategies \( s_i = \{ s_i^1, s_i^2, \ldots, s_i^{N_i} \} \). Specify a material payoff function \( O: s_1 \times s_2 \rightarrow X_1 \times X_2 \). Let \( \Delta(X_i) \) denote the space of lotteries over \( X_i \). Let \( \Sigma_i \) denote the space of mixed strategies of player \( i \), and extend \( O \) to be from mixed strategies to lotteries accordingly. Let each player have preferences \( \succeq_i^0 \) over \( \Delta(X_i) \).

Our model builds on the important work of Segal and Sobel (2007). These authors study players who, conditional on an anticipated strategy profile \( \sigma^* = (\sigma_i^*, \sigma_j^*) \), have preferences \( \succeq_{i,\sigma^*} \) defined over their own strategies \( \Sigma_i \) (rather than just preferences over material outcomes). Segal and Sobel axiomatically characterize the conditions for representing such preferences with payoff functions that are linear combinations of a player’s expected utility for material outcomes and his counterparts’ expected utility for material outcomes. We situate our model within this framework, and examine players whose preferences give rise to a particular functional form for the total payoff, \( v_{i,\sigma^*}(\sigma_i) \), that player \( i \) receives by playing strategy \( \sigma_i \) in the context of the anticipated strategy profile \( \sigma^* \):

\[
v_{i,\sigma^*}(\sigma_i) = u_i(\sigma_i, \sigma_j^*) + \lambda_i M_{i,\sigma^*}(\sigma_i) M_{ij,\sigma^*}(\sigma_j^*) u_j(\sigma_i, \sigma_j^*). \tag{1}
\]

The players’ total payoffs are the sum of their material and psychological payoffs. Player \( i \)’s material payoff is his expected utility \( u_i(\sigma_i, \sigma_j^*) \). His psychological payoff includes his counterpart’s expected utility for her material payoff, \( u_j(\sigma_i, \sigma_j^*) \), weighted by several parameters. First, \( \lambda_i \geq 0 \) indexes player \( i \)’s other-regard. The greater is \( \lambda_i \), the more player \( i \) cares about reciprocating kindness with kindness and unkindness with unkindness. Second, as we next detail, \( M_{i,\sigma^*} \) and \( M_{ij,\sigma^*} \) are player \( i \)’s assessments of his and his counterpart’s motivations. Combining motivation scores multiplicatively allows for reciprocity.
The motivation scores $M_{i,\sigma^*}$ and $M_{ij,\sigma^*}$ correspond to assessments of kindness or unkindness that are appropriately “discounted” to account for self-interest. The degree to which a player’s behavior is assessed to be kind or unkind depends on whether it materially helps or hurts his counterpart, and on whether it materially helps or hurts himself. For example, a strategy that helps a counterpart seems kind, but it seems less kind if it also helps the player himself. A strategy that hurts a counterpart seems unkind, but it seems less unkind if it also helps the player himself.

Consider player $i$’s assessment of her own motivation, $M_{i,\sigma^*}(\sigma_i)$. Let $p_\sigma(s_i)$ denote the probability that $\sigma_i$ assigns to each pure strategy $s_i$, and let $c_\sigma(s_i)$ denote the set of all $s_i$ for which $p_\sigma(s_i) > 0$. Similarly, let $p_{\sigma^*}(s_j)$ denote the probability that $\sigma^*_j$ assigns to each pure strategy $s_j$, and let $c_{\sigma^*}(s_j)$ denote the set of all $s_j$ for which $p_{\sigma^*}(s_j) > 0$. Next, for each pure strategy profile $\{s_i, s_j\}$ in $c_\sigma(s_i) \times c_{\sigma^*}(s_j)$, let $A_{i,\sigma^*}(s_i, s_j)$ denote the set of all alternative pure strategies $s'_i$ in $s_i$ which yield utilities $u_j(s'_i, s_j)$ such that $u_j(s'_i, s_j) \neq u_j(s_i, s_j)$. With each element of each $A_{i,\sigma^*}(s_i, s_j)$, associate a basic motivation score $b_{i,\sigma^*}(s'_i, s_j)$ given by

$$b_{i,\sigma^*}(s'_i, s_j) = \begin{cases} 
(+1 - \theta_i) & u_i(s_i, s_j) \geq u_i(s'_i, s_j) \text{ and } u_j(s_i, s_j) > u_j(s'_i, s_j) \\
1 & u_i(s_i, s_j) < u_i(s'_i, s_j) \text{ and } u_j(s_i, s_j) > u_j(s'_i, s_j) \\
(-1 + \theta_i) & u_i(s_i, s_j) > u_i(s'_i, s_j) \text{ and } u_j(s_i, s_j) < u_j(s'_i, s_j) \\
-1 & u_i(s_i, s_j) \leq u_i(s'_i, s_j) \text{ and } u_j(s_i, s_j) < u_j(s'_i, s_j)
\end{cases}$$

where the parameter $\theta_i \in [0, 1]$ captures $i$’s degree of skepticism. Denote the cardinality of a set by horizontal bars. Then for $|A_{i,\sigma^*}(s_i, s_j)| > 0$, let the partial motivation score $m_{i,\sigma^*}(s_i, s_j)$ be given by

$$m_{i,\sigma^*}(s_i, s_j) = \frac{\sum_{s'_i \in A_{i,\sigma^*}(s_i, s_j)} b_{i,\sigma^*}(s'_i)}{|A_{i,\sigma^*}(s_i, s_j)|}.$$
If $|A_{i\sigma^*}(s_i, s_j)| = 0$, player $i$ cannot affect his counterpart’s material payoffs. In such cases, set $m_{i,\sigma^*}(s_i, s_j) = 0$. That is, we presume that if a player can neither help nor hurt his counterpart, he cannot be perceived as kind or unkind.

Finally, to compute the motivation score $M_{i,\sigma^*}(\sigma_i)$, the model weights the partial motivation scores by the probability that the mixed strategy profile $\{\sigma_i, \sigma_j^*\}$ assigns to each pure strategy profile $\{s_i, s_j\}$ in $c_i(\sigma_i) \times c_j^* (\sigma_j^*)$:

$$M_{i,\sigma^*}(\sigma_i) = \sum p_{\sigma_i}(s_i) p_{\sigma_j}(s_j) m_{i,\sigma^*}(s_i, s_j).$$

$M_{ij,\sigma^*}(\sigma_j^*)$ captures $i$’s assessment of $j$’s anticipated strategy and is calculated similarly. Let $p_{\sigma_i}(s_i)$ denote the probability that $\sigma_i^*$ assigns to each pure strategy $s_i$, and $p_{\sigma_j}(s_j)$ the probability that $\sigma_j^*$ assigns to each pure strategy $s_j$, and let $c_i^*(\sigma_i^*)$ and $c_j^* (\sigma_j^*)$ denote the sets of all $s_i$ for which $p_{\sigma_i}(s_i) > 0$ and all $s_j$ for which $p_{\sigma_j}(s_j) > 0$, respectively. For each pure strategy profile $\{s_i, s_j\}$ in $c_i^*(\sigma_i^*) \times c_j^* (\sigma_j^*)$, let $A_{j,\sigma^*}(s_i, s_j)$ denote the set of all alternative pure strategies $s_j'$ which yield expected utilities $u_i(s_i, s_j')$ such that $u_i(s_i, s_j') \neq u_i(s_i, s_j)$. With each element of $A_{j,\sigma^*}(s_i, s_j)$, associate a basic motivation score $b_{ij,\sigma^*}(s_j')$ computed in analogy to $b_{i,\sigma^*}(s_i)$ above. For instance, if playing $s_j$ materially helps both players compared to an alternative strategy $s_j'$, the latter is assigned a basic motivation score of $1 - \theta_i$. Note that player $i$’s basic motivation scores for player $j$ feature $\theta_i$ (not $\theta_j$). Again, for $|A_{j,\sigma^*}(s_i, s_j)| = 0$, let $m_{ij,\sigma^*}(\sigma_j^*) = 0$, and for $|A_{j,\sigma^*}(s_i, s_j)| > 0$, let the partial motivation score $m_{ij,\sigma^*}(s_i, s_j)$ be given by

$$m_{ij,\sigma^*}(s_i, s_j) = \frac{\Sigma_{s_j' \in A_{j,\sigma^*}(s_i, s_j)} b_{ij,\sigma^*}(s_j')}{|A_{j,\sigma^*}(s_i, s_j)|}.$$

The motivation score $M_{ij,\sigma^*}(\sigma_j^*)$ that player $i$ assigns to player $j$’s anticipated strategy is then given by
Segal and Sobel (2007) define Nash equilibrium as an anticipated strategy profile \(\sigma^* = (\sigma_i^*, \sigma_j^*)\) in which \(\sigma_i^*\) and \(\sigma_j^*\) are the players’ preferred strategies conditional on \(\sigma^*\), so that neither player has an incentive to unilaterally deviate. We use the term “skeptical equilibrium” to describe those Nash equilibria that arise when preferences can be represented via the total payoffs in Equation 1. Lemma 1 in Segal and Sobel (2007) can be used to show that if players have preferences of the kind we examine, every 2-player game has at least one skeptical equilibrium.

Modeling Approach

Three features of our modeling approach deserve mention. First, note that motivation scores take into account all alternative strategies. In particular, strategies that may appear implausible, such as cooperation in response to defection in a sequential prisoners’ dilemma, can impact motivation scores and thereby influence resulting equilibria. We believe this approach is reasonable, because there may be signaling value in foregoing such strategies. For instance, by foregoing cooperation in response to defection, a second-mover reveals that she is not a pure altruist.

Second, our model does not refine a game’s set of skeptical equilibria by considering how players might revise their assessments of motivations off the equilibrium path. For example, having the first-mover cooperate and the second-mover plan to cooperate in response to either first-move cooperation or first-move defection can be a skeptical equilibrium of the standard, sequential prisoners’ dilemma in Figure 1 (as long as \(\lambda_2\) is sufficiently high, and \(\theta_2\) sufficiently low). In this equilibrium, the second-mover views the first-mover as kind due to his cooperation, and thus reciprocates. Her cooperative response to first-move defection is off the path of play. Were the first-mover to actually defect, it would not be in the spirit of the model for the second-mover to subsequently cooperate. But we simply do not consider this issue. Dufwenberg and Kirchsteiger (2004) provide a history-dependent model that explores...
SKEPTICAL RECIPROCITY AND PRINCIPLED DEFECTION

the relevant considerations. Future work could aim to combine the skepticism in our model with the history-dependence in theirs.

Third, as discussed in the text, the special case of \( \theta_i = \theta_j = 0 \) constitutes a model of strong reciprocity. In that case, motivation scores depend only on whether players help or hurt one another relative to their alternative strategies.

Positive Reciprocity in the Standard, Sequential Prisoners’ Dilemma

Consider the game in Figure 1. Let \( s^* = \{C; cd\} \) be the anticipated strategy profile. That is, the first-mover is expected to cooperate, and the second-mover is expected to reciprocate both first-move cooperation and first-move defection (here and in any strategy profile that follows, upper-case letters denote first-move actions, the first lower-case letter denotes the second-move action in response to cooperation, and the second lower-case letter the second-move action in response to defection). We next show that if the second-mover is sufficiently other-regarding (\( \lambda_2 \) is sufficiently large) and not overly skeptical (\( \theta_2 \) is sufficiently small), then \( s^* \) is a skeptical equilibrium.

Begin by calculating \( M_1 \) and \( M_{21} \). For convenience, assume that the function from dollars to material utilities is simply the identity for both players. The first-mover plays \( C \), which given \( s^* \) yields each player a material payoff of 4. He could deviate to \( D \), which given \( s^* \) would yield each player a material payoff of 2. By playing \( C \) rather than \( D \), the first-mover is thus helping himself (4 > 2) and his counterpart (4 > 2). Furthermore, \( D \) is the only possible deviation from \( s^* \) by which he could alter the players’ material payoffs. So \( M_1 = 1 - \theta_1 \), and \( M_{21} = 1 - \theta_2 \). Next, calculate \( M_2 \) and \( M_{12} \). The second-mover has two possible deviations by which she could alter the players’ material payoffs: She could deviate to \( dd \) or \( dc \). Either deviation yields her a material payoff of 6 and the first-mover a material payoff of 0. Relative to these deviations, her planned strategy of \( cd \) helps the first-mover (4 > 0) but hurts her (4 < 6). Both are thus assigned partial motivation scores of +1, so \( M_2 = M_{12} = (1 + 1) / 2 = 1 \).
Each player’s total payoff under \( s^* \) is thus \( v_i = 4 + 4\lambda_i(1 - \theta_i) \). For \( s^* \) to be a skeptical equilibrium, each player’s total payoff from playing \( s_i^* \) must be at least as great as her total payoff were she to unilaterally deviate to some alternative strategy. For the first-mover, the only alternative strategy is \( D \). If the second-mover is playing her part of \( s^* \) (i.e., \( cd \)), then by playing \( D \) instead of \( C \), the first-mover would hurt both himself (2 < 4) and his counterpart 2 (2 < 4). So deviating to \( D \) yields a motivation score \( M_1 = -1 \), and a total payoff \( v_1 = 2 + 2\lambda_1(-1) \). Since \( \lambda_1 > 0 \), the first-mover has no incentive to deviate.

The second-mover can deviate to either \( dd \) or \( dc \), both of which would lead to material payoffs of 0 for the first-mover and 6 for the second-mover. Relative to \( s^* \), these deviations would hurt the first-mover (0 < 4) and help the second-mover (6 > 4), so they yield \( M_2 = -1 + \theta_2 \) and a total payoff \( v_2 = 6 + 0\lambda_2(-1 + \theta_2)^2 \). The second-mover would thus be willing to play her part of \( s^* \), that is \( cd \), as long as \( 4 + 4\lambda_2(1 - \theta_2) \geq 6 \), which reduces to \( \lambda_2(1 - \theta_2) \geq \frac{1}{2} \).

In sum, \( \{C; cd\} \) is a skeptical equilibrium only if the second-mover is sufficiently other-regarding and not overly skeptical. The more skeptical the second-mover (i.e., the larger \( \theta_2 \)), the more difficult it is for the strategy profile \( \{C; cd\} \) to be an equilibrium.

**Positive Reciprocity in the De-Coupled Prisoners’ Dilemma**

Consider the de-coupled prisoners’ dilemma corresponding to Figure 1. We now show that the conditions under which the second-mover will prefer to reciprocate cooperation in this game impose weaker restrictions on \( \lambda_2 \) and \( \theta_2 \) than the \( \{C; cd\} \)-equilibrium in the standard, sequential game.

Because the first-mover believes that the second-mover cannot affect the players’ material payoffs, \( M_{12} \) is equal to zero, and there is no possible psychological payoff for the first-mover. In the following, we therefore focus on the second-mover.

The second-mover assesses the first-mover’s motivation relative to a degenerate game in which the first-mover chooses between material payoffs of (2, 0) and (0, 4). In this degenerate game, given any
context of play, the first-mover choosing (2, 0) engenders $M_{21} = -1 + \theta_2$ and the first-mover choosing (0, 4) engenders $M_{21} = +1$.

In assessing her own motivation score, on the other hand, the second-mover takes into account the complete game. Cooperating in response to cooperation yields $M_2 = +1$ and a total payoff $v_2 = 4 + 4\lambda_2$. Defection in response to cooperation yields $M_2 = +1 - \theta_i$ and a total payoff $v_2 = 6 + 0\lambda_2(1-\theta_i)$. The second-mover will thus reciprocate cooperation as long as $4 + 4\lambda_2 \geq 6$, which reduces to $\lambda_2 \geq \frac{1}{2}$. Recall that the corresponding condition in the standard, sequential game had $\lambda_2 (1-\theta_2) \geq \frac{1}{2}$.

That is, the presence of skepticism ($\theta_2 > 0$) makes second-mover positive reciprocity more likely in the decoupled game than in the corresponding standard, sequential game. In contrast, under strong reciprocity ($\theta_2 = 0$), the second-mover will reciprocate cooperation in either game as long as $\lambda_2 > \frac{1}{2}$.

**Negative Reciprocity in the Standard, Sequential Prisoners’ Dilemma**

We next show that $s^* = \{D; dd\}$ is a skeptical equilibrium of the standard, sequential game for any values of $\lambda_2$ and $\theta_2$.

Mutual defection yields each player a material payoff of 2. The first-mover could deviate to $C$, which given $s^*$ would yield material payoffs of 0 for him and 6 for the second-mover. By playing $D$ rather than $C$, the first-mover is thus helping himself ($2 > 0$) and hurting his counterpart ($2 < 6$). Since $C$ is the only possible deviation from $s^*$ by which he could alter the players’ material payoffs, it follows that $M_1 = -1 + \theta_1$, and $M_{21} = -1 + \theta_2$. The second-mover has two possible deviations by which she could alter the players’ material payoffs: she could deviate to $dc$ or $cc$. Both lead to material payoffs of 0 for her and 6 for the first-mover. Her planned strategy of $dd$ therefore hurts the first-mover ($2 > 6$) but helps her ($2 > 0$), so $M_2 = -1 + \theta_2$, and $M_{12} = -1 + \theta_1$.

Because $M_1$, $M_{12}$, $M_2$, and $M_{21}$ are all negative, the products $M_1M_{12}$ and $M_1M_{21}$ are both positive. Both players garner positive psychological utilities under $s^* = \{D; dd\}$. In this light, it is easy
to verify that $s^*$ is an equilibrium. Any unilateral deviation would lower the deviating player’s material utility to zero. It would also yield the player a positive motivation score. Given the counterpart’s negative motivation score, the deviating player would then have negative psychological utility. In other words, any unilateral deviation would lower both a player’s material and psychological payoff.

*Negative Reciprocity in the De-Coupled Prisoners’ Dilemma*

While the de-coupled game should yield more positive reciprocity than the standard, sequential game, the two games should yield equivalent negative reciprocity. Recall that by choosing the material payoffs $(2, 0)$ rather than $(0, 4)$, decoupled game first-movers engender $M_{2f} = -1 + \theta_2$. A second-mover faced with a first-move choice of $(2, 0)$ in the de-coupled game thus finds herself in exactly the same situation as a second-mover faced with first-move defection in the standard, sequential game.

*A Formalization of the Endogenous Sequencing Prisoners’ Dilemma*

Suppose there are $t = 1, \ldots, T$ discrete time periods, with $t = 1$ denoting the beginning of the countdown, and $t = T$ its end. We formalize the game in discrete time because our model applies to finite two-player games. Specify each player’s strategy $s_i$ by (a) a first-move action $f_i$ in $\{C; D\}$, (b) a time $t_i$ at which to make the first move (where $t_i = T$ means that $i$ actually does not make a first move before the timer elapses, in which case $f_i$ corresponds to the action the player would take in an eventual simultaneous game), and (c) two reaction vectors $r^{c}_i$ and $r^{d}_i$, each of length $T$ and with all entries in $\{c; d\}$, that indicate player $i$’s response to first move cooperation (defection) by player $j$ in each time period. We refer to the player with the smaller $t_i$ as the first-mover. Assume that when the players have equal $t_i$, which player becomes the actual first-mover is resolved by a fair coin flip. The players’ material payoffs are determined by the first-mover’s action $f_1$ and the corresponding entry in the second-mover’s relevant reaction vector. Material payoffs are not affected by the timer and are the same as in the standard, sequential game in Figure 1.
A Reciprocating Equilibrium in the Endogenous Sequencing Prisoners’ Dilemma

Consider a strategy profile $s^*$ with $f_1 = f_2 = C$, $t_1 = t_2 = t^*$, and both $r^*_i = r^*_j = c$ and $r^{d}_i = r^{d}_j = d$ for all time periods. We now examine the conditions under which such strategy profiles can be skeptical equilibria. Two important observations emerge. First, psychological payoffs in such strategy profiles are maximal when both players opt to first-move cooperate immediately (i.e., when $t_1 = t_2 = 1$). Second, “fast” skeptical equilibria of this type support a wider range of parameters $\theta_2$ and $\lambda_2$ than the $\{C, cd\}$-equilibrium in the standard, sequential game. For instance, when $t_1 = t_2 = 1$, $s^*$ is a skeptical equilibrium for $\lambda_2 \geq \frac{1}{2}$, no matter how skeptical the players are.

We begin by calculating the total payoff obtained by each player. Each player’s material payoff is 4. To calculate psychological payoffs, examine possible pure strategy deviations from $s^*$. There are four possible sets of deviations for player $i$:

1. deviations to early first-move defection, in which $f_i = D$ and $t_i < t^*$,
2. deviations to early first-move cooperation, in which $f_i = C$, and $t_i < t^*$,
3. deviations to defection on cooperation, in which the $t^*$-th element of $r^c_i$ is $d$, and $t_i > t^*$, and
4. deviations to cooperation on defection, in which $r^{d}_i = c$ at some $t_i$.

(There are also deviations in which the first-mover is randomly chosen because both players act at $t = t^*$, but $i$ first-move defects, or defects on cooperation, or both. Their omission below does not affect our conclusions.)

Since only (1) and (3) alter the players’ material payoffs compared to $s^*$, only these sets of deviation affect the players’ motivation scores. Deviations in (1) lead to material payoffs of 2 for each player. Compared to these strategies, player $i$ helps both herself and player $j$ by playing $s^*$, yielding partial motivation scores of $+1 - \theta_i$. Deviations in (3), on the other hand, lead to a material payoff of 6 for player $i$ and 0 for player $j$. Compared to these strategies, player $i$ is hurting herself to help player $j$ by
playing \( s^* \), yielding partial motivation scores of +1. To compute \( M_i \) and \( M_{ij} \), we examine how many alternative strategies there are of each type.

Consider deviations of type (1), to early first-move defection. There are \( t^* - 1 \) time periods before period \( t^* \), and \( i \) could first-move defect in any of them. Because \( i \)'s first-move action \( f_i = D \) would then drive the outcome of the game, his response vectors \( r_{i}^{c} \) and \( r_{j}^{c} \) are unconstrained in these strategies of first-move defection. There are \( 2^T \) different combinations of response vectors. Overall, there are thus \( (t^* - 1)2^T \) unique strategies by which player \( i \) could deviate to first-move defection.

Next, consider deviations of type (3), to defection on cooperation. To ensure that \( j \) moves first, \( i \) must delay his first-move action until after \( t^* \). There are \( T-t^* \) time periods after \( t^* \); \( i \) would be free to first-move defect or first-move cooperate in any of them (as \( i \)'s first-move action would not be binding). The response vectors \( r_{i}^{c} \) and \( r_{j}^{c} \) are unconstrained with the exception of \( r_{i}^{c} \) in time period \( t^* \), which leaves \( 2^{2T-1} \) different combinations of response vectors. Overall, there are thus \( 2(T-t^*)2^{2T-1} = (T-t^*)2^{2T} \) unique strategies by which player \( i \) could deviate to defection on cooperation.

This allows us to calculate \( M_i \) and \( M_{ij} \) and hence the total payoffs for each player:

\[
v_{i,s^*} = 4 + \lambda_i \left[ \frac{(1-\theta \cdot (t_{i}^* - 1) + (T-t_{i}^*))}{(T-1)} \right] \left[ \frac{(1-\theta \cdot (t_{i}^* - 1) + (T-t_{i}^*))}{(T-1)} \right] 4
\]

Each time period before player \( i \)'s first move at \( t_{i}^* \) thus contributes \((1-\theta) \) to \( j \)'s total motivation score, and each time period after the first move at \( t_{i}^* \) contributes +1 (and vice versa for \( t_{j}^* \)). As a result, psychological payoffs decrease in \( t_{i}^* \) and \( t_{j}^* \). The intuition for \( M_i \) and \( M_{ij} \), explained in greater detail in the text, is that a player’s cooperative first move seems more genuine the earlier it occurs because it is tantamount to renouncing any attempt to take advantage of potential first-move cooperation by the counterpart.

For \( s^* \) to be a skeptical equilibrium, each player \( i \)'s total payoff from playing \( s_i^* \) must be at least as great as her total payoff from her alternative strategies given \( s^* \). We next compute the total payoffs of
SKEPTICAL RECIPROCITY AND PRINCIPLED DEFECTION

first-move defection and of delayed defection on cooperation (i.e., the two sets of alternative strategies described above).

Relative to first-move defection, player \( i \) has three possible types of deviations: first-move cooperation before \( t^* \), delaying her first move until after \( t^* \) and reciprocating cooperation with cooperation, and delaying her first move until after \( t^* \) and defecting on cooperation. The partial motivation scores assigned to these three types of deviations are \(-1\), \(-1\), and \(+1\), respectively, and accounting for the number of unique strategies of each type yields a motivation score of \((1 - t^*) / (2T - t^* - 1)\) for \( i \)'s first-move defection, and of \((-1 + \theta_i)\) for delayed defection on cooperation. Both of these motivation scores are less than or equal to zero, so first-move defection yields a lower total payoff than delayed defection on cooperation. Consequently, a strategy profile \( s^* \) as defined above is a skeptical equilibrium for all parameter values \( \lambda_i \) and \( \theta_i \) for which the total payoff exceeds 6 (the total payoff of delayed defection on cooperation).

Because the motivation scores \( M_i \) and \( M_{ij} \) decrease in \( t_j^* \) and \( t_i^* \), respectively, faster first moves lead to higher total payoffs. Total payoffs are maximal if both players first-move cooperate immediately \((t_1 = t_2 = 1)\), in which case \( v_{i,s^*} = 4 + \lambda_i4 \). This strategy profile supports a wider range of parameter values for \( \lambda_2 \) as a skeptical equilibrium than \{C, cd\} in the standard, sequential game. Moreover, the payoffs no longer depend on \( \theta_i \): It can be an equilibrium no matter how skeptical the players are.
Table 1. Motivation scores for a player’s strategy as a function of how it impacts her counterpart as well as the player herself.

<table>
<thead>
<tr>
<th></th>
<th>Help Self</th>
<th>Hurt Self</th>
</tr>
</thead>
<tbody>
<tr>
<td>Help Other</td>
<td>$+1 - \theta_i$</td>
<td>+1</td>
</tr>
<tr>
<td>Hurt Other</td>
<td>$-1 + \theta_i$</td>
<td>-1</td>
</tr>
</tbody>
</table>
Table 2. Summary statistics for the standard, sequential and de-coupled games in Experiment 1.

<table>
<thead>
<tr>
<th></th>
<th>1st-mover C</th>
<th>2nd-mover reciprocity of C with C</th>
<th>2nd-mover reciprocity of D with D</th>
<th>2nd-mover reciprocity overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Sequential</td>
<td>34/45 (75.6%)</td>
<td>60/83 (72.3%)</td>
<td>21/27 (77.8%)</td>
<td>81/110 (73.6%)</td>
</tr>
<tr>
<td>De-Coupled</td>
<td>20/39 (51.3%)</td>
<td>40/47 (85.1%)</td>
<td>33/47 (70.2%)</td>
<td>73/94 (77.7%)</td>
</tr>
</tbody>
</table>
Table 3. Summary statistics for the standard, sequential and endogenous-sequencing games in Experiment 2.

<table>
<thead>
<tr>
<th></th>
<th>1&lt;sup&gt;st&lt;/sup&gt;-mover C</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt;-mover reciprocity of C with C</th>
<th>2&lt;sup&gt;nd&lt;/sup&gt;-mover reciprocity of D with D</th>
<th>Simultaneous game if timer expired, C</th>
<th>Median // Mean time elapsed before 1&lt;sup&gt;st&lt;/sup&gt;-move (% of total time)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Sequential</td>
<td>32/47 (68%)</td>
<td>22/32 (69%)</td>
<td>16/17 (94%)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Endogenous-Sequential, Aggregated</td>
<td>54/70 (77%)</td>
<td>40/52 (77%)</td>
<td>14/17 (82%)</td>
<td>1/7 (14%)</td>
<td>N/A (15% // 25%)</td>
</tr>
<tr>
<td>Endogenous-Sequential, 20 Seconds</td>
<td>17/21 (81%)</td>
<td>10/15 (67%)</td>
<td>4/5 (80%)</td>
<td>0/2 (0%)</td>
<td>6.0 // 8.2 seconds (30% // 41%)</td>
</tr>
<tr>
<td>Endogenous-Sequential, 60 Seconds</td>
<td>21/29 (72%)</td>
<td>17/20 (85%)</td>
<td>6/8 (75%)</td>
<td>1/5 (20%)</td>
<td>9.0 // 10.8 seconds (15% // 18%)</td>
</tr>
<tr>
<td>Endogenous-Sequential, 120 Seconds</td>
<td>16/20 (80%)</td>
<td>13/17 (76%)</td>
<td>4/4 (100%)</td>
<td>N/A</td>
<td>13.0 // 24.0 seconds (11% // 20%)</td>
</tr>
</tbody>
</table>
Table 4. Positive reciprocity in the standard, sequential game and in games of various speeds under endogenous sequencing (as measured by the % of the countdown elapsed prior to a first-move) in Experiment 2. The statistical tests in the two right-hand columns compare the positive reciprocity rate under endogenous sequencing with that of the standard, sequential game.

<table>
<thead>
<tr>
<th>2nd-movers responding to 1st-mover C:</th>
<th>1st-mover C</th>
<th>2nd-mover reciprocity of C with C</th>
<th>Fisher’s Exact p-value vs. seq. game</th>
<th>Logistic Regression: $b_{\text{Game}} (SE), p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Sequential</td>
<td>32/47 (68%)</td>
<td>22/32 (69%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 5% of Countdown Elapsed Prior to 1st-Move</td>
<td>8/11 (72.7%)</td>
<td>9/9 (100%)</td>
<td>.083</td>
<td>N/A</td>
</tr>
<tr>
<td>≤ 10% of Countdown Elapsed Prior to 1st-Move</td>
<td>22/25 (88.0%)</td>
<td>21/22 (95%)</td>
<td>.019</td>
<td>2.36 (1.27) p = .064</td>
</tr>
<tr>
<td>≤ 15% of Countdown Elapsed Prior to 1st-Move</td>
<td>30/35 (85.7%)</td>
<td>27/30 (90%)</td>
<td>.061</td>
<td>1.99 (1.06) p = .060</td>
</tr>
<tr>
<td>≤ 20% of Countdown Elapsed Prior to 1st-Move</td>
<td>35/40 (85.6%)</td>
<td>31/34 (91%)</td>
<td>.031</td>
<td>2.12 (1.04) p = .041</td>
</tr>
<tr>
<td>&gt; 20% of Countdown Elapsed Prior to 1st-Move</td>
<td>19/30 (63.3%)</td>
<td>9/18 (50%)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Figure 1. A sequential prisoners’ dilemma.
Figure 2. First-movers’ behavior in the standard, sequential (left) and endogenous-sequencing (right) games in Experiment 2.
Figure 3. Second-movers’ responses to first-mover cooperation, in the standard, sequential (left) and endogenous-sequencing (right) games in Experiment 2.
Supplementary Materials: Experimental Instructions

**Experiment 1**, Standard, sequential game: Screen 1

**INTRODUCTION**
In this study, you have been paired with another participant here today.

One of you will be Participant 1, and one of you will be Participant 2. Which role you have will be randomly determined by the computer in a few moments.

Participant 1 will be asked to make a choice first. Participant 2 will be notified of Participant 1’s choice. Then Participant 2 will be asked to make a choice. Your choices will impact how much money each of you leaves the study with. We will pay you both in cash at the end of today’s session.

Please enter your SONA ID number in the field below.

[Field for SONA ID number]

Move to the next screen

Screen 2

**SET-UP**
We will explain the details of this study step-by-step.

To start off, each of you has been credited with $2.

![Currency notes]
**SKEPTICAL RECIPROCITY AND PRINCIPLED DEFECTION**

**Screen 3**

**YOUR DECISION**
Each participant can either **KEEP** his or her $2 or **GIVE** his or her $2 to the other participant.

![Images of 'Keep' and 'Give']

Importantly, if a participant decides to give his or her $2 to the other participant, the other participant will receive both that $2 as well as an additional, extra $2.

**NOTIFYING PARTICIPANT 2**
Participant 1 will make his or her choice first. Participant 2 will then be notified whether Participant 1 chose to keep his or her $2, or give the $2. At that point, Participant 2 will make his or her choice.

Because Participant 1 moves first, his or her choice may influence Participant 2's choice.

**Move to the next screen**

**Screen 4**

**REVIEW**
Participant 1 and Participant 2 each have to decide whether to KEEP or to GIVE.

If a participant **keeps** his or her $2, then he or she receives these $2. The other participant is notified and receives $0.

If a participant **gives** the $2, then he or she receives $0. The other participant is notified and receives a total of $4 (the $2 given to him or her plus the additional, extra $2).

Finally, as we mentioned, because Participant 1 decides first, his or her choice may influence Participant 2's choice.

**Move to the next screen**
COMPREHENSION TEST

Suppose Participant 1 keeps his or her $2 and Participant 2 keeps his or her $2. Then…
- Participant 1 will get a total of $4, and Participant 2 will get a total of $4.
- Participant 1 will get a total of $2, and Participant 2 will get a total of $2.
- Participant 1 will get a total of $0, and Participant 2 will get a total of $6.

Suppose Participant 1 gives his or her $2 and Participant 2 gives his or her $2. Then…
- Participant 1 will get a total of $4, and Participant 2 will get a total of $4.
- Participant 1 will get a total of $2, and Participant 2 will get a total of $2.
- Participant 1 will get a total of $0, and Participant 2 will get a total of $6.

Suppose Participant 1 gives his or her $2 and Participant 2 keeps his or her $2. Then…
- Participant 1 will get a total of $4, and Participant 2 will get a total of $4.
- Participant 1 will get a total of $2, and Participant 2 will get a total of $2.
- Participant 1 will get a total of $0, and Participant 2 will get a total of $6.

Which of the following is correct?
- Before deciding to keep or give, Participant 2 will find out what Participant 1 chose to do.
- Before deciding to keep or give, Participant 2 will NOT find out what Participant 1 chose to do.

Move to the next screen

ROLE ASSIGNMENT
The computer has randomly determined that you will be Participant 2, and your counterpart will be Participant 1.

By going on to the next screen, you indicate that you are ready— the study will begin as soon as everybody is ready. If anything about the study or these instructions is unclear to you, please raise your hand and ask the experimenter for an explanation.

I'm ready to start!
Screen 7f

Please decide whether to keep your $2 or give your $2 to Participant 2. We will then notify Participant 2 of your decision.

Keep  give

Keep your $2  Give your $2

Screen 8s

Participant 1 has decided to give his or her $2 to you. Please decide whether to keep your $2 or give your $2 to Participant 1.

Keep  give

Keep your $2  Give your $2
Screen 9s (choice-first)

For second-movers in the ratings-first condition, “Your decision has been recorded.” on the first line is replaced by “Participant 1 has decided to KEEP (GIVE) his or her $2 (to you).”

Your decision has been recorded.

Now please assess Participant 1’s motives in GIVING rather than KEEPING:

By GIVING rather than KEEPING, to what extent do you think...

Participant 1 is **BEING GENUINELY NICE?**

Not at all  ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ A lot

Participant 1 is **CALCULATING HIS/HER SELF-INTEREST?**

Not at all  ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ A lot

Move to the next screen
Experiment 1, De-Coupled game: Screen 1

**INTRODUCTION**
In this study, you have been paired with another participant here today.

One of you will be Participant 1, and one of you will be Participant 2. Which role you have will be randomly determined by the computer in a few moments.

Participant 1 will be asked to make a choice that impacts how much money each of you leaves the study with. Participant 2 will be notified of Participant 1’s choice. We will pay you both in cash at the end of today’s session.

Please enter your SONA ID number in the field below.

Move to the next screen

Screen 2

**SET-UP**
We will explain the details of this study step-by-step.

To start off, Participant 1 will be credited with $2.

Move to the next screen
SKEPTICAL RECIPROCITY AND PRINCIPLED DEFLECTION

Screen 3

PARTICIPANT 1’S DECISION
Participant 1 can either KEEP his or her $2 or GIVE his or her $2 to Participant 2.

Importantly, if Participant 1 decides to give his or her $2 to Participant 2, Participant 2 will receive both that $2 as well as an additional, extra $2.

NOTIFYING PARTICIPANT 2
Participant 2 will be notified of whether Participant 1 chose to KEEP his or her $2, or GIVE the $2.

Move to the next screen

Screen 4

REVIEW
If Participant 1 keeps his or her $2, then he or she receives these $2. Participant 2 is notified and receives $0.

If Participant 1 gives the $2, then he or she receives $0. Participant 2 is notified and receives a total of $4 (the $2 given to him or her plus the additional, extra $2).

Move to the next screen
SCREEN 5

COMPREHENSION TEST
Suppose Participant 1 keeps his or her $2. Then...
- Participant 1 will get $4, and Participant 2 will get $4.
- Participant 1 will get $2, and Participant 2 will get $2.
- Participant 1 will get $2, and Participant 2 will get $0.
- Participant 1 will get $0, and Participant 2 will get $4.

Suppose Participant 1 gives his or her $2. Then...
- Participant 1 will get $4, and Participant 2 will get $4.
- Participant 1 will get $2, and Participant 2 will get $2.
- Participant 1 will get $2, and Participant 2 will get $0.
- Participant 1 will get $0, and Participant 2 will get $4.

Which of the following is correct?
- Participant 2 will find out whether Participant 1 elected to keep or give his or her $2.
- Participant 2 will NOT find out whether Participant 1 elected to keep or give his or her $2.

Move to the next screen

SCREEN 6f

ROLE ASSIGNMENT
The computer has randomly determined that you will be Participant 1, and your counterpart will be Participant 2.

By going on to the next screen, you indicate that you are ready-- the study will begin as soon as everybody is ready. If anything about the study or these instructions is unclear to you, please raise your hand and ask the experimenter for an explanation.

I'm ready to start!
SKEPTICAL RECIPROCITY AND PRINCIPLED DEFECTION

Screen 6s

ROLE ASSIGNMENT
The computer has randomly determined that you will be Participant 2, and your counterpart will be Participant 1.

By going on to the next screen, you indicate that you are ready—the study will begin as soon as everybody is ready. If anything about the study or these instructions is unclear to you, please raise your hand and ask the experimenter for an explanation.

Screen 7f

Please decide whether to keep your $2 or give your $2 to Participant 2. We will then notify Participant 2 of your decision.

Keep your $2  Give your $2
Screen 8s (second-mover, choice-first)

Participant 1 has decided to GIVE his or her $2 to you.

Move to the next screen

Screen 9s

SURPRISE CHOICE FOR PARTICIPANT 2
We are now crediting you, Participant 2, with a new $2.

Participant 1 chose to give his or her $2 to you. Now, you will have to decide whether to KEEP the new $2 or GIVE the new $2 to Participant 1.

Thus, like Participant 1 you will make a decision that impacts how much money you and Participant 1 each leave the study with.

Move to the next screen
YOUR DECISION
You will have to either KEEP your $2 or GIVE your $2 to Participant 1.

Keep  give

Importantly, if you decide to give your new $2 to Participant 1, he or she will receive both the new $2 as well as an additional, extra $2.

NOTIFYING PARTICIPANT 1
After you decide to KEEP or GIVE, Participant 1 will be notified. We will let Participant 1 know that we gave you this surprise choice about a new $2. We will also let him/her know what you decided to do.

COMPREHENSION TEST
As a reminder, Participant 1 decided to give his or her initial $2 to you.

Suppose you keep the new $2. Then counting both Participant 1’s and Participant 2’s decisions, at the end of the session...
  √ Participant 1 will get a total of $4, and you as Participant 2 will get a total of $4.
  √ Participant 1 will get a total of $2, and you as Participant 2 will get a total of $2.
  √ Participant 1 will get a total of $6, and you as Participant 2 will get a total of $6.

Suppose you give the new $2 to Participant 1. Then counting both Participant 1’s and Participant 2’s decisions, at the end of the session...
  √ Participant 1 will get a total of $4, and you as Participant 2 will get a total of $4.
  √ Participant 1 will get a total of $2, and you as Participant 2 will get a total of $2.
  √ Participant 1 will get a total of $6, and you as Participant 2 will get a total of $6.

Move to the next screen
Screen 12s

As a reminder, Participant 1 decided to give his or her $2 to you. Please decide whether to keep your new $2 or give your new $2 to Participant 1.

Keep  give

Keep your $2  Give your $2

Screen 13s (second-mover, choice first)

For second-movers in the ratings-first condition, this screen replaces Screen 8s, and the single line on Screen 8s “Participant 1 has decided…” replaces “Your decision has been recorded.”

Your decision has been recorded.

Now please assess Participant 1’s motives in GIVING rather than KEEPING:

By GIVING rather than KEEPING, to what extent do you think ....

Participant 1 is BEING GENUinely NICE? Not at all  ☐ ☐ ☐ ☐ ☐ ☐ ☐  A lot
Participant 1 is CALCULATING HIS/HER SELF-INTEREST? Not at all  ☐ ☐ ☐ ☐ ☐ ☐ ☐  A lot

Move to the next screen
Experiment 2, Standard, sequential game: Screen 1

INTRODUCTION
In this study, you have been paired with another participant here today.

The two of you will each be asked to make a decision. How much money each of you leaves the study with will be determined by these decisions. We will pay you in cash at the end of the study.

SET-UP
To start off, each of you has been credited with $2.

Please enter your participant number in the field below.

Move to the next screen

Screen 2

We will explain the details of this study step-by-step.

YOUR DECISION
Each participant can either KEEP his or her $2 or GIVE his or her $2 to the other participant.

Keep give

Importantly, if a participant decides to give his or her $2 to the other participant, the other participant will receive both that $2 as well as an additional, extra $2.

Move to the next screen
SKEPTICAL RECIPROCITY AND PRINCIPLED DEFECTION

Screen 3

PROCEDINGS
You and the other participant will decide on keeping versus giving one after the other. Who goes first and who goes second will be determined randomly by the computer.

GOING FIRST
If you go FIRST, you will decide whether to keep your $2 or give your $2 to the other participant before (s)he makes his or her decision.

NOTE: The other participant will be notified of whether you keep or give before making his or her own decision. So your decision might influence the other participant’s decision.

GOING SECOND
If you go SECOND, you will be notified of the other participant’s decision before deciding whether to keep your $2 or give your $2 to the other participant.

Move to the next screen

Screen 4

EXAMPLE
Suppose the other participant goes first and gives his or her $2 to you. You go second and keep your $2.

Then you will leave the study with a total of $6: your $2 that you have kept, plus the $2 the other participant has given to you, plus the additional extra $2.

The other participant will leave the study with $0: (s)he has given his or her $2 to you, whereas you have kept your $2.

COMPREHENSION TEST
Suppose you go first and give your $2 to the other participant, and the other participant goes second and keeps his or her $2.

Then how much cash will you leave the study with? [ ]

How much cash will the other participant leave the study with? [ ]

Move to the next screen
A SECOND EXAMPLE
Suppose you go first and give your $2 to the other participant. The other participant goes second, and also gives his or her $2 to you.

Then you will leave the study with a total of $4: the $2 the other participant has given to you, plus an additional extra $2.

Likewise, the other participant will also leave the study with a total of $4: the $2 you have given to him or her, plus another additional extra $2.

COMPREHENSION TEST
Suppose you go first and keep your $2, and the other participant goes second and also keeps his or her $2.

Then how much cash will you leave the study with?  
How much cash will the other participant leave the study with?  

Move to the next screen

ORDER
The computer has randomly determined that you will go FIRST, and the other participant will go second.

COMPREHENSION TEST
When making a decision on keeping or giving...  
- ...the other participant decides, I am notified of his or her decision, and then I decide;  
- ...I decide, the other participant is notified of my decision, and then (s)he decides.

By moving on to the next screen, you can indicate that you are ready-- the study will begin as soon as both you and the participant you have been paired with are ready. If anything about the study or these instructions is unclear to you, please raise your hand and ask the experimenter for an explanation.

I'm ready to start!
Screen 7f

Please decide whether to keep your $2 or give your $2 to the other participant.

Keep  give
Keep your $2  Give your $2

Screen 8s

The other participant has decided to give his or her $2 to you.

Keep  give
Keep your $2  Give your $2
**Experiment 2**: Endogenous-sequecing game, 20-second timer, Screen 1

**INTRODUCTION**
In this study, you have been paired with another participant here today. The two of you will each be asked to make a decision. How much money each of you leaves the study with will be determined by these decisions. We will pay you in cash at the end of the study.

**SET-UP**
To start off, each of you has been credited with $2.

Please enter your participant number in the field below.

[Enter participant number]

**Screen 2**

We will explain the details of this study step-by-step.

**YOUR DECISION**
Each participant can either **KEEP** his or her $2 or **GIVE** his or her $2 to the other participant.

Keep    give

Importantly, if a participant decides to give his or her $2 to the other participant, the other participant will receive both that $2 as well as an additional, extra $2.
A SECOND EXAMPLE
Suppose you give your $2 to the other participant, and the other participant gives his or her $2 to you.

Then you will leave the study with a total of $4: the $2 the other participant has given to you, plus an additional extra $2.

Likewise, the other participant will also leave the study with a total of $4: the $2 you have given to him or her, plus another additional extra $2.

COMPREHENSION TEST
Suppose you keep your $2, and the other participant also keeps his or her $2.

Then how much cash will you leave the study with?  
How much cash will the other participant leave the study with?

Move to the next screen

EXAMPLE
Suppose you keep your $2, and the other participant gives his or her $2 to you.

Then you will leave the study with a total of $6: your $2 that you have kept, plus the $2 the other participant has given to you, plus the additional extra $2.

The other participant will leave the study with $0: (s)he has given his or her $2 to you, whereas you have kept your $2.

COMPREHENSION TEST
Suppose you give your $2 to the other participant, and the other participant keeps his or her $2.

Then how much cash will you leave the study with?  
How much cash will the other participant leave the study with?

Move to the next screen
TIMING
You will also have to think about the timing of your decision. Once the two of you are ready, a 20-second timer will begin counting down. At any point during these 20 seconds, each of you can (a) hold off on making a decision, i.e., WAIT, (b) KEEP your $2, or (c) GIVE your $2.

WAIT vs. Keep or give

NOTIFICATION
You will be notified if and when the other participant makes a decision, and you will be told whether (s)he has kept his or her $2 or has given it to you.

By waiting, you may get the chance to see the other participant's decision before you make your decision. On the other hand, by moving first, you might influence the other participant's decision. So the timing issue is about who makes the first move.

WHAT IF THE TIMER RUNS OUT
If the 20-second timer runs out, we will require both you and the participant with whom you have been paired to decide whether to keep or give your $2 without knowing each other's decision.

COMPREHENSION TEST
If the timer runs out...
- The study is over.
- We each have to decide on giving or keeping without knowing each other's decision.

By moving on to the next screen, you can indicate that you are ready— the study will begin as soon as both you and the participant you have been paired with are ready. If anything about the study or these instructions is unclear to you, please raise your hand and ask the experimenter for an explanation.
SKEPTICAL RECIPROCITY AND PRINCIPLED DEFECTION

Screen 7

NOTIFICATION STATUS
The other participant is still waiting and has not decided yet whether to keep his or her $2 or give it to you.

00:20

Keep give

Keep your $2 Give your $2

Screen 8

The other participant has decided to give his or her $2 to you.

Keep give

Keep your $2 Give your $2