Principled Defection:

On Caring that Fails to Activate and Non-Cooperative Behavior

Johannes Müller-Trede

IESE Business School

University of Navarra

Yuval Rottenstreich

Rady School of Management

University of California, San Diego

Abstract

Many theories identify selfishness and lack of caring about others as the fundamental impediments to cooperation. We highlight a different source of non-cooperative behavior: People's caring can be abundant but fail to activate. We present an attribution-based, game-theoretic model in which people may defect rather than cooperate even if they place little weight on self-interest, place much weight on reciprocating others, and recognize that they have been treated well. In the model, people consider others' motives and often perceive two broad possibilities. Someone may treat another individual well out of genuine kindness or out of tactical self-interest, hoping to elicit and profit from a reciprocal response. The model formalizes the notion that when people construe positive treatment they receive as "just business" (i.e., motivated by tactical self-interest), their caring remains dormant, and they do not reciprocate. When they interpret positive treatment they receive as genuinely kind, their caring is activated, may be substantial, and they may reciprocate. We term non-reciprocity engendered by attributions of tactical motives "principled defection," and we experimentally corroborate its prevalence. Our work indicates that existing research underestimates people's taste for reciprocity. It yields novel perspectives on generosity in ultimatum games and unraveling in finitely-repeated interactions. It stands in opposition to the influential social heuristics hypothesis; a synthesis of our analysis and related theorizing on the norm of self-interest, reactive egoism, and sophisticated social inference offers an alternative explanation for findings cited as support for social heuristics.

Keywords: Cooperation, Reciprocity, Attribution, Social Inference, Self-Interest, Social Preferences, Social Heuristics

Principled Defection: On Caring that Fails to Activate and Non-Cooperative Behavior

Elliott and Michelle are acquaintances who work as freelance web developers. Each of them is about to commence a valuable project. They happen to cross paths. In talking, they realize that Elliott's project would be of even greater value to Michelle, and vice versa. They also realize that each of them could hand their project off to the other individual. A few days later, Elliott contacts Michelle and gives his project to her. She accepts it and works on it but also keeps her own project. In other words, though Elliott gives up his project so that Michelle can have it, she retains her project for herself.

In this article, we are inspired by the question of what explains Michelle's non-cooperative response. Extant accounts of cooperation and reciprocity provide a straightforward answer. They highlight the role of self-interest. By these accounts, after Elliott has sacrificed to treat her positively, Michelle's decision not to answer in kind must be selfish.

There are two classes of such accounts. Inspired by biological theories of genetic fitness (Trivers, 1971), early accounts explain behavior entirely on the basis of self-interest (Axelrod & Hamilton, 1981; Kreps, Milgrom, Roberts, & Wilson, 1982; see also Delton, Krasnow, Cosmides, & Tooby, 2011; Rand et al., 2014). They assert that people will only bear costs to help others if there is a tactical reason to do so. For instance, if giving her project to Elliott could encourage him to give her additional projects in the future, Michelle might reconsider and transfer her project to him. Without such inducements, Michelle will retain her project.

Newer accounts often invoke both selfish and social motivations. Many different social motivations guide behavior, including altruism (Campbell, 1972; Hoffman, 1981; Krebs, 1970), justice (Brockner & Wiesenfeld, 1996; Lerner, 1982; Thibaut & Walker, 1975; Tyler, 1994), and the focus of our work, reciprocity (Campbell, 1975; Gouldner, 1960; Keysar, Converse, Wang,

& Epley, 2008; Komorita, Parks, & Hulbert, 1992; Pruitt, 1968; Regan, 1971). When people are socially motivated, they can genuinely value treating others well; they are not pro-social solely in tactical pursuit of their own self-interest (Caporael, Dawes, Orbell, & Van de Kragt, 1989; Fehr, Fischbacher, & Gächter, 2002; Fehr & Gintis, 2007; Gintis, 2000; Penner, Dovidio, Piliavin, & Schroeder, 2005; Yamagishi, Li, Takagishi, Matsumoto, & Kioynari, 2014). Indeed, the newer accounts posit a tension: They emphasize social motivations as a driver of cooperative behavior and self-interest as an obstacle to it. For instance, if Michelle appreciates Elliott's giving his project to her, she may wish to reciprocate by giving her project to him. If she instead gives pronounced precedence to her own payoffs, she will abstain from reciprocating.

It is notable that despite their vast differences, both classes of accounts identify selfishness and a lack of caring about others as the fundamental impediments to cooperation. In early accounts, people are uninterested in helping others; they help only if ulterior motives warrant doing so. In newer accounts, people do not help if their self-interest outweighs their social motivation.

We consider a different source of non-cooperative behavior. In what follows, we introduce and experimentally examine a game-theoretic model of reciprocity that builds on intuitive assumptions about people's attributions of each other's motives. The model allows for a tension between selfish and social motivations. But it recognizes that a person may defect on cooperation even if she places little weight on self-interest, places much weight on reciprocity, and understands that her counterpart has treated her well. Thus, the model suggests that in many situations the fundamental impediment to cooperation is not that people are self-interested and lacking in social motivation. Instead, it highlights an antecedent factor: People's social motivation, which may be abundant, may fail to activate.

In our analysis, people consider their counterparts' motives and often see two broad possibilities. These possibilities roughly correspond to the patterns emphasized by each class of extant accounts. Someone may act positively toward another individual because he genuinely cares about how he treats her, reflecting a social motivation like altruism, justice, or reciprocity. He may also act positively out of tactical self-interest, hoping to elicit and profit from her subsequent reciprocal cooperation (Everett, Pizarro, & Crockett, 2016; Jordan, Hoffman, Nowark, & Rand, 2016; Kelley & Stahelski, 1970). Elliott, for example, may transfer his project to Michelle to help her—or because he hopes doing so will prompt her to transfer her project to him. Our model formalizes the notion that people frequently do not reciprocate positive behavior that might be attributable to tactical self-interest. They feel differently about someone attempting to maximize his own interests compared to someone driven by how he treats others or by a concern for their welfare (Bigman & Tamir, 2016; Blount, 1995; Brehm & Cole, 1966; McCabe, Rigdon, & Smith, 2005; McCabe & Smith 2000; Schopler & Thompson, 1968). For example, if Michelle interprets Elliott's behavior as an attempt to acquire her project, she may not see him as meriting her reciprocity.

Put another way, if Michelle construes Elliott's behavior as "just business" rather than genuinely kind, if she attributes his cooperation to tactics, then her potentially abundant social motivation to treat him well may not be activated. From this perspective, the critical situational factor shaping Michelle's behavior is not that keeping Elliott's project is materially beneficial to her. It is that Elliott did not establish that he is socially motivated. If he instead were able to establish that he is trying to help Michelle rather than influence her, her own caring would be activated, and she might well reciprocate.

We use the term "principled defection" for non-reciprocation of positive treatment that is engendered by attributions of a counterpart's potentially tactical behavior. By characterizing

such behavior as principled, we contrast it with defection driven by what amounts to self-prioritization or greed. Our experiments corroborate the prevalence of principled defection, by showing that when a person's cooperation is unambiguously genuine rather than tactical, reciprocity rates are relatively high. Based on these results, we argue that the fundamental impediment to cooperation is often untapped caring, not self-interest.

Our work complements Miller's (1999) argument that people frequently act non-cooperatively not because they are truly self-interested, but because they believe social norms mandate self-interest (see also Epley & Dunning, 2000; Fetchenhauer & Dunning, 2010; Heath, 1999; Kiesler, 1966; Markle, 2011; Miller & Ratner, 1998; Ratner & Miller, 2001; Vohs, Baumeister, & Chin 2007; Saito, 2015). Our argument is that people frequently act non-cooperatively not because they are selfish and uncaring, but because they do not see a place for their caring when others' behavior may not be driven by caring.

Furthermore, our work may be contrasted with research showing that expectations of others' non-cooperative intent can underlie one's own non-cooperation. Kerr (1983; see also Orbell & Dawes, 1981; Yamagishi & Soto, 1986) reported that people who anticipate free riding from collaborators on a shared task often reduce their own effort, because they are afraid of feeling like a "sucker." In studies of reactive egoism, Epley, Caruso, and Bazerman (2006; see also Croson, 2000; Halevy, 2017; Kennedy & Pronin, 2008; Pierce, Kilduff, Galinsky, & Sivanathan, 2013) found that explicitly considering a counterpart's perspective can accentuate the tendency to anticipate self-interested behavior and to preemptively engage in such behavior oneself. Note, however, that if Elliott gives Michelle his project, he has unquestionably benefited her. In considering Michelle's response, we are thus highlighting someone who has received positive treatment, not someone who anticipates non-cooperative treatment. Accordingly, we stress attributions rather than negative expectations.

The literature in social psychology is replete with evidence and theories of cynical, skeptical, and suspicious attributions (e.g., Bem, 1967; Epley & Dunning, 2000; Fein, 1996; Fein, Hilton, & Miller, 1990; Ham & Vonk, 2011; Hilton, Fein, & Miller, 1993; Jones, Davis, & Gergen, 1961; Kruger & Gilovich, 1999). To be clear, our contribution does not lie in the observation that people are sensitive to ambiguous motives; rather, by formalizing this notion and pursuing its implications, we come to several insights about cooperation and reciprocity that have not been evident through the lens of prior theories.

First, to reiterate, we highlight that contrary to numerous theories, the fundamental impediment to cooperation is often not selfishness and lack of caring. It is untapped social motivation. In essence, we make a distinction between the strength and activation of social motivation. By recent theories, non-reciprocation of positive behavior arises when relatively weak social motivation is trumped by relatively pronounced self-interest. We offer an additional possibility: Non-reciprocation can arise when behavior construed as "just business" fails to activate people's potentially strong social motivation.

Second, a vast literature that spans the social and biological sciences employs experimental and simulated games to investigate pro-social behavior (for reviews, see Fehr & Gächter, 2000; Malmendier, te Velde, & Weber, 2014). This literature has generated a stark dichotomy of views about reciprocity. One prominent stream asserts that people have an inherent taste or preference for reciprocity (Fehr, Fischbacher, & Gächter, 2002; Bowles & Gintis, 2002; Gintis 2000). Another prominent stream suggests that people have little interest in reciprocity per se, especially positive reciprocity (Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Offerman, 2002; Nowak, 2006; Rand & Nowak, 2013). Understanding the impact of attributions can inform the debate between these views. In particular, our analysis suggests several ways in which existing research may underestimate people's taste for reciprocity.

Third, our analysis also yields new perspectives on several intensely-studied phenomena that involve pro-social behavior. One such phenomenon is generosity in ultimatum games (Camerer & Thaler, 1995; Güth & Tietz, 1990; Oosterbeek, Sloof, & Van de Kuilen, 2004). Another is the tendency for cooperation to unravel over the course of finitely-repeated interactions (Axelrod & Hamilton, 1981; Andreoni & Miller, 1993; Chater, Vlaev, & Grinberg, 2008; Erev & Roth, 2002; Kreps et al., 1982; Morehous, 1966; Rapoport & Chammah, 1965).

Fourth, our model is consistent with theories that emphasize deliberative and sophisticated aspects of social inference (Fein, 1996; Gilbert, Krull, & Pelham, 1988; Gilbert, Pelham, & Krull, 1988; Trope 1986; Trope & Gaunt, 1999, 2000; see also Lieberman, Gaunt, Gilbert, & Trope, 2001). Moreover, our data buttress such perspectives: Many participants in our experiments do not fall prey to correspondence errors; they do not necessarily construe cooperation as indicative of social motivation. Yet, they are sensitive to cues that enable them to become convinced of a counterpart's social motivation, and they reciprocate on that basis (cf. Barclay & Willer, 2007; Inesi, Gruenfeld, & Galinsky, 2012; Roberts, 1998). Our work thus stands in opposition to the influential social heuristics hypothesis, which casts cooperation and reciprocity as products of simple, automatic thinking (Rand 2016; Rand, Greene, & Nowak, 2012; Rand et al., 2014). Indeed, we later present a synthesis of our model, dual process theories of attribution, Epley et al.'s (2006) work on reactive egoism, and Miller's (1999) norm of self-interest. This synthesis provides an alternative explanation for findings cited as support for the social heuristics hypothesis. It also predicts several patterns of behavior that run counter to the social heuristics hypothesis.

From Here

The remainder of this article is organized as follows. We begin by introducing our model and distinguishing it from alternative game-theoretic accounts of reciprocity. We do so via a set of games which can give rise to pro-social behavior. In some cases, that behavior may be unambiguously attributed to social motivation. In others, it could reflect material self-interest and is thus ambiguously motivated. With this set of games, only our model accommodates an intuitive pattern of behavior that includes principled defection in response to ambiguously-motivated pro-social behavior. Moreover, only our model accommodates a closely related pattern that emerges when a positive response to pro-social behavior is itself ambiguously-motivated. Experiments 1 and 1A corroborate our model's unique predictions.

We subsequently return to Elliott and Michelle. Their interaction roughly corresponds to a sequential prisoners' dilemma, perhaps the quintessential setting for studying cooperation and reciprocity and hence a natural domain in which to consider principled defection. Accordingly, Experiments 2 and 3 each contrast a standard, sequential prisoners' dilemma with a variant of this dilemma. In the standard game, the first-mover (Elliott) cannot establish his social motivation by cooperating, which gives room for principled defection by the second-mover (Michelle). In each of the two variants, an altered situational factor enables a cooperating first-mover to establish his social motivation, so principled defection is less relevant. Comparisons between the standard game and the variants therefore yield estimates of the prevalence of principled defection. Like Experiment 1, Experiments 2 and 3 suggest that principled defection is common.

Model

Our models builds on the classic attributional notion of discounting (Einhorn & Hogarth, 1986; Fishbein & Ajzen, 1975; Jones, 1979; Kelley 1973; Nisbett & Ross, 1980). We presume

that a person's inclination to reciprocate depends on her judgment of the kindness or unkindness of her counterpart's behavior, and that she moderates her judgment whenever his behavior is materially profitable for him. This set-up allows our model to explain, for example, that a counterpart's helpful or generous behavior may not be fully kind, because however helpful or generous it is, it may reflect tactics rather than genuine caring.

Mathematically, we build on a framework introduced by Segal and Sobel (2007). We formally state the model in the Appendix. Here, we explicate its workings via the games in Figures 1 and 2. Together, Figures 1A, 1B, and 1C demonstrate a pattern that features principled defection in response to ambiguously-motivated pro-social behavior. Figures 2A and 2B illustrate a closely related pattern that emerges from ambiguous motives on the part of the individual responding to pro-social behavior.

Figures 1A and 1B depict *handoff games*, which we designed by drawing on related examples in Rabin (1993, Example 6) and Dufwenberg and Kirchsteiger (2004; Game G7). They parallel situations in which one individual can opt into or out of a relationship with another, but if he opts in, the second individual controls how the relationship works out. An individual may be working on a project, for instance, and could retain the project and continue working on it. Or, alternatively, he could hand it off to someone who will increase its overall return. If the second person is given the project, she can conduct it in a way that respects the initial person's concerns and is generous in giving him credit. Or she could do the opposite, ignore his concerns and be miserly about crediting him.

In Figure 1B, the first-mover can unilaterally garner $3 and not engage with the second-mover, who as a result would garner $0. Or, the first-mover can hand off control to the second-mover. At that point, the second-mover has the opportunity to split $7 between them. She can act generously, granting $5 to the first-mover and $2 to herself, or miserly, granting the reverse

payoffs. In Figure 1A, the first-mover's unilateral option is worth $6 to him; all else remains as in Figure 1B.

Figure 1C is a simplified dictator game (Camerer & Thaler, 1995; Dana, Weber, & Kuang, 2006; Kahneman, Knetsch, & Thaler, 1986). In this game, the erstwhile first-mover does not make a decision. The erstwhile second-mover simply selects between the generous and miserly splits of $7.

Motivations and Motivation Scores

Suppose the individuals in Figure 1B anticipate that the first-mover will hand off control to the second-mover, and that the second-mover will respond with generosity. How will they assess the motivation underlying the second-mover's strategy?

The model takes the anticipated strategy profile as given and generates an assessment in three steps. First, it identifies all alternative pure strategies that would change at least one individual's material payoff. In the game of Figure 1B, there is just one alternative: The second-mover could be miserly. Second, it 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. The second-mover's generosity belongs in the upper-right cell: Relative to the alternative, it materially "helps" the other player, the first-mover, because he receives $5 and would receive less, $2, if the second-mover were miserly. Furthermore, relative to the alternative, the second-mover's strategy materially "hurts" the second-mover herself. She receives $2 but would receive $5 by defecting. Third, based on this classification, the model assigns the strategy a "motivation score" ranging from +1, which corresponds to unambiguous kindness, to –1, which corresponds to unambiguous unkindness. Because the second-mover has forsaken a greater material payoff to improve the lot of the first-mover, her generosity is deemed

a +1. Thanks to her sacrifice, in the model her social motivation stands undoubted. She could not have been motivated by material self-interest.

What about the first-mover's decision to hand off control? Our analysis suggests that in Figure 1B, this strategy can be viewed questioningly. First, the only alternative has him act unilaterally. Second, handing off control yields the players \$5 and \$2, respectively, rather than the \$3 and \$0 they would receive if the first-mover acted unilaterally. So while the first-mover materially "helps" the second-mover (\$2 > \$0), he also materially "helps" himself (\$5 > \$3). His handoff is thus classified in the upper-left cell. Third, on that basis, it receives a motivation score of $1 - \theta_i$, where θ_i is an individual-specific discounting parameter satisfying $0 \leq \theta_i \leq 1$. It dilutes assessments of the first-mover's kindness (or unkindness) to account for the possibility that he is acting out of calculated self-interest rather than genuine social motivation. Greater θ_i reflects greater discounting by individual i.[1]

The remainder of Table 1 is similarly derived. If a player sacrifices to reduce a counterpart's material payoff, her strategy is classified into the lower-right and scored –1 (cf. Fehr & Gächter, 2002; Guala, 2012). Because she effectively paid to hurt the other player, the strategy is unambiguously mean. On the other hand, reducing a counterpart's material payoff while increasing one's own receives a motivation score of $-1 + \theta_i$. Because such behavior may be driven by self-interest, it is not seen as inherently mean. A second-mover's miserly response to a first-mover's handoff provides one example. It materially hurts the first-mover (relative to

---

1 In the game of Figure 1B, each player has 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 comparing the anticipated strategy with every alternative pure strategy. This aspect of the model surfaces in Experiment 2 and is important in generating predictions in Experiment 3. We discuss it in detail in that context.

the alternative of generosity). But it also maximizes the second-mover's own material payoff, so her intentions need not be unkind. In sum, θ dilutes attributions of both kindness and unkindness.[2]

Player i's total utility v_i is given by $v_i = u_i + \lambda_i u_j M_i M_{ij}$. Here, u_i is the player's utility for his material outcome. The latter term reflects his utility from reciprocity or the lack thereof. It 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 social motivation; the greater is λ_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 kind, a player's total utility rises as she is increasingly kind in return. Likewise, if a counterpart is unkind, a player's total utility rises as she is increasingly unkind in return.

Our model may be applied to any two-player game with finite action sets. Moreover, the total utility functions we specify satisfy a set of conditions identified by Segal and Sobel (2007) that guarantee the existence of at least one Nash equilibrium. 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 utility. Many games will, of course, permit multiple equilibria. For instance, if the second-mover's λ is

---

2 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 and Kirchsteiger (2004) in viewing mixed strategies as reflecting incomplete information about population behavior and not as an individual's conscious decision to randomize (see also Segal & Sobel, 2007).

sufficiently large and her θ is sufficiently small, then the game of Figure 1B supports an equilibrium in which the first-mover hands off to the second-mover, who then reciprocates with generosity. A unilateral action by the first-mover that is met with miserliness by the second-mover, however, also form an equilibrium, for any values of λ and θ.

Altering the Unilateral Action Impacts Attributions and thus the Activation of Social Motivation

How does increasing the first-mover's incentive for unilateral action from $3 (Figure 1B) to $6 (Figure 1A) impact attributions, motivation scores, and behavior? If the players continue to anticipate a handoff and ensuing generosity, the second-mover's material payoff is unaffected. It remains $2. But the first-mover's social motivation is now signaled more clearly: By handing off control, he ends up with $5 rather than $6. He thus now hurts himself to help the second-mover, and his motivation score is +1 rather than $1 - \theta$. Because foregoing the handoff in favor of unilateral action is materially more profitable than receiving generous behavior, a handoff cannot be motivated by tactical self-interest and must instead reflect genuine caring. The second-mover should therefore be more likely to reciprocate with generosity in Figure 1A than Figure 1B.

The possibility of greater second-mover generosity in Figure 1A than 1B highlights the failure to activate social motivation as a fundamental impediment to pro-social behavior. In Figure 1B, where the first-mover's motivation score for opting-in is only $+1 - \theta$, even a highly socially-motivated second-mover (i.e., someone who has large λ) may discount enough (via sufficiently large θ) to effectively construe the first-mover's handoff as "just business." She may then respond with miserliness, which constitutes principled defection. In Figure 1A, where the first-mover's motivation score is +1, the second-mover can be assured that the first-mover's handoff is not driven by mere material self-interest and is not "just business." Instead, it is

driven by social motivation. The second-mover may then respond with generosity. In this way, a non-zero discounting parameter θ allows for people who act non-cooperatively not because they are selfish and uncaring, but because they do not see a place for their caring when others' behavior may not be driven by caring.

Note that under the special case of $\theta = 0$ and no discounting, the model reduces to a pure selfish-social tension. It essentially assumes that social motivation is always activated, so that behavior is determined by the relative strength of self-interest versus social motivation, as captured by λ. This special case does not allow for different behavior across Figures 1A and 1B. A second-mover with small λ is relatively more selfish and will be miserly in both games; a second-mover with large λ is relatively less selfish and will be generous in both games.

As we have mentioned, there are clearly additional social motivations beyond a desire for reciprocity. With that in mind, consider the simplified dictator game in Figure 1C, which eliminates the erstwhile first-mover's decision. Because he has no chance to act, he does not have an opportunity to treat the erstwhile second-mover positively and cannot earn any reciprocal treatment. He may still benefit, however, from any other social motivations that drive his counterpart. Two such motivations may be particularly pertinent. His counterpart may be altruistic, willing to engage in self-sacrificing, generous behavior that is not conditioned on any antecedent or expected behavior by a counterpart (Fehr & Fischbacher, 2003; Fletcher & Zwick, 2007; Krebs, 1970). Or his counterpart could act generously because she believes she is supposed to behave that way, expects that the first-mover or the experimenters may judge her on that basis, and is averse to being judged negatively (Dana, Cain, & Dawes, 2006). In sum, in Figure 1C, the erstwhile second-mover's choice between generous and miserly behavior may be seen as tapping a baseline level of social motivation. A taste for reciprocity will add to this baseline. Our model therefore makes the intuitive, overall prediction that as we move from

Figure 1C to 1B to 1A, and first-movers have an increasingly attractive unilateral option, second-movers should become increasingly likely to act generously.

Comparisons with Extant Game-Theoretic Accounts of Reciprocity

The games in Figure 1 illustrate how our model identifies three distinct types of situations, only one of which can give rise to principled defection. In Figure 1C, the erstwhile first-mover cannot impact the second-mover. He is therefore judged neutrally and as not meriting reciprocity. In Figure 1A, by handing off, the first-mover helps the second-mover and hurts himself. He is thus judged as unquestionably kind and fully deserving of reciprocity. In Figure 1B, by handing off, the first-mover helps the second-mover but also helps himself. His motives are therefore ambiguous. Consequently, his kindness is discounted, and he is viewed as intermediately deserving of reciprocity, which opens the door to principled defection.

Extant game-theoretic accounts of reciprocity recognize that self-interest can render motivations ambiguous and that kindness judgments should reflect this fact. But in contrast to our model, the foundation of these accounts is not the psychology of attribution. Thus, instead of discounting, they rely on other methods for handling ambiguity (Rabin, 1993; Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006; see also Sobel, 2005). A key drawback of these methods is that, unlike discounting, they essentially resolve rather than maintain ambiguity. In our model, people maintain divergent hypotheses: A counterpart's behavior could be self-interested or genuinely caring. They balance these hypotheses by judging behavior intermediately deserving of reciprocity. However, in extant accounts, people must commit to one hypothesis and dismiss the other. That is, any behavior is interpreted as either entirely undeserving or fully meriting reciprocity; there is no intermediate option. As we next detail, this

feature makes it difficult for extant accounts to accommodate some commonplace patterns that include principled defection.

In Rabin's (1993) theory, anyone who stands to gain materially from his own actions is deemed fully undeserving of positive reciprocity, irrespective of the effects his actions have on others. Put in our terms, behavior that helps a counterpart but also helps oneself is judged entirely and definitively self-interested. Not surprisingly, and as Rabin (1993, p. 1296) himself points out, this approach fares poorly in many settings. In our handoff games, it implies that a first-mover forgoing $3 is perceived as entirely undeserving of positive reciprocity, just like a passive, erstwhile first-mover who does not get a chance to act. Thus, Rabin's (1993) theory implies that second-movers will act equivalently across Figures 1B and 1C, because any generosity by second-movers in either figure must reflect altruism or some other social motive rather than reciprocity.

Dufwenberg and Kirchsteiger's (2004) theory views a person as fully deserving of positive reciprocity only if he both helps his counterpart and his chosen actions expose him to a potential material loss relative to his alternative actions. Thus, a first-mover who foregoes $3 to choose a handoff is cast as fully deserving of positive reciprocity. He both helps the second-mover by handing off and leaves himself vulnerable to receiving only $2 if the second-mover is miserly. The theory views a person as fully undeserving of positive reciprocity if his actions do not leave him exposed. To illustrate, consider a modified handoff game in which the outside option is only worth $1 to the first-mover. In that setting, a first-mover who hands off is still helping the second-mover, but he will do better than $1 no matter how the second-mover responds. Put in our terms, the theory deems some behaviors that help a counterpart as entirely socially motivated and other behaviors that help a counterpart as entirely self-interested. This approach can also fare poorly. In Figure 1, it implies that first-movers forgoing $3 and $6 are

judged equivalently. In both cases, a hand off leaves the first-mover vulnerable to receiving the lesser $2 payout and thus fully merits reciprocity. Thus, Dufwenberg and Kichsteiger's (2004) theory implies that second-movers will behave equivalently across Figures 1A and 1B.[3]

Finally, Falk and Fischbacher's (2006) theory of reciprocity and social comparisons focuses on considerations of equity. In the theory, a person is judged as kind if he provides his counterpart with as much or more than he himself receives. He is judged as unkind if he provides his counterpart with less than he himself receives. Judgments are tempered if an individual cannot influence the players' relative standing. For instance, someone who has no choice but to provide his counterpart with more than he himself receives is viewed as less kind than someone who could have taken the greater share for himself. This set-up distinguishes

---

[3] Mathematically, both Rabin (1993) and Dufwenberg and Kirchsteiger (2004) employ notions of efficiency to handle ambiguous motivations. Loosely speaking, only strategies that satisfy a version of Pareto efficiency "count" in assessing a person's motives. Using Pareto efficiency in such a way is ingenious. It allows these theories to implicitly account for whether someone benefits from his own actions, even though they score an individual's motivation (or "kindness") solely on the basis of the material payoff his actions make available to his counterpart. This is because within the set of Pareto efficient strategies, giving a counterpart more implies taking less for oneself.

There are, however, two fundamental limitations to these theories' efficiency-based approach. First, it is not sensitive to changes in a counterpart's material payoffs that do not alter the set of efficient strategies. Second, it rules out any psychological payoffs whenever there is only a single strategy in the efficient set. Our model is not susceptible to these limitations. Experiment 1 concerns the first limitation. Experiment 1A will concern the second limitation.

between Figure 1C and the two handoff games, because the erstwhile first-mover in Figure 1C cannot affect the players' outcomes and is therefore judged neutrally. Contingent on the players' relative standing being held constant, however, it is beyond the scope of the theory to distinguish behaviors that do or do not improve a player's own outcome compared to his alternatives. In other words, Falk and Fischbacher's (2006) theory does not address ambiguous motivation. It therefore makes equivalent predictions across Figures 1A and 1B.

In sum, by resolving or not addressing ambiguity rather than recognizing and maintaining it, extant accounts do not allow for the possibility that a person is somewhat but not fully deserving of reciprocity. They thus do not adequately account for commonplace patterns of behavior that implicate principled defection, like increasing second-mover generosity as a first-mover obtains increasingly attractive unilateral options. To reiterate, in the games in Figure 1, Rabin's (1993) theory equates Figures 1B and 1C, and Dufwenberg and Kirchsteiger's (2004) and Falk and Fischbacher's (2006) theories equate Figures 1A and 1B. Only our model predicts that the frequency of second-mover generosity in Figure 1B will fall in between that of Figures 1A and 1C. We test this prediction in Experiment 1.

Experiment 1: Handoff Games

Method

Participants. 542 undergraduates at UCSD's Rady School of Management participated in our experiment (46.5% female, M_{age} = 21.0; 12 participants did not provide demographic information). We recruited native English speakers who had not previously taken part in any related task. Participants received course credit, and their choices were incentivized with the monetary payoffs shown in Figure 1.

Power calculations based on a pilot study suggested that a sample size of approximately 500 would be appropriate; we scheduled weekly batches of experimental sessions until we exceeded this target. At the end of the experiment, we asked participants two simple comprehension questions. Almost all participants (519, or 95.8%) answered both questions correctly; only a single participant answered both questions incorrectly. Mean performance on the comprehension questions did not differ significantly across conditions ($p = .61$). We did not exclude any participants from the analyses below; our results and their statistical significance are unchanged if only participants who correctly answered both questions are considered.

Procedure. We conducted sessions of up to 22 participants in a large room with 32 private computer stations. Participants were seated in an arrangement that maximized physical separation. Our study was the first of several, unrelated tasks they completed.

We initially randomly assigned sessions to either of the handoff games (Figures 1A and 1B). In addition, because our hypotheses concern second-mover generosity, we ran more second-movers than first-movers. We randomly assigned participants to a role via a procedure that (i) grouped up to three second-movers with each first-mover, while (ii) distributing participants as evenly as possible over the smallest possible number of groups. For example, a session with eight participants would yield two groups of four (each composed of a first-mover and three second-movers), whereas a session with nine participants would yield three groups of three (each composed of one first-mover and two second-movers). Once we obtained a sufficient sub-sample for the handoff games, we turned to sessions that instantiated the dictator game in Figure 1C. In these sessions, all participants played as erstwhile second-movers.

Materials. The study was programmed in z-Tree (Fischbacher, 2007). This software displayed the instructions on each participant's computer screen while they were read aloud by

the experimenter. After the instructions had been read, participants wore headphones to prevent them from hearing each other's vocal reactions and typing noises.

The first-mover's unilateral action in the handoff games was framed as "opting-out" and "receiving $3" (Figure 1B) or "receiving $6" (Figure 1A) for himself, while the second-mover would be "receiving $0." The handoff was framed as "opting-in," in which case the second-mover would choose between the divisions of $7 in Figure 1. Second-movers in all three games (Figure 1A-C) were asked to "choose between two possible divisions," described as "$5 for the first-mover, and $2 for the second-mover" or vice versa. None of our instructions included the terms unilateral, handoff, control, generous, or miserly.

After participants made their decisions and were informed of their payout, they provided demographic information and answered the comprehension questions. The entire sequence of instructions, in the form of screenshots, is in the Supplemental Materials.

Results and Discussion

Not surprisingly, more first movers handed off control (i.e., opted-in) given the middling $3 unilateral action in Figure 1B than the more attractive $6 unilateral action in Figure 1A, (53/66, or 80.3% vs. 29/50, or 58.0%), $\chi^2(1, 116) = 6.83$, $p < .01$. If first-movers who eschew $6 are socially motivated to spare the second-mover from receiving nothing, we may infer that roughly one quarter of those who eschew $3 are driven by tactics (because [.80 – .58] / .80 = .28). That is, roughly one in four Figure 1B first-movers may hand off control because they hope to elicit generosity from the second-mover and accrue $5 for themselves.

More critically, consider second-movers. As uniquely predicted by our model, more of them act generously as we move from Figure 1C, across 1B, to 1A. In Figure 1C, in which their counterpart is passive, only 28.7% (39/136) of erstwhile second-movers act generously. In

Figure 1B, where they respond to an ambiguously-motivated first-mover, 48.0% (60/125) of second-movers act generously. This increase in generosity is significant $\chi^2(1, 261) = 10.33$, $p < .01$. Finally, in Figure 1A, where they respond to an unambiguously socially-motivated first-mover, 63.1% (41/65) of second-movers act generously. This additional increase in generosity is also significant, $\chi^2(1, 190) = 3.90$, $p = .048$.

To calculate the observed rate of principled defection, note that from Figure 1B to 1A, second-mover miserliness falls from 52.0% to 36.9%. Eliminating ambiguity about a first-mover's motives thus decreases the frequency of second-mover non-reciprocation by between 29% (since [.52 – .37] / .52 = .29). In other words, between a quarter and a third of the observed non-reciprocity may not reflect selfishness or lack of caring. It may reflect social motivation that fails to activate when a counterpart's behavior can be construed as "just business." To be sure, this estimate is subject to complex sampling error; it is hardly precise. Nevertheless it provides an initial indication that principled defection may be prevalent.

Positive Reciprocity with Ambiguously-Motivated Responses

Experiment 1 distinguished our model from extant theories, by investigating a setting in which only our model accommodates a pattern that includes principled defection in response to ambiguously-motivated pro-social behavior. We now further distinguish our model, by showing the flip side: Only our model allows for positive reciprocity when a response to pro-social behavior is itself ambiguously-motivated.

Consider Figure 2A. Like Figures 1A and 1B, it depicts a Handoff game in which someone holding a project can retain it for himself (to garner $4 while his counterpart receives nothing). Or he can give it to his counterpart. In a qualitative departure from Figures 1A and 1B, however, the counterpart does not decide between responding generously or miserly.

Instead, she must consider whether to dedicate herself to the project. If she does, the project will go well, and both individuals will profit (receiving $1 and $3, respectively). If she does not, then it will go poorly, and neither individual will accrue any profit.

Critically, if she is handed the project, the responder's motivations for dedicating herself to it are ambiguous. Relative to the alternative, her dedication materially aids the person who gave her the project (he will garner $1 rather than $0). But it is also materially profitable for her (she garners $3 rather than $0). It could be construed as either socially motivated or "just business," and it therefore receives a motivation score of $+1 - \theta$. In contrast, recall that by acting generously rather than miserly, the second-mover in Figures 1A and 1B sacrificed to help the first-mover, so her motivation was unambiguous and scored +1.

Despite the ambiguity surrounding the responder's dedication, our model suggests that the two individuals' interaction could include a handoff, ensuing dedication, and resultant feelings of positive reciprocity. Such an interaction can form an equilibrium as long as neither person discounts too strongly. In our view, this is eminently sensible: We believe that settings akin to Figure 2A provide fertile ground for positive reciprocity. More generally, we believe it is exceedingly common for reciprocally kind interactions to feature responses that are ambiguous rather than purely socially motivated.

Comparisons with Extant Game-Theoretic Accounts of Reciprocity

Extant accounts unfortunately do not allow for positive reciprocity when a responder's motives are ambiguous. This property also stems from their resolving rather than maintaining ambiguity. Recall that in Rabin's (1993) theory, any action that could be ambiguously motivated is instead deemed entirely self-interested. But if dedication is viewed as self-interested in Figure 2A, then reciprocal kindness is not possible. Dufwenberg and Kirchsteiger (2004) view anyone

whose actions do not leave her materially vulnerable as undeserving of reciprocity. But a strategy of dedication insulates the responder in Figure 2A against any material loss: Whether or not she eventually receives the project, she does not fare any worse if she plans to be dedicated than if she plans otherwise. Finally, in Falk and Fischbacher's (2006) theory, because dedication provides the responder with higher earnings than the other individual, it renders her unkind and precludes positive reciprocity.

Experiment 1A

To corroborate the presence of positive reciprocity given ambiguously-motivated responders, we had participants play either Figure 2A or 2B. We relied on the same participant pool and procedures as in Experiment 1, with the exception that we now matched each second-mover with multiple first-movers, since our principal hypothesis concerned the latter. Screenshots from the experiment may be found in the Supplemental Materials.

The instructions mentioned neither a project nor dedication. They stated that the initial individual could "keep" an entire $4 endowment for himself and thereby end the game with his counterpart receiving nothing. Or, he could "divide" the endowment, assigning $1 to himself and $3 to his counterpart. In the Figure 2A condition, the responder could either "accept" or "reject" the $1/$3 division. If she accepted it, each individual would receive their assigned cash amount. If she rejected it, neither individual would receive any money. The dictator game in Figure 2B served as a control condition. The erstwhile responder in that condition was passive. She did not have an accept/reject decision to make. Whatever the initial individual selected was implemented.

By our analysis, many social motivations, including the pursuit of reciprocity, may drive someone to select divide over keep in Figure 2A. But reciprocity is not among the social

motivations that can arise in Figure 2B, because it requires an active responder (just as it requires an active first-mover in Experiment 1). Our analysis therefore accords with more frequent division in Figure 2A than 2B. In contrast, because they do not allow for reciprocity given ambiguous responses, extant accounts do not accord with such a pattern.

Consistent with our analysis, only 29.3% of participants (24/82) divided in Figure 2B, whereas 44.9% (40/89) did so in Figure 2A, $\chi^2(1, 171) = 4.48$, $p = .034$. Virtually all responders in Figure 2A accepted (23/25 = 92.0%).

In sum, by highlighting attributions and the discounting of ambiguous behavior, our model can accommodate two intuitive patterns of behavior that extant theories cannot. The first includes principled defection in response to ambiguously-motivated pro-social behavior. The second emerges when a positive response to pro-social behavior is itself ambiguously-motivated. Experiments 1 and 1A corroborate both patterns. Furthermore, by indicating that principle defection is prevalent, Experiment 1 provides initial evidence of untapped social motivation as a fundamental impediment to cooperative behavior.

Principled Defection and Prisoners' Dilemmas

We now return to Elliott and Michelle. As we have mentioned, their interaction roughly corresponds to a sequential prisoners' dilemma, which is perhaps the quintessential setting for studying cooperation and reciprocity. Accordingly, Experiments 2 and 3 each contrast a standard, sequential prisoners' dilemma with a variant of it to derive additional estimates of the prevalence of principled defection. In addition, our findings in Experiment 2 lead to a discussion of why existing research may underestimate people's taste for reciprocity, while Experiment 3 serves as a lead-in to an alternative explanation for findings cited as support for the social heuristics hypothesis (Rand 2016; Rand, Greene, & Nowak, 2012; Rand et al., 2014).

Experiment 2: Sequential and De-Coupled Prisoners' Dilemmas

Recall that in our introductory example, Elliott gave his project to Michelle knowing that she had a project she could give him. She could thus attribute his motives in part to tactical self-interest, which could weaken her inclination to reciprocate. In contrast, suppose that as Elliott decides whether to give his project to Michelle, he does not realize that Michelle has a project she could give him. If he then nevertheless gives his project to her, Michelle can confidently attribute Elliott's behavior to social motivation. After all, he would have no reason to act tactically. This should heighten her inclination to reciprocate. Experiment 2 considers this prediction.

The sequential prisoners' dilemma in Figure 3 roughly captures a situation in which Elliott knows about Michelle's project. In this game, the first-mover can claim $2 for himself or generate $4 for the second-mover, who can then likewise either claim $2 for herself or generate $4 for the first-mover. If both players "defect" by miserly claiming their own $2, they each end up with $2. If they instead "cooperate" by generously creating $4 for the counterpart, they each end up with $4. If one defects and the other cooperates, they receive $0 and $6, respectively. Defecting corresponds to retaining one's project, while cooperating corresponds to handing it off to the other person, to whom it is more valuable. Finally, each player realizes the first-mover makes his decision knowing the second-mover will have a similar decision.

Our model can succinctly capture the ambiguity of first-move cooperation. Suppose the players anticipate that the first-mover will cooperate and that the second-mover will reciprocate both cooperation and defection. Then first-move cooperation yields each player $4, and first-

move defection yields each player $2. It therefore "helps" the second-mover but also "helps" the first-mover and receives a motivation score of $+1 - \theta_i$.[4]

A "de-coupled" variant of the sequential prisoners' dilemma roughly captures a situation in which Elliott is not aware that Michelle has a project she could give him. It differs from the standard game in only one respect: Both players are initially unaware of the second-mover's decision. At the outset, all they know is that one player faces a decision which affects both of them, and that the other player will be notified of the action he selects. Only later, after the first player has selected an action and the second player has been notified, are they informed that the second player faces her own decision.

First-Move Cooperation is Less Ambiguous under De-Coupling

By our analysis, a first-move cooperator in the de-coupled game "hurts himself" (by 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 rather than $+1 - \theta$, reflecting no discounting of his social motivation. Compared to the standard game, second-movers should thus be more inclined to reciprocate cooperation in the de-coupled game. This parallels the observation that Michelle

---

4 Though the derivation of motivation scores resembles that of Figure 1B, a new point delivered by Figure 3 is that conditioning players' judgments on an entire anticipated strategy profile is a crucial feature of the model and sensibly so. That is, anticipated actions off the path of play necessarily factor into the assessment of motivations. For instance, in the sequential prisoners' dilemma, first-move cooperation is only ambiguous if the second-mover is expected to reciprocate defection as well as cooperation. Both expectations are necessary for cooperation to be materially beneficial for the first-mover, which in turn engenders second-mover discounting of his genuineness.

will be more likely to reciprocate Elliott if he gives her the project without believing that she has a project she could give him.

First-Move Defection

Along with first-move cooperation and subsequent responses, it is instructive to consider first-move defection and subsequent responses.[5] A second-mover who is sufficiently motivated by altruism or impression management could respond to defection with cooperation. On the other hand, both material self-interest and negative reciprocity may induce her to respond to defection with defection.

While our analysis indicates that de-coupling removes ambiguity about first-move cooperation, it indicates that de-coupling does not remove ambiguity about first-move defection. Note that in equilibrium, a first-move defector in the standard game anticipates that the second-mover will defect on cooperation rather than reciprocate it. Given this anticipated strategy profile, the first-mover's defection "hurts the other player" (who will receive $2 rather than $6), but it also "helps himself" (by taking $2 rather than $0). In the de-coupled game, 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). In both settings, first-move defection could thus be interpreted as driven by a desire to be unkind to the second-mover. Or it could be interpreted as

---

5 First-move defection very roughly corresponds to Elliott not giving his project to Michelle. However, it is an observable action, whereas Elliott may simply not act, and never contact Michelle, if he does not intend to give her his project. Moreover, if he does not act, she might take the initiative and contact Elliott to give her project to him. Experiment 3 examines a setting in which parties may not act rather than act and in which either party may take the initiative.

"merely" self-interested. In both settings, it thus accrues a motivation score of $-1 + \theta$ and should therefore induce the same degree of negative reciprocity.

In sum, our model makes a dual prediction: Relative to the standard game, de-coupled game second-movers will be more likely to reciprocate cooperation but not defection.

Experiment 2 and Alternative Game-Theoretic Accounts of Reciprocity

Rabin (1993) matches our model's predictions and indeed invokes an ambiguity-based intuition. Recall that in Rabin's (1993) theory, behavior that might be tactical, like standard game first-move cooperation, is underserving of positive reciprocity. Only behavior that cannot be tactical, like de-coupled first-move cooperation, merits positive reciprocity. Assuming some second-mover cooperation reflects altruism or other social motivations, it follows that second-movers in both games will sometimes respond to cooperation with cooperation, but that de-coupled second-movers will do so more frequently. Rabin's theory views both de-coupled and standard first-move defectors as meriting negative reciprocity and thus implies they will be reciprocated equivalently.

Dufwenberg and Kirchsteiger (2004) also match our model's predictions but cite a different psychological mechanism. By cooperating rather than defecting, a first-mover grants his counterpart \$4 rather than \$0 in the de-coupled game and \$4 rather than \$2 in the anticipated equilibrium of the standard game. Dufwenberg and Kirchsteiger (2004) thus view him as kinder and as more likely to elicit reciprocity. The underlying intuition is that the same action (creating \$4 for a counterpart by cooperating) is kinder if it benefits a relatively poor person than a relatively rich person. This intuition is certainly compelling in many settings. To gather some indication of whether it obtains in our experiment, we will ask second-movers to rate the kindness of first-movers. Finally, de-coupled and standard first-move defectors reduce their

counterpart's outcome by the same amount, so that in Dufwenberg and Kirchsteiger's (2004) view they should be treated equivalently.

Falk and Fischbacher (2006) predict no difference in second-mover reciprocity rates across the standard and de-coupled games. Their theory focuses on issues of equity and social comparison. The juxtaposition of de-coupled and standard sequential prisoners' dilemmas is not germane to these concerns.

Method

Participants. A total of 297 undergraduates (54.9% female, mean age = 20.8 years) at UCSD's Rady School of Management took part in our experiment. We used the same recruitment criteria and incentive structure as in Experiment 1.

Power calculations based on a pilot study suggested that during the period in which the experiment was conducted, approximately two weeks' worth of participants would generate a sufficient sample size. Toward the end of the study, after we had collected their responses to all dependent measures (described below), we asked participants two comprehension questions. Most of them (233, or 78.4%) answered both questions correctly, and only two participants answered both questions incorrectly. Participants in the de-coupled condition performed slightly better than their counterparts in the sequential condition, but the difference was not significant ($p = .22$). We did not exclude any participants from the analyses below; our results are qualitatively unchanged if only participants who correctly answered both questions are included.

Procedure. We conducted sessions of between seven and sixteen participants in the same room and using the same methods as in Experiment 1. Each session was randomly assigned to either the sequential or de-coupled condition. Because our central hypotheses concerned second-

movers, we implemented the same procedure as in Experiment 1 to match each first-mover with up to three second-movers.

Materials. Complete instructions for both conditions, in the form of screenshots, are in the Supplemental Materials. We used a framing that did not employ the terms "cooperate" and "defect." In what follows, we detail several other crucial aspects of the instructions.

At the outset of the standard game, we credited each participant 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 would add $2 to the amount conveyed to the second-mover, so that the second-mover would receive a total of $4. We further told participants that once the first-mover selected an action, the second-mover would be notified of that action and then make her decision between "keeping" and "giving."

At the outset of the de-coupled game, we credited only the first-mover with $2. The first-mover then made his choice between "keeping" and "giving;" we told participants that once the first-mover selected an action, the second-mover would be notified of that action. We did not inform the participants that the second-mover would eventually face the same choice. That choice was later introduced by informing participants that the second-mover was now also being credited with $2 and had his or her own decision to make.

We also collected ratings data concerning perceptions of motives. In both conditions, after they had been notified of the first-mover's action, second-movers first made their own choice and subsequently assessed their counterpart's motives. For the latter, second-movers were asked "to what extent do you think Participant 1 was acting out of kindness by giving [keeping] rather than keeping [giving]?" and "to what extent do you think Participant 1 was acting out of self-interest by giving [keeping] rather than keeping [giving]?" The questions were

displayed simultaneously, with the kindness question always on top. Participants responded on 7-point scales with endpoints labeled "Not at all" and "A lot." The intuition formalized by Dufwenberg and Kirchsteiger's (2004) theory that generous behavior is kinder when it is directed at a counterpart who is less well-off, suggests that first-mover cooperators will be perceived as kinder in the de-coupled game (where absent their cooperation, the second-mover will ostensibly end up with no money). Our model does not predict this difference in kindness; it only predicts that first-move cooperators will be perceived as less selfish in the de-coupled game.

Results and Discussion

The data corroborate our predictions. As a preliminary, consider first-movers. Table 2 shows that 59.3% of standard 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). In contrast, only 37.1% of de-coupled game first-movers cooperated. The difference is marginally significant, $\chi^2(1, 89) = 3.75$, $p = .053$. It could reflect more frequent tactical behavior in the standard game; it could also reflect the pursuit of positive reciprocity in the standard but not the de-coupled game (where at the time he acts, reciprocity appears unattainable to a de-coupled first-mover, because he does not expect his counterpart to have a decision of her own).

Next, consider second-movers who faced first-move cooperation. In the standard game, 62.5% reciprocated with cooperation of their own. This number also accords with previous studies (Clark & Sefton, 2001; Hayashi, Ostrom, Walker, & Yamagishi, 1999; Kiyonari, Tanida, & Yamagishi, 2000; Tversky & Shafir, 1992). By contrast, in the de-coupled game, 84.4% reciprocated. This difference across games is statistically significant $\chi^2(1, 77) = 4.85$, $p = .028$.

The increase in reciprocal cooperation is consistent with the notion that a cooperative first-mover's motives are ambiguous in the standard, but not the de-coupled game. In turn, the second-mover's social motivation to reciprocate is dampened in the standard, but not in the de-coupled game. To estimate the rate of principled defection implied by this analysis, note that de-coupling reduces second-mover defection on cooperation from 37% to 16%. This reduction suggests that approximately half of defections in the sequential game may be principled (because [.37 - .16] / .37 = .57). As before, this estimate is subject to complex sampling error. In addition, it does not reflect the possibility that de-coupling may trigger psychological mechanisms beyond our analysis; it does not account for Dufwenberg and Kirchsteiger's (2004) view, for instance, that de-coupled first-movers are simply more generous than standard first-movers. Respecting these caveats, the sizable effect nevertheless suggests that principled defection may be prevalent. Much non-reciprocation of positive treatment need not stem from selfishness and lack of caring. Instead, it may emerge because social motivation remains dormant given attributions of potentially tactical behavior.

As predicted, de-coupling did not increase negative reciprocity on the part of second-movers who faced first-move defection. In the standard game, 75.0% reciprocated with defection of their own, which is again in line with several 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, a virtually identical 74.8% reciprocated with defection, $\chi^2(1, 74) = 0.0006$, $p = .98$.

Finally, second-movers' ratings lend support our predictions. First-move cooperators were rated less self-interested in the de-coupled ($M = 2.9$, $SD = 1.6$) than in the standard game ($M = 3.7$, $SD = 1.7$), $t(75) = 2.00$, $p = .048$. They were rated slightly but not significantly kinder in the de-coupled than in the standard game (Ms = 5.6 vs. 5.4, SDs = 1.3 vs. 1.6), $t(75) = 0.81$, p

= .42. This null effect is arguably inconsistent with Dufwenberg and Kirchsteiger's (2004) view that de-coupled first-move cooperators are more generous than standard first-move cooperators. Ratings of first-move defectors showed a muted pattern. De-coupled game first-move defectors were rated slightly but not significantly more self-interested (*M*s = 6.3 vs. 5.9, *SD*s = 1.1 vs. 1.7), $t(129) = 1.45$, $p = .15$, and slightly but not significantly less kind (*M*s = 1.8 vs. 2.2, *SD*s = 1.1 vs. 1.7), $t(129) = .81$, $p = .42$.

De-Coupled Gift Exchange

Stanca, Bruni, and Corazzini (2009) studied de-coupling in the gift-exchange paradigm, a well-known generalization of the sequential prisoners' dilemma in which cooperation is a matter of degree rather than all-or-none (Fehr, Kirchsteiger, & Riedl, 1993, 1998). They report data that are somewhat difficult to interpret but generally in line with Experiment 2. Moreover, considering their paradigm and results reveals a key property of our model that we have not yet discussed and that further highlights the role of attributions and discounting in reciprocity.

Stanca et al. (2009) staked each of two players twenty tokens that were redeemable for cash. The first-mover could send a "gift" of any portion of his stake to the second-mover. The tokens sent would be tripled. For example, if the first-mover sent ten tokens, the second-mover received thirty tokens. Before learning what the first-mover elected to do, the second-mover listed her potential responses: she indicated what gift 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.). Any tokens she sent would also be tripled.

Consistent with tactical self-interest, Stanca et al. (2009) reported greater first-mover gift-giving in the standard game. In apparent contrast to our Experiment 2, however, Stanca et al.

(2009) reported that, collapsing across all possible first-mover gifts, the average second-mover return gift was only modestly greater under de-coupling (see their Tables 2 and 7). Moreover, the statistical significance of this finding hinged on the specific method of comparison.

The limited impact of de-coupling is intriguing. It begs the question of whether Stanca et al.'s (2009) data speak against the importance of attributions and discounting for reciprocity. In the next subsection, we show that they do not. On the contrary, a key property of our model offers a natural explanation for the juxtaposition of Stanca et al.'s (2009) modest results and our more pronounced results.

Discounting Can Influence the Perceived Valence of an Action

In many settings, our model implies that an action's perceived valence is fixed. In the handoff games of Experiment 1, for instance, a first-mover handoff is always perceived as at least somewhat kind, while the first-mover acting unilaterally is always perceived as at least somewhat unkind. Likewise, in the de-coupled and sequential prisoners' dilemmas of Experiment 2, cooperation is always viewed positively and defection negatively. However, in other settings—notably including gift-exchange—our model implies that an act's perceived valence will depend on how much the perceiver discounts for self-interest. A behavior judged positively by one person may be judged negatively by someone else.

Consider a de-coupled first-mover in Stanca et al.'s (2009) experiment who gives five of his twenty tokens to the second-mover. Is he acting kindly or unkindly? 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. Averaging over the fifteen more generous and five less generous gifts yields a total motivation score of $[15(-1 + \theta) +$

5(+1)] / 20 = –½ + ¾θ. Whether the first-mover has been kind or unkind is therefore open to interpretation. To somebody who discounts very little, the overall motivation score is negative: She takes gifts at face value and is disappointed in merely small measures of generosity, like the sending of five tokens. To somebody who discounts a lot, the overall motivation score is positive: She is keenly aware of greed and self-interest as motivational forces and is impressed by even small measures of generosity.

The same first-mover gift may then elicit disparate return gifts, depending on whether a second-mover discounts a lot and perceives the first-mover's action as kind or discounts a little and perceives it as unkind. As a result, the relationship between initial and return gifts will be noisy in both the standard and de-coupled gift-exchange games, making it difficult to detect an effect of de-coupling. That is, rather than revealing a limited impact of attributions and discounting, Stanca et al.'s (2009) data may reflect a robust impact of attributions and heterogeneous discounting, and a resultant attenuated relationship between initial and return gifts in gift-exchange.

The difficulties in measuring reciprocity in the gift-exchange paradigm do not arise with the simpler structure of the sequential prisoners' dilemma. Cooperation is always kind in the simpler structure; it cannot be unkind. Likewise, defection is always unkind; it cannot be kind. In other words, sequential prisoners' dilemmas are a more natural setting than gift-exchange for confirming patterns like principled defection.

Do People Have an Inherent Interest in Reciprocity?

A vast literature that spans the social and biological sciences investigates pro-social behavior via experimental and simulated games (for reviews, see Fehr & Gächter, 2000; Malmendier, te Velde, & Weber, 2014). In line with the juxtaposition of our results and Stanca

et al.'s (2009) results, this literature includes a stark dichotomy of views about reciprocity. One influential stream asserts that people have an inherent taste or preference for reciprocity. Fehr, Fischbacher, and Gächter (2002, p.3; see also Bowles & Gintis, 2002; Gintis 2000; Gray, Ward, & Norton, 2014), for instance, offer what they term the strong reciprocity hypothesis:

> "[W]e provide strong evidence in favor of … strong reciprocity … A person is a strong reciprocator if she is willing to sacrifice resources (a) to be kind to those who are being kind … and (b) to punish those who are being unkind ... even if this … provides neither present nor future material rewards."

In contrast, another influential stream suggests that people have little interest in reciprocity per se, especially positive reciprocity (Bolton & Ockenfels, 2000; Delton, Krasnow, Cosmides, & Tooby, 2011; Nowak, 2006; Offerman, 2002; Rand & Nowak, 2013). For example, Charness and Rabin (2002, p. 850) write that:

> "[m]ost of our evidence strongly replicates others' findings that positive reciprocity has virtually no explanatory power in many of the conventional games studied."

Our model accords with the former view. Indeed, our model may be seen as a formalization of strong reciprocity. By emphasizing attributions and discounting, we attempt to psychologically enrich this hypothesis. Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006) likewise offer specific formalizations of strong reciprocity.

The latter view, which downplays the importance of reciprocity, has been catalyzed in part by an impactful series of papers noting that many one-shot interactions confound reciprocity with other concerns. Consider distributional issues like inequity aversion. A second-mover in a sequential prisoners' dilemma who responds to first-move cooperation with her own cooperation may not be driven by a taste for helping those who help her. She may simply be interested in equality (Bolton, 1991; Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999; Ochs & Roth, 1989). Or consider efficiency. A second-mover who cooperates on cooperation may do so because she

does not want to be "wasteful" and thus strives to maximize the parties' joint return (Engelmann & Strobel, 2004).

Note that the data we present are not susceptible to counter-explanations that straightforwardly invoke distributional or efficiency concerns. For instance, explaining the key finding of Experiment 2 on the basis of inequity or waste aversion requires the assumption that de-coupled second-movers are more sensitive to such issues than standard second-movers.

More broadly, we believe that understanding the impact of attributions and discounting can inform the debate between those who accept the notion of an inherent interest in reciprocity and those who question it. In particular, our model and the results we have presented so far suggest three overlapping ways in which existing research may underestimate people's inherent interest in reciprocity.

First, extant theories fail to appreciate some instances of positive reciprocity. Recall Figure 2A, which captures a situation in which someone can give a valuable project to a counterpart. If the counterpart dedicates herself to the project, both individuals will profit to some degree, but if not, neither individual will accrue any profit. This situation, we have argued, can be fertile ground for reciprocity. Although it is in her material self-interest, the person receiving the project may react with dedication in part out of a desire to reciprocate, and the initial individual may anticipate his counterpart being socially motivated in this way. In general, we believe reciprocally kind interactions frequently involve a responding individual who is ambiguously motivated. But as we have also discussed, extant theories do not recognize such interactions as instances of positive reciprocity.

Blindness to such instances of positive reciprocity has shaped the interpretation of prominent empirical work. The ultimatum game, for instance, has been an intense focus of investigation for nearly forty years. In its canonical instantiation, two players, a proposer and

responder, divide a pot of money. The proposer offers a specific split. The responder can accept or reject this offer. If she accepts it, it is implemented. If she rejects it, neither player receives anything. Experiments reveal two key stylized facts: proposers are frequently generous, offering an equal or nearly equal split, and responders frequently reject miserly offers of less than about a quarter of the pot (for reviews and meta-analyses see Camerer & Thaler, 1995; Güth & Tietz, 1990; Oosterbeek, Sloof, & Van de Kuilen, 2004; Tisserand, 2014).

Both our model and alternative game-theoretic accounts view rejections of miserly offers as instances of negative reciprocity. But the alternative theories do not view any offers, no matter how generous, as instigations of potential positive reciprocity. Only our model does. Positive reciprocity is achievable in our model but not extant theories, because when a responder accepts an offer, her motives are ambiguous. She benefits the proposer, by allowing him to garner a positive payoff rather than zero, but she similarly benefits herself. Reflecting extant theories, many authors characterize responders' willingness to forfeit their own money to reject miserly proposers as particularly clear evidence of a taste for meeting unkindness with unkindness (e.g., Fehr, Fischbacher, & Gächter, 2002; Fehr & Gintis, 2007; Gintis, 2000; Rabin, 1993). The same authors do not mention that some generous offers may reflect proposers willingly sacrificing money in pursuit of the good feelings generated by the responder accepting. Rather than positive reciprocity, several other motivations are typically cited for generous offers, including altruism (Camerer & Thaler, 1995), fairness (Nowak, Page, & Sigmund, 2000), image maintenance (Dana, Cain, & Dawes, 2006), and, of course, self-interest (proposers may resort to generous offers because they fear that miserly offers would be rejected; see, e.g., Hoffman, McCabe, Shachat, & Smith, 1994). In our concluding discussion, we further consider the intuitive possibility that proposers' pursuit of positive reciprocity engenders generous offers in the ultimatum game.

Second, extant theories fail to appreciate the degree to which people who do not reciprocate a counterpart's behavior may nevertheless be interested in reciprocity. Principled defection offers a prime example. Extant theories classify second-movers in a prisoners' dilemma who do not reciprocate a cooperative first move as relatively uninterested in positive reciprocity. As we have discussed at length, our model allows for the possibility that these individuals do have a pronounced interest in positive reciprocity—but that attributions of a first-mover's potentially tactical behavior mute their expression of this interest.

The logic of principled defection also speaks to field experiments of gift-exchange. In these experiments, employers grant employees unexpected pay raises, to which employees can respond with greater work effort—or not. Some researchers report that employees do boost their effort, but others report no boost in effort (Cohn, Fehr, & Götte, 2015; Cohn, Fehr, Herrmann, & Schneider, 2014; Esteves-Sorenson, 2017; Falk, 2007; Gilchrist, Luca, & Malhotra, 2016; Gneezy & List, 2006; Kube, Maréchal, & Puppe, 2012). The heterogeneity of these results presents a puzzle that has drawn considerable attention. As Cooper and Kagel (2016) write in the *Handbook of Experimental Economics*, "[i]t remains an important question to determine why gift exchange manifests itself in some settings and not in others (p. 274)."

Our analysis suggests that principled defection provides an answer. Employees may respond differently to pay raises they perceive as genuine tokens of appreciation rather than an employer tactic for eliciting increasing productivity. That is, they may act on their interest in positive reciprocity only if they are sufficiently convinced their employer has acted out of social motivation. This intuitive hypothesis is keenly appreciated in studies of human resources management (Dabos & Roussea, 2004; Eisenberger, Armeli, Rexwinkel, Lynch, & Rhoades, 2001; Porter, Pearce, Tripoli, & Lewis, 1998). Indeed, in an influential textbook, Baron and Kreps (1999) write that the impact of "gifts to employees is likely to depend on ... the gifts being

seen as ... sincere and benevolent rather than selfishly motivated. (p. 109)." The empirical literature, however, has by our reading only distinguished between gifts that employers send intentionally and gifts that employees receive for other reasons. It has not accounted for potentially varied attributions regarding intentionally sent gifts.

Third, as our earlier discussion of Stanca et al. (2009) indicates, reciprocity will often have a subtle and therefore easily dismissed empirical signature (see also Cox, 2004). Gift-exchange is just one example of a prominent game that allows for varying degrees of cooperative versus uncooperative behavior. Other examples include trust games (Berg, Dickhaut, & McCabe, 1995), lost wallet games (Dufwenberg & Gneezy, 2000; Woods & Servátka, 2016), and moonlight games (Abbink, Irlenbusch, & Renner, 2000). In such games, the correlation between how much a first-mover sends and how a second-mover responds tends to be modest. A recent meta-analysis of trust games, for instance, pegs the correlation in the range of .2 to .3 (Johnson & Mislin, 2011). In individual studies, the correlation frequently falls shorts of statistical significance (e.g., Berg, Dickhaut, & McCabe, 1995; Pillutla, Malhotra, & Murnighan, 2002). These modest effects and their lack of robustness are often interpreted as strong evidence against a genuine interest in positive reciprocity. Our model suggests they may instead reflect the workings of attributions and discounting and be entirely consistent with interest in reciprocity.

Indeed, because our model allows for the perceived valence of an action to vary when cooperation is a matter of degree, interactions involving first-mover transfers of less than 50% of an endowment will be difficult to classify even qualitatively. The same transfer and ensuing response may be reciprocally kind, reciprocally unkind, or as a mix of kind and unkind, depending on the discounting and attributions of the individuals involved.

On the other hand, our analysis also points to a testable asymmetry. While transfers of less than 50% may frequently be perceived as either kind or unkind, transfers of 50% or greater

will always perceived as kind, no matter how extensive the discounting. It follows that second-movers may be less disparate in responding to larger transfers than small transfers. This conjecture accords with several studies showing that first-movers who send large fractions of their holdings often receive more consistent (and more charitable) treatment than those who send small fractions (e.g., Glaeser, Laibson, Scheinkman, & Soutter, 2000; Pillutla, Malhotra, & Murnighan, 2002; Schotter & Sopher, 2006).

In sum, the impact of attributions and discounting renders documentation of reciprocity a complex enterprise (see also Cooper & Kagel, 2016, p. 241). Researchers may fail to recognize some instances of reciprocity, fail to appreciate the extent to which people who do not reciprocate their counterparts' behavior are nevertheless interested in reciprocity, and dismiss subtle empirical patterns that belie pronounced underlying interest in reciprocity. A person who is a strong reciprocator, in the sense of having an inherent preference for reciprocity, but who also engages in attributions and discounting, will at times behaviorally resemble a person with scant or merely tactical interest in reciprocity. As a result, there may be a tendency to underestimate people's inherent preference for reciprocity.

Experiment 3: Who Makes the First Move?

In our introductory example, Elliott acted first. He gave his project to Michelle, thereby putting the ball in her court. Though we have not emphasized it, as we constructed the example, Elliott's making the first move was not a given. One could imagine Elliott hesitating, and Michelle in quick time making the first move herself by giving her project to him. Or both individuals might be reluctant to make the first move, so that a longer period of inaction passes, and the question emerges of whether anyone will eventually seize the opportunity to act.

In other words, the sequential prisoner's dilemma depicted in Figure 3 may capture many elements of Elliott and Michelle's interaction, but it does not capture decisions regarding whether to make the first move. Such decisions are intimately intertwined with considerations of reciprocity. Indeed, in his famous sociological treatment of reciprocity, Gouldner (1960) notes that the question of "who goes first" is endemic to exchange. Much like us, Gouldner (1960, p. 177) considers two people, each of whom holds an item prized by the other. He observes that when the two consider a transaction:

> "Each may then feel that it would be advantageous to lay hold of the other's valuables without relinquishing his own. Furthermore, suppose that each party suspects the other of precisely such an intention … each is likely to regard the impending exchange as dangerous and to view the other with some suspicion. Each may then hesitate to part with his valuables before the other has first turned his over … each may say to other, 'You first!' Thus the exchange may be delayed or altogether flounder and the relationship may be prevented from developing.
>
> "… reciprocity may serve as a starting mechanism in such circumstances by preventing or enabling the parties to break out of this impasse."

Accordingly, we now contrast the standard, sequential prisoners' dilemma with what we term an "endogenous sequencing" prisoners' dilemma. In this new game, each player can again "defect" by claiming $2 for himself or "cooperate" by generating $4 for the other player. Ex ante, however, nothing distinguishes the players: There is no assignment of players to the roles of first- 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. That is, 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 are placed in a simultaneous prisoners' dilemma.

Less Ambiguous First-Move Cooperation under Endogenous Sequencing

Our model suggests that the kindness of a first-move cooperator will seem less ambiguous under endogenous sequencing than in the standard game, and that this effect will be magnified the more rapidly he cooperates. To introduce the underlying intuition, consider a first-move cooperator who anticipates that his counterpart will reciprocate both cooperation and defection. This individual has more alternative strategies that alter the player's payoffs under endogenous sequencing than in the standard game. In both cases, he could first-move defect. But under endogenous sequencing, he could also pursue the maximal possible payoff, by waiting rather than making the first-move, hoping his counterpart first-move cooperates, and then defecting on her.

Because of his extra alternative, this first-mover cooperator *can* credibly signal his social motivation. In eschewing pursuit of the maximal payoff, he "hurts himself" and "helps" his counterpart. Bearing a cost to benefit his counterpart lends credence to the possibility that he is socially motivated rather than self-interested.

In turn, the kindness of an endogenous sequencing first-move cooperator will be less ambiguous than the kindness of a standard game first-move cooperator. As we have discussed, the latter individual cannot signal his social motivation. He has no way to bear a cost to benefit his counterpart. His only alternative is first-move defection, and relative to that alternative first-move cooperating helps his counterpart but also helps himself.

Finally, consider the speed with which a person first-move cooperates under endogenous sequencing. As he acts more rapidly, he more emphatically eschews pursuit of the maximal possible payoff. Immediately cooperating sends the clearest signal: There is no chance the player was trying to wait out his counterpart and planning to defect if she made a cooperative

initial move. Waiting dilutes the signal. It generates ambiguity about the player's motives—maybe he would have defected if his counterpart had already cooperated. Thus, the faster someone first-move cooperates, the less ambiguity there is about the kindness underlying his cooperation.

Recent social psychological research accords with the notion that quick actions can be strong signals. Critcher, Pizarro, and Inbar (2013) report that decisions that may be morally good are judged especially positively when made quickly; likewise, decisions that may be morally bad are judged especially negatively when made quickly. Taking time to ponder a decision evidently 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.

Equilibria, Motivation Scores, and Empirical Predictions

In the Supplemental Materials, we analyze a set of strategy profiles that epitomize reciprocity and formally capture the aforementioned intuition. In particular, we suppose that both players anticipate that the other would first-move cooperate at some point in time, and would reciprocate both first-move cooperation and first-move defection. These strategy profiles extend the standard, sequential game profile we discussed earlier, in which the first-mover cooperates and the second-mover reciprocates both cooperation and defection. They involve players who are maximally willing to be a part of mutual cooperation as well as mutual defection.

We show that in the resulting equilibria, the first-move cooperator accrues a motivation score between $+1 - \theta$ and +1 instead of the $+1 - \theta$ accrued by a first-move cooperator in the standard, sequential game. Moreover, the earlier he must act to make the first move, the closer

his motivation score draws to +1. A similar analysis holds for the second-mover, whose overall motivation score also approaches +1 as the first move comes earlier.

While our model also permits other equilibria, the foregoing suggests two empirical predictions. First, many games will feature rapid, first move cooperation. Players will be willing to make themselves vulnerable to a counterpart's defection, because they realize that doing so best allows them to pursue positive reciprocity. Second, and most critically, rapid first-move cooperation will be reciprocated at a higher rate than first-move cooperation in the standard game. This prediction also reflects the strong signaling value of quick first-move cooperation. If second-movers doubt the social motivation of a rapid first-move cooperator less than that of a standard game first-move cooperator, they should be less inclined toward principled defection.

Experiment 3 and Alternative Game-Theoretic Accounts of Reciprocity

Within the class of equilibria we have outlined, both Rabin (1993) and Dufwenberg and Kirchsteiger (2004) judge first-move cooperation as kinder under endogenous sequencing than in the standard game. In the standard game, the first-mover's cooperation grants the second-mover $4 when she would otherwise garner $2; under endogenous sequencing, it grants her $4 rather than the $0 she would accrue were the first-mover to wait, see her cooperate, and then defect on her. Both theories thus predict that first-move cooperation will be reciprocated at a higher rate under endogenous sequencing. On the other hand, within the class of equilibria we have outlined, both theories view all cooperative first moves as equally kind—regardless of their timing. We note, however, that our model as well as Rabin's (1993) and Dufwenberg and Kirchsteiger's (2004) can also reach the predictions we have highlighted by drawing on less parsimonious mixed strategy equilibria that do not reflect the intuition that rapid first-move

cooperation is a strong signal of social motivation. Falk and Fischbacher (2006) do not make such predictions. Like the distinction between the de-coupled and standard game, the distinction between endogenous sequencing and the standard game is not germane to their theory.

Method

Participants. A total of N = 235 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 129 of them (54.9%) were female. As in our other experiments, they received course credit for participating and were paid according to their game outcome. Our prior experience at the Rady lab suggested that during the period when the experiment took place, two weeks' worth of sessions would yield sufficiently many participants. We thus conducted sessions during two Monday through Friday periods.

Participants answered five training questions during the course of the instructions (and, unlike in our other experiments, they did so before responding to the dependent measures). They correctly answered an average of 4.0 of the five questions. On average, they answered slightly fewer questions correctly in the endogenous sequencing condition, but the difference was not significant ($p = .37$). 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.

Procedure. We conducted sessions of between six and twenty participants in the same room and using the same methods as in Experiments 1, 1A, and 2. Each session was randomly assigned to either the sequential or endogenous sequencing condition. 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 we do not include their data below. For

robustness, we ran endogenous sequencing sessions with timer lengths of 20, 60, and 120 seconds.

Materials. As before, the experiment was programmed in z-Tree (Fischbacher, 2007). We maintained the framing of Experiment 2, which did not use the terms defection and cooperation. Each player was initially credited 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. The complete instructions for both conditions, in the form of screenshots, may be found in the Supplemental Materials.

Results and Discussion

Behavior in the standard game was roughly in line with Experiment 2. As Table 3 shows, 66.7% of first-movers cooperated, 67.7% of second-movers responding to cooperation reciprocated with cooperation, and 93.8% of second-movers responding to defection reciprocated with defection.

More importantly, behavior under endogenous sequencing was consistent with our analysis. 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 4 provides initial evidence that games typically ended quickly because one of the players elected to first-move cooperate: 34 of the 39 first-moves that occurred within the initial 20% of the countdown were cooperative (87.2%; 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 of the 25 first-moves occurring within the initial 10% of the countdown were cooperative (88.0%).

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 standard game. Figure 5 illustrates that the data support this prediction: 30 of the 33 second-movers responding to cooperative first moves within the initial 20% of the countdown reciprocated cooperation (90.9%; right panel, leftmost bar). This reciprocity rate significantly exceeds that of the standard game (30/33 vs. 21/31, $p = .029$ 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 of the 22 responses to cooperative first moves occurring within the initial 10% of the countdown were themselves cooperative (95.5%). These results are consistent with our claim that potent signals of genuine kindness can catalyze 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 very quickly. Thus, they are directly comparable to second-movers in the standard 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, an OLS regression shows that the likelihood of second-movers responding to cooperation with cooperation decreased substantially for each percentage point of the countdown elapsed prior to the cooperative first move ($b_{Elapsed}$ = -.89, SE =.21, t = -4.16, $p < .001$). To put this estimate into context, the regression model estimates reciprocity rates of 97% for instantaneous first-move cooperation and just 8% for first-move cooperation just prior to the countdown expiring.

As in our earlier experiments, it is useful to estimate the prevalence of principled defection in the standard game. To do so, we can assume that in the standard game, defection on cooperation can reflect both self-interest and reactions to ambiguity. By contrast, a rapid cooperative first move under endogenous sequencing all but eliminates that ambiguity. Then with the conservative 20 second-cutoff for defining rapid first moves, almost three quarters of failures to reciprocate cooperation in the standard game may be principled (since 1 - [[1 - .91]/[1 - .68]] = .72). As always, this estimate must be taken with a grain of salt; there is sampling error to consider, and the estimate of course conditions on our model's assumptions. Respecting these caveats, the estimate is nevertheless suggestive of the prevalence of principled defection. Recall that in Experiments 1 and 2, roughly one third and one half of failures to reciprocate cooperation may reflect principled defection. These numbers, while disparate, are all consistent with the notion that attributions of potential tactical behavior dampen social motivation which in turn gives rise to principled defection. We thus cite untapped social motivation as a fundamental impediment to cooperation and reciprocity.

Sophisticated Social Judgment versus Social Heuristics

We mentioned earlier that our analysis is at odds with the influential social heuristics hypothesis (Rand, 2016; Rand et al., 2014). This hypothesis belongs to the class of theories

which assume that people are self-interested and not socially motivated. It explains non-tactical cooperation as an overgeneralized, learned response. Because tactical cooperation is frequently profitable, for instance, some people may develop a rapid, automatic tendency to cooperate more widely. Such a heuristic can be overridden by slower, reflective mental processes, but it can govern behavior when the ample attention and cognitive capacity necessary for reflective processes are absent (Rand, Green, & Nowak, 2012). By the social heuristics hypothesis, then, instances of non-tactical cooperation and positive reciprocity are essentially mistakes incurred by self-interested people when they are not sufficiently vigilant. These instances do not reflect a bona fide interest in treating others well or in repaying kindness with kindness.

Social heuristics are shaped by experience. They enact whichever behavior has typically been payoff maximizing in past situations. They may thus favor cooperation, as we have just noted. But depending on a person's history, they could take on many forms (Rand, 2016; Rand & Nowak, 2013). In any case, heuristic processes do not tailor behavior to the specific situation at hand. That task falls to reflective processes, which attempt to maximize self-interest by drawing on whatever strategy best fits a particular situation. As Rand (2016, p. 1192) puts it, "intuition favors behavior that typically maximizes payoffs, whereas deliberation favors behavior that maximizes one's payoff in the current situation." In sum, the claim is that people's instincts may come to include cooperation, but these instincts are often overridden by inherently self-interested, reflective thinking. This position has inspired a lively debate about the cognitive underpinnings of cooperation and reciprocity (for recent views congruent with the social heuristics hypothesis see, e.g., Halali, Bereby-Meyer, & Meiran, 2014; Halevy & Chou, 2014;

for recent alternative views and criticisms, see Kessler & Meier, 2014; Krajbich, Bartling, Hare, & Fehr, 2015; Tinghög et al., 2013).[6]

In Experiment 3, heuristic cooperation could contribute to rapid first-move cooperation under endogenous sequencing. Importantly, however, social heuristics cannot accommodate our key results. Second-movers show an increased willingness to reciprocate rapid, endogenous sequencing cooperation in Experiment 3. Similarly, Experiments 1 and 2 reveal increased positive reciprocity given, respectively, a highly attractive unilateral action for the initial cooperator and de-coupling. Experiment 1A demonstrates increased generosity given an active than passive recipient of this generosity. All these results can be explained by an analysis of attribution, discounting, and ambiguous motives. None of them seems amenable to a hypothesis implicating differential levels of reflective thinking. For instance, why would second-movers' reliance on reflection vary with a $3 versus $6 unilateral payout for the first-mover; or why would de-coupled second movers be less reflective than standard second-movers?

On the flip side, a synthesis of our analysis and dual process theories of social inference (Gilbert, Krull, & Pelham, 1988; Gilbert, Pelham, & Krull, 1988; Trope 1986; see also Lieberman et al., 2002) can naturally explain data cited as corroboration of the social heuristics hypothesis. The leading empirical support for social heuristics comes from studies showing that factors like time pressure and cognitive load, which impede reflective processes and thereby engender reliance on heuristic processes, increase cooperation rates in non-tactical settings (Rand, 2016). This finding is taken as evidence that effortful, deliberate thinking is more self-

---

[6] Some time ago, Dawes (1980, p. 190) advanced a prediction largely opposite that of social heuristics. He argued that impeding reflective processes will narrow people's focus to material payoffs, which are often obvious and salient, to the exclusion of social considerations, which require more careful thought to understand and appreciate.

interested than heuristic thinking. However, research on the cognitive and neurological underpinnings of social inference has identified several ways in which curtailing reflection impacts many social judgments, including attributions and expectations (Fein, 1996; Lieberman, 2007; Lieberman et al., 2002; Trope & Alfieri, 1997; Trope & Gaunt, 1999, 2000). In the remainder of this section, we argue that findings cited as evidence for self-interested reflection may instead emerge from established properties of impaired social judgment.

We next flesh out our alternative explanation, by outlining how it can be distinguished from the social heuristics hypothesis in sequential games akin to our experiments. Rand's (2016) meta-analysis includes five studies of non-tactical, second-mover behavior in such settings. Subsequently, we consider how our explanation and social heuristics can be distinguished in simultaneous-play games. The vast majority of the studies of non-tactical cooperation meta-analyzed by Rand (2016) examine such settings.

According to dual process theories of social inference, attributions result from rapid, automatic impressions that are adjusted and refined by slower, effortful thinking. Discounting is generally seen as predominantly slow and effortful; that is, as primarily a reflective adjustment (Fein, 1996; but see Ham & Vonk, 2004). Impeding reflective processes thus causes attributions to hew more closely to initial impressions (Gilbert, Krull, & Pelham, 1988; Gilbert, Pelham & Krull, 1988; Lieberman et al., 2002; Trope, 1986). Put in our terms, impeding reflective processes drives θ towards zero.

Our analysis and dual process theories thus jointly point to an ambiguity-driven prediction: Taxing mental resources should render judgments of ambiguously-motivated cooperation more neutral; by our model, it will push motivation scores from $+1 - \theta$ toward +1. By contrast, judgments of unambiguously-motivated cooperation are not discounted even when mental resources are plentiful, so taxing mental resources should not impact them much; by our

model their motivation score would equal +1 in both cases. Impeding reflective processes should therefore increase reciprocal cooperation in response to ambiguous but not unambiguous cooperation.

The social heuristics hypothesis cannot readily accommodate this ambiguity-driven prediction. Because it assumes that reflective processes have a self-interested bias relative to heuristic processes, it implies that impeding these processes should always increase reciprocal cooperation.

Handoff games provide one setting for an empirical test. By our explanation but not social heuristics, impeding reflective processes should increase reciprocal cooperation in Figure 1B, where the unilateral action is worth $3 to the first-mover. It should not increase reciprocal cooperation in Figure 1A, where the unilateral action is worth $6.

Our explanation also yields an ambiguity-driven prediction concerning reciprocal defection that is inconsistent with the social heuristics view. Taxing mental resources will prevent people from discounting their initial impression of first-move defection. They will thus judge this behavior more negatively when mental resources are scarce and be more inclined to respond with their own defection. In other words, impeding reflective processes has the potential to increase defection in response to defection. This prediction can be tested in any social dilemma in which some second-movers cooperate in the face of defection. In the sequential prisoners' dilemma of Experiment 2, for instance, only 75% of second-movers who faced defection responded in kind (the others may have been guided by altruism or some other social motivation). Taxing mental resources should increase this figure.

We now turn to one-shot, simultaneous-play games. In this setting, a player cannot attempt to tactically influence his counterpart, by the simple fact that she will not have knowledge of his actions prior to deciding on her own actions. The principal reason we have

emphasized for discounting is therefore simply not relevant. In addition, any attributions must be contingent ("if she were to do X, I think it'd be because of Y"). Indeed, in this setting the What? (a counterpart will do) may take the place of the Why? (a counterpart does what she does). That is, expectations rather than attributions likely constitute the critical social judgment that can be stymied by impeding reflective processes.

In this light, consider work by Croson (2000), Caruso, Epley, and Bazerman (2006), and Guerra and Zizzo (2004; Figure 2). These authors prompt participants to form expectations about their counterparts' behavior in simultaneous games. Their results are consistent. In simultaneous prisoners' dilemmas, Croson (2000) observes that participants who are explicitly asked to form expectations are less likely to cooperate. In a control condition of her experiment, the cooperation rate is 77%, but participants who form expectations estimate a 45% cooperation rate, and 55% of them cooperate. The discrepancy between the 77% cooperation rate in the control condition and the 45% estimate thereof presumably contributes to the diminished 55% cooperation rate, since participants who do not foresee cooperation from a counterpart are unlikely to cooperate. In a hypothetical prisoners' dilemma, Caruso, Epley, and Bazerman (2006) similarly find that participants instructed to engage in perspective-taking predict that their counterparts are significantly less likely to cooperate and then cooperate less themselves. Finally, in the game depicted in Figure 6, which is akin to a simultaneous handoff game, Guerra and Zizzo (2004) report that when asked to provide their expectations, row players are less likely to choose Top, which corresponds to the first-mover handing off rather than retaining in the sequential game and pursues the mutually beneficial outcome Top, Left.[7]

---

7 Studies of settings in which cooperation may be tactically profitable find less consistent effects of eliciting expectations (Blanco, Engelmann, Koch, & Normann, 2014; Caruso, Epley, & Bazerman, 2006; Croson, 2000, Experiment 1; Gächter & Renner, 2010).

Such findings accord with the notion that people who think thoroughly about their counterpart's intentions tend to believe their counterpart will behave relatively non-cooperatively. This notion dovetails with Epley et al.'s (2006; Experiments 1 and 5) observation that in social dilemmas, perspective-taking renders a counterpart's egoism salient and pronounced (see also Kruger & Gilovich, 1999; Pierce et al., 2013). It is also in line with Miller's norm of self-interest (1999; see also Epley & Dunning, 2000; Heath, 1999, Markle 2011); by the descriptive element of this norm, people are prone to anticipate non-cooperative behavior from others.

Taxing mental resources should curtail such thinking, along with the reactive egoism and aversion to being the sucker (Kerr, 1983) it frequently engenders. In sum, we suggest that impairing participants' reflective processing increases cooperation because it prevents them from anticipating and reacting to their counterparts' non-cooperation, not because it induces them to rely on relatively cooperative heuristics.

To empirically distinguish our explanation from the social heuristics hypothesis, note that our explanation implies that taxing mental resources should not impact behavior when there is no reason to engage in reactive egoism, sucker aversion, or the like. In contrast, if reflective processes have a self-interested bias relative to heuristic processes, taxing mental resources should always increase cooperation in non-tactical settings. These competing predictions can be tested in any setting in which some participants have stronger reasons for reactive egoism and related patterns than others. Simultaneous handoff games provide such a setting.

Consider Guerra and Zizzo's (2004) game in Figure 6. In this game, expectations regarding a counterpart's cooperative or non-cooperative behavior matter only for row players, whose payoffs parallel those of the first-mover in Figure 1B. If a row player anticipates that his counterpart will cooperate by choosing Left, it is in his interest to choose Top; if he expects his

counterpart to act non-cooperatively by choosing Right, it is in his interest to choose Bottom. There is thus room for his reactive egoism. On the other hand, the column player's own choice is only consequential if the row player chooses Top in pursuit of the mutually beneficial outcome Top, Left. There is nothing she should do or can do to protect her self-interest if he chooses Bottom instead. Thus, there is no reason for her to engage in reactive egoism. Indeed, Guerra and Zizzo (2004) report that row players were less likely to play Top after providing expectations of their counterpart, whereas column players were not impacted by the expectation manipulation. Future research could examine time pressure or cognitive load in this setting. Our explanation, but not the social heuristics hypothesis, predicts that the consequences of taxing mental resources should parallel those of the expectation manipulation.

The social heuristics hypothesis begins with the presumption that people are self-interested. It views instances of non-tactical cooperation as mistakes that arise when reflective processes do not adequately check heuristic processes. By casting pro-social behavior as the product of frugal, automatic thinking, it accords with theories that depict people as "cognitive misers" who often minimize effortful social judgment (e.g., Gilbert, 1989; Gilbert & Osborne, 1989; Quattrone, 1982). Our analysis, on the other hand, begins with the presumption that people are not solely self-interested. They are also socially motivated. It further presumes that people are concerned with and actively judge others' motives. It therefore casts many instances of non-tactical cooperation as products of deliberate thinking, which accords with theories that posit a large role for effortful, sophisticated social judgment (Fein, 1996; Gilbert, Krull, & Pelham, 1988; Trope, 1986). We have shown that an explanation which synthesizes our analysis and such theories can explain the evidence cited in support of the social heuristics hypothesis. We have also identified experiments that can distinguish between our explanation and the social heuristics hypothesis. Beyond the specific issue of social heuristics, we hope the study of

experimental games can further our understanding of simple versus sophisticated social inferences, in the domains of cooperation and reciprocity and more generally (see also Apps, Rushworth, & Change, 2016; Gluth & Fontanesi, 2016; Hein, Morishima, Leiberg, Sul, & Fehr, 2016; Yamagishi, Takagishi, Fermin, Kanai, Li, & Matsumoto, 2016).

General Discussion

We have introduced and experimentally examined an attributional theory of reciprocity that distinguishes between good treatment that is unambiguously caring and good treatment that may be driven by tactical self-interest. In a game-theoretic framework, we formalize the notion that for many people, being treated well is not enough. When the good treatment they receive could be tactical rather than genuinely caring, these people's social motivation remains dormant; rather than reciprocating, they may engage in what we term "principled defection." On the other hand, when the good treatment they receive is unambiguously caring, their social motivation is activated and may be substantial; they may reciprocate at high levels. Because our data suggest that principled defection is commonplace, we argue that untapped social motivation is often a fundamental impediment to cooperation and reciprocity. People frequently act non-cooperatively not because they are selfish and uncaring, but because they do not see a place for their caring when others' behavior may not be driven by caring.

Our work provides a new perspective on several empirical phenomena and the scholarly debates they inspire. It suggests that the impact of attributions and discounting renders documentation of reciprocity a complex enterprise. A person who has a pronounced taste for reciprocity, but who also engages in attributions and discounting that may mute his social motivation, will at times behaviorally resemble a person with scant or merely tactical interest in reciprocity. As a result, there may be a tendency to underestimate people's taste for reciprocity.

Furthermore, we have suggested that the leading evidence for an assumed instinct for cooperation and associated social heuristics is equally consistent with an alternative explanation that emphasizes effortful, sophisticated social judgments. In the remainder of this Discussion, we examine the implications of our analysis for repeated interactions and for research on the well-known ultimatum game. We also consider the conceptual and formal limitations of our work as well as a prescription it offers.

Repeated Interactions: Unraveling Does Not Constitute Evidence Against a Taste for Reciprocity

We have focused on one-shot interactions, but the notion that attributions may dampen social motivation is also relevant to finitely-iterated interactions. In such settings, cooperation rates are often initially high but then "unravel" as the final interaction approaches (Andreoni & Miller, 1993; Morehous, 1966; but see Rapaport & Chammah, 1965; Chater, Vlaev, & Grinberg, 2008; Erev & Roth, 2001). This phenomenon may provide an additional example of extant theories' failure to appreciate the degree to which people who do not engage in reciprocal behavior may nevertheless have a taste for reciprocity. Most theories that assume people have a taste for responding to kindness with kindness cannot account for unraveling, because they imply that a history of past cooperation should engender subsequent cooperation. Unraveling is thus sometimes framed as damning evidence against a taste for reciprocity (e.g., Selten & Stoecker, 1986). Indeed, unraveling is often taken as proof of tactical roots for cooperation and reciprocity. People may cooperate tactically early on and in the middle portion of a game, in order to facilitate and maintain a subsequent stretch of materially profitable, reciprocal cooperation. Late in a game, however, with fewer and fewer subsequent periods, the incentive to continue this tactic wanes (Axelrod & Hamilton, 1981; Kreps et al., 1982). Our model, though it presumes a taste for reciprocity, is entirely consistent with unraveling. To someone who

discounts for self-interest, even a long history of cooperation may reflect ulterior motives and thus not merit late-period reciprocity.

The Ultimatum Game: Positive Reciprocity as a Reason for Generous Offers

We argued earlier that, like other social motivations such as altruism or fairness, positive reciprocity may contribute to the making of generous offers in ultimatum games—but that this intuitive possibility has been overlooked, because it does not fit with extant theories. To further examine the matter, consider two types of equilibria in which the proposer offers a relatively equitable split. In "fearful" equilibria, the proposer is generous because he anticipates that the responder would reject miserly splits. In other words, he is compelled to act generously because of his counterpart's power to reject miserly offers. The stick of negative reciprocity is key. Both our model and extant theories allow for fearful equilibria. In "nice" equilibria, the proposer anticipates that the responder would accept most miserly splits. His generosity thus does not reflect fear; it reflects social motivation. He wishes to treat his counterpart well. In other words, the proposer acts generously in pursuit of the good feelings that are generated if the responder accepts his offer. The carrot of positive reciprocity is key. Our model allows for nice equilibria, but extant theories do not.

We suspect that many investigations of proposer generosity betray a conceptualization of the ultimatum game akin to fearful rather than nice equilibria. For example, Hoffman, McCabe, Shachat, and Smith (1994; see also Harrison & McCabe, 1992; Marlowe, 2004, p. 186) considered altruism as a source of generous offers and argued against it. They observed that proposers were substantially more generous than "dictators" who also split a pot with another individual but could not have their offer rejected. Because altruism should appeal to both proposers and dictators, Hoffman et al. (1994, pp. 347–348) inferred that "offers in ultimatum

games … appear to be determined primarily by strategic … considerations… rather than … preference." In essence, they concluded that proposer's generosity reflects the self-interested avoidance of negative reciprocity inherent to fearful equilibria.

Nice equilibria, however, that implicate positive reciprocity, fit equally well with Hoffman et al.'s (1994) data. As we emphasized in Experiment 1A, because reciprocity is contingent on a counterpart's behavior, it should impact proposers but not dictators. That is, the good feelings engendered by the reciprocal acceptance of a generous offer are one reason proposers might make such offers. This reason is not relevant to dictators, who cannot tap such feelings.

To be clear, we do not claim that fearful dynamics and negative reciprocity are rare in the ultimatum game. They are surely common. Our claim is simply that nice dynamics driven by positive reciprocity may be common too. The intuitive possibility that proposers' pursuit of positive reciprocity engenders generous offers may have been overlooked because it does not fit with extant theories.

Modeling Interpretations and Limitations

We have adopted an interpretation of our model that allows for cross-person variation in both the strength of social motivation and the propensity to discount. 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 in a sequential prisoners' dilemma who have 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 standard game which renders the first-mover's motives unclear, only the second-mover who discounts less—the individual with a lower θ—may reciprocate.

Our data are also consistent with an alternative interpretation that allows for cross-person variation in the strength of social motivation but not the propensity to discount. Under this interpretation, every individual has the same non-zero θ. As a result, differences in reciprocal behavior can be explained by differences in λ only. For instance, a second-mover with intermediate λ and a second-mover with high λ may both reciprocate in a de-coupled game. But in a sequential game, only the more other-regarding second-mover—the individual with a high λ—may reciprocate.

Allowing cross-person variation in both λ and θ best comports with the notion of principled defection. Under this interpretation, a second-mover with abundant social motivation could be less likely to reciprocate cooperation than a second-mover with lesser social motivation, if the former individual also discounts a lot. Allowing for cross-person variation only in λ does not as fully comport with principled defection, because it implies that more socially motivated second-movers are always more likely to reciprocate than less socially motivated second-movers. Nevertheless, both interpretations are fundamentally consistent with principled defection. Both allow for non-zero θ, so that failures to reciprocate cooperation can reflect lack of clarity about motives and untapped social motivation rather than mere self-interest.

By interpreting our model in terms of cross-person variation in θ, we do not mean to imply that discounting is a stable personality trait (Ackermann, Fleiß, & Murphy 2014). Some situations likely accentuate or temper the propensity to discount. Framing a social dilemma as the "Wall Street Game", for example, may induce less discounting for self-interest than framing it as the "Community Game" (Liberman, Samuels, & Ross, 2004). Person-by-situation interactions could also be important. Relatedly, future work could examine the implications of viewing the extent of discounting as a summary of an individual's beliefs about whether others

are socially motivated. In a game of incomplete information, higher values of θ could correspond to greater confidence about a counterpart having low λ.

We have summarized an individual's propensity to discount by the single parameter θ. But this propensity may have at least two different drivers: perceptions of ambiguity in others' motives, and reactions to perceived ambiguity in others' motives (cf. Kruglanski & Webster, 1996). To explicate, note that individuals can vary in their tolerance of perceived self-interest. Some people may reciprocate even if they perceive substantial self-interest by their counterpart, while other people might not reciprocate even if they perceive only minimal self-interest. In other words, two individuals who agree in their perceptions of a counterpart's motives may nevertheless arrive at different decisions about whether to reciprocate her actions. Our findings are consistent with both perception-driven and reaction-driven discounting.

Two mathematical features of our model reflect simplifications worthy of mention. First, motivation scores are ordinal. A player who helps both his counterpart and himself, for instance, receives the same score regardless of how much he or his counterpart stand to gain. Second, and perhaps more importantly, when players are either mutually kind or mutually unkind, their total utilities increase in each other's material payoffs. Ceteris paribus, each player is thus better off by granting the other a greater material gain. This feature is compelling in the case of positive reciprocity. It may not be as compelling in the case of negative reciprocity. At the cost of some added complexity, the realism of both features could be addressed by refining the motivation scores and (or) how they combine into weights on a counterpart's material payoff. An appropriately refined model would preserve the spirit of our approach. It would also preserve the representation of players' preferences as linear combinations of their material payoffs and thus remain within Segal and Sobel's (2007) framework that we draw upon.

Finally, our model reduces extensive form games to their normal form. It thus cannot address issues concerning how players might revise assessments of motivations off the equilibrium path. This is a shortcoming of our model. Dufwenberg and Kirchsteiger's (2004) theory, in contrast, is fully history-dependent and explores the relevant issues. Future work could aim to combine the discounting formalized in our model with their theory's history-dependence.

The Nature of Self-Interest

Our model reduces self-interest to material self-interest. Doing so provides analytic tractability but is a conceptual oversimplification. Self-interest need not be material. People can be self-interested in connection with social issues. For instance, they sometimes bear costs to maintain a positive self-image (Dana, Cain, & Dawes, 2006). Relatedly, research on group engagement (see Tyler, 2013; Tyler & Blader, 2001, 2003), social identity (Abrams & Hogg, 1988; Ellemers, Spears, & Doosje, 2002; Tajfel & Turner, 1979; 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. While we have addressed discounting for material self-interest, we believe people also discount for non-material self-interest. It may be possible to examine whether the extent to which people discount varies systematically across the two.

Moving beyond a simplified, material notion of self-interest suggests replacing the dichotomy of self-interest versus social motivation with the dichotomy of self-interest versus other-regard. There is intriguing work which holds that the tradeoff between self and others is often less sharp than suggested by analyses like ours (see, e.g., Rand et al., 2014). Kameda, Tsukasaki, Hastie, and 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 profitable in the short-run, it must be profitable in the long-run. This belief reflects the view that morality and self-interest cannot truly conflict, because moral behavior is rewarded karmically.

The aforementioned work notwithstanding, some researchers ask whether humans can be truly other-regarding (Batson, 2011; Sahlins, 2008, 2013). In this vein, note that we model people as maximizers of individual utility functions. These utility functions depend on others' payoffs, but their maximization could be interpreted in terms of a broader notion of self-interest that includes both non-material and material concerns. While this and related properties of utility functions have led some researchers to caution against their use (e.g., Caporeal, Dawes, Orbell, & van de Kragt, 1989; Colman, 2003), we do not interpret our approach as endorsing a view of people as ultimately selfish. Like many researchers who investigate altruism (e.g., Krebs, 1970), we remain agnostic regarding this fundamental question.

Conclusion

We have argued that a person may fail to act pro-socially not because she is selfish and uncaring, but because she believes even those who treat her well may be fundamentally self-interested, and she does not see a place for her caring when others are self-interested. In this light, consider Elliott and Michelle one last time. As he ponders whether to give Michelle his project, Elliott may fear Michelle's self-interest. Even if he has altruistic reasons for handing his project off to her, the possibility that she will be selfish and not reciprocate may be sufficiently aversive to dissuade him from doing so. In some situations, Elliott's fear could be well-placed. But our work suggests that rather than worrying about Michelle's self-interest, it would

frequently behoove Elliott to instead worry about her perceptions of his self-interest. If he can communicate that he genuinely cares about how he treats her, Michelle's caring may be activated, and it may be substantial.

References

Abbink, K., Irlenbusch, B., & Renner, E., (2000). The moonlighting game—An experimental study on reciprocity and retribution. *Journal of Economic Behavior and Organization*, 42, 265–277.

Abrams, D., & Hogg, M. A. (1988). Comments on the motivational status of self-esteem in social identity and intergroup discrimination. *European Journal of Social Psychology*, 18(4), 317-334.

Ackermann, K. A., Fleiß, J., & Murphy, R. O. (2016). Reciprocity as an individual difference. *Journal of Conflict Resolution*, 60(2), 340-367.

Andreoni, J., & Miller, J. H. (1993). Rational cooperation in the finitely repeated prisoner's dilemma: Experimental evidence. *The Economic Journal*, 103(418), 570-585.

Apps, M. A., Rushworth, M. F., & Chang, S. W. (2016). The Anterior Cingulate Gyrus and Social Cognition: Tracking the Motivation of Others. *Neuron*, 90(4), 692-707.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science*, 211(4489), 1390-1396.

Balliet, D., Mulder, L. B., & Van Lange, P. A. (2011). Reward, punishment, and cooperation: a meta-analysis. *Psychological Bulletin*, 137(4), 594-615.

Barclay, P., & Willer, R. (2007). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society of London B: Biological Sciences*, 274(1610), 749-753.

Baron, J. (1997). The illusion of morality as self-interest: A reason to cooperate in social dilemmas. *Psychological Science*, 8(4), 330-335.

Baron, J. N., & Kreps, D. N. (1999). *Strategic human resources: Frameworks for general managers.* New York, NY: Wiley Publishers

Batson, C. D. (2011). *Altruism in humans*. Oxford, UK: Oxford University Press.

Bear, A., & Rand, D. G. (2016). Intuition, deliberation, and the evolution of cooperation. *Proceedings of the National Academy of Sciences*, 113(4), 936-941.

Bem, D. J. (1967). Self-perception: An alternative interpretation of cognitive dissonance phenomena. *Psychological Review*, 74(3), 183-200.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1), 122-142.

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, *145*(12), 1654.

Blanco, M., Engelmann, D., Koch, A. K., & Normann, H. T. (2014). Preferences and beliefs in a sequential social dilemma: a within-subjects analysis. *Games and Economic Behavior*, *87*, 122-135.

Blount, S. (1995). When social outcomes aren′t fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, *63*(2), 131-144.

Brehm, J. W., & Cole, A. H. (1966). Effect of a favor which reduces freedom. *Journal of Personality and Social Psychology*, 3(4), 420-426.

Bolton, G. E. (1991). A comparative model of bargaining: Theory and evidence. *The American Economic Review*, 1096-1136.

Bolton, G. E., Brandts, J., & Ockenfels, A. (1998). Measuring motivations for the reciprocal responses observed in a simple dilemma game. *Experimental Economics*, 1(3), 207-219.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American economic review*, 90(1), 166 – 193.

Bowles, S., & Gintis, H. (2002). Behavioural science: homo reciprocans. *Nature*, *415*(6868), 125.

Brockner, J., & Wiesenfeld, B. M. (1996). An integrative framework for explaining reactions to decisions: interactive effects of outcomes and procedures. *Psychological Bulletin*, 120(2), 189-208.

Camerer, C. F., & Thaler, R. H. (1995). Anomalies: Ultimatums, dictators and manners. *Journal of Economic Perspectives*, 9(2), 209-219.

Campbell, D. T. (1972). On the Genetics of Altruism and the Counter-Hedonic Components in Human Culture. *Journal of Social Issues*, 28(3), 21-37.

Campbell, D. (1975). On the conflicts between biological and social evolution and between psychology and moral tradition. *American Psychologist*, 30(12), 1103 – 1126.

Caporael, L. R., Dawes, R. M., Orbell, J. M., & Van de Kragt, A. J. (1989). Selfishness examined: Cooperation in the absence of egoistic incentives. *Behavioral and Brain Sciences*, 12(4), 683-699.

Caruso, E. M., Epley, N., & Bazerman, M. H. (2006). The costs and benefits of undoing egocentric responsibility assessments in groups. *Journal of Personality and Social Psychology*, 91(5), 857-871.

Cassar, L. and Meier, S. (2017). Intentions for Doing Good Matter for Doing Well: The (Negative) Signaling Value of Prosocial Incentives. NBER Working Paper.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3), 817-869.

Chater, N., Vlaev, I., & Grinberg, M. (2008). A new consequence of Simpson's paradox: Stable cooperation in one-shot prisoner's dilemma from populations of individualistic learners. *Journal of Experimental Psychology: General*, 137(3), 403-421.

Clark, M. S., & Mills, J. (1979). Interpersonal attraction in exchange and communal relationships. *Journal of Personality and Social Psychology*, 37(1), 12 – 24.

Clark, K., & Sefton, M. (2001). The sequential prisoner's dilemma: Evidence on reciprocation. *The Economic Journal*, 111(468), 51-68.

Cohn, A., Fehr, E., & Götte, L. (2014). Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science*, 61(8), 1777-1794.

Cohn, A., Fehr, E., Herrmann, B., & Schneider, F. (2014). Social comparison and effort provision: Evidence from a field experiment. *Journal of the European Economic Association*, 12(4), 877-898.

Colman, A. M. (2003). Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences*, 26(2), 139-153.

Cooper, D. J., & Kagel, J. H. (2016). Other-regarding preferences. *The Handbook of Experimental Economics*, *2*, 217.

Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*(2), 260-281.

Critcher, C., Inbar, Y., & Pizarro, D. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, 4(3), 308-315.

Cropanzano, R., Byrne, Z. S., Bobocel, D. R., & Rupp, D. E. (2001). Moral virtues, fairness heuristics, social entities, and other denizens of organizational justice. *Journal of Vocational Behavior*, 58(2), 164-209.

Croson, R. T. (2000). Thinking like a game theorist: factors affecting the frequency of equilibrium play. *Journal of Economic Behavior & Organizatio*n, 41(3), 299-314.

Dana, J., Cain, D. M., & Dawes, R. M. (2006). What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100(2), 193-201.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1), 67-80.

Dabos, G. E., & Rousseau, D. M. (2004). Mutuality and reciprocity in the psychological contracts of employees and employers. *Journal of Applied Psychology*, 89(1), 52-72.

Dawes, R. M. (1980). Social dilemmas. *Annual Review of Psychology*, 31(1), 169-193.

Delton, A. W., Krasnow, M. M., Cosmides, L., & Tooby, J. (2011). Evolution of direct reciprocity under uncertainty can explain human generosity in one-shot encounters. *Proceedings of the National Academy of Sciences*, 108(32), 13335-13340.

Dufwenberg, M., & Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, 30(2), 163-182.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. Games and Economic *Behavior*, *47*(2), 268-298.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging Probable Cause. *Psychological Bulletin*, 99(1), 3-19.

Eisenberger, R., Armeli, S., Rexwinkel, B., Lynch, P. D., & Rhoades, L. (2001). Reciprocation of perceived organizational support. *Journal of Applied Psychology*, 86(1), 42-51.

Ellemers, N., Spears, R., & Doosje, B. (2002). Self and social identity. *Annual Review of Psychology*, 53(1), 161-186.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, 94(4), 857-869.

Epley, N., Caruso, E. M., & Bazerman, M. H. (2006). When Perspective Taking Increases Taking. *Journal of Personality and Social Psychology*, 91(5), 872-889.

Epley, N., & Dunning, D. (2000). Feeling" Holier Than Thou": Are Self-Serving Assessments Produced by Errors in Self-or Social Prediction?. *Journal of Personality and Social Psychology*, 79(6), 861-875.

Erev, I. & Roth, A.E. (2001), "On simple reinforcement learning models and reciprocation in the prisoner dilemma game". In Gigerenzer, G. and Selten, R. (Eds.), The Adaptive Toolbox. 215-232 Cambridge, MA: MIT Press.

Esteves-Sorenson, C. (2017). Gift exchange in the workplace: Addressing the conflicting evidence with a careful test. *Management Science*. https://doi.org/10.1287/mnsc.2017.2801

Everett, J., Pizarro, D., & Crockett, M. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*(6), 772.

Falk, A. (2007). Gift exchange in the field. *Econometrica*, 75(5), 1501-1511.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2), 293-315.

Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, *13*(1), 1-25.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415(6868)*, 137-140.

Fehr, E., & Gintis, H. (2007). Human motivation and social cooperation: Experimental and analytical foundations. *Annual Review of Sociology*, *33*, 43-64.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1993). Does fairness prevent market clearing? An experimental investigation. *The Quarterly Journal of Economics*, 437-459.

Fehr, E., Kirchsteiger, G., & Riedl, A. (1998). Gift exchange and reciprocity in competitive experimental markets. *European Economic Review*, *42*(1), 1-34.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3), 817-868.

Fein, S. (1996). Effects of suspicion on attributional thinking and the correspondence bias. *Journal of Personality and Social Psychology*, 70(6), 1164-1180.

Fein, S., & Hilton, J. L. (1994). Judging others in the shadow of suspicion. *Motivation and Emotion*, 18(2), 167-198.

Fein, S., Hilton, J. L., & Miller, D. T. (1990). Suspicion of ulterior motivation and the correspondence bias. *Journal of Personality and Social Psychology*, 58(5), 753-764.

Fetchenhauer, D., & Dunning, D. (2010). Why so cynical? Asymmetric feedback underlies misguided skepticism regarding the trustworthiness of others. *Psychological Science*, 21(2), 189 – 193.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2), 171-178.

Fishbein, M., & Ajzen, I. (1975). *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley.

Fiske, A. P. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, 99(4), 689-723.

Fletcher, J. A., & Zwick, M. (2007). The evolution of altruism: Game theory in multilevel selection and inclusive fitness. *Journal of Theoretical Biology*, 245(1), 26-36.

Gächter, S., & Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, *13*(3), 364-377.

Gilbert, D. T. (1989). Thinking lightly about others: Automatic components of the social inference process. In: *Unintended Thought* (pp. 189-211), New York, NY: Guilford Press.

Gilbert, D. T., Krull, D. S., & Pelham, B. W. (1988). Of thoughts unspoken: Social inference and the self-regulation of behavior. *Journal of Personality and Social Psychology*, 55(5), 685-694.

Gilbert, D. T., & Osborne, R. E. (1989). Thinking backward: Some curable and incurable consequences of cognitive busyness. *Journal of Personality and Social Psychology*, 57(6), 940-949.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On Cognitive Busyness: When Person Perceivers Meet Persons Perceived. *Journal of Personality and Social Psychology*, 54(5), 733-740.

Gilchrist, D. S., Luca, M., & Malhotra, D. (2016). When 3+ 1> 4: Gift structure and reciprocity in the field. *Management Science*, 62(9), 2639-2650.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, 206(2), 169-179.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The Quarterly Journal of Economics*, 115(3), 811-846.

Gluth, S., & Fontanesi, L. (2016). Wiring the altruistic brain. *Science*, 351(6277), 1028-1029.

Gneezy, U., & List, J. A. (2006). Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments. *Econometrica*, 74(5), 1365-1384.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, 161-178.

Gray, K., Ward, A., & Norton, M. (2014). Paying it forward: Generalized reciprocity and the limits of generosity. *Journal of Experimental Psychology: General*, *143*(1), 247 – 254.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, 35(1), 1-15.

Guerra, G., & Zizzo, D. J. (2004). Trust responsiveness and beliefs. *Journal of Economic Behavior & Organization*, *55*(1), 25-30.

Güth, W., & Tietz, R. (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology*, 11(3), 417-449.

Halali, E., Bereby-Meyer, Y., Meiran, N. (2014). Between self-interest and reciprocity: The social bright side of self-control failure. *Journal of Experimental Psychology: General*, 143(2), 745-754

Halevy, N. (2017). Preemptive strikes: Fear, hope, and defensive aggression. *Journal of Personality and Social Psychology*, *112*(2), 224-237.

Halevy, N., & Chou, E. Y. (2014). How decisions happen: Focal points and blind spots in interdependent decision making. *Journal of Personality and Social Psychology*, 106(3), 398-417.

Ham, J., & Vonk, R. (2011). Impressions of impression management: Evidence of spontaneous suspicion of ulterior motivation. *Journal of Experimental Social Psychology*, 47(2), 466-471.

Harrison, G. W., & McCabe, K. (1992). Testing noncooperative bargaining theory in experiments. *Research in Experimental Economics*, 5, 137-169.

Hayashi, N., Ostrom, E., Walker, J., & Yamagishi, T. (1999). Reciprocity, trust, and the sense of control a cross-societal study. *Rationality and Society*, 11(1), 27-46.

Heath, C. (1999). On the social psychology of agency relationships: Lay theories of motivation overemphasize extrinsic incentives. *Organizational Behavior and Human Decision Processes*, 78(1), 25-62.

Hein, G., Morishima, Y., Leiberg, S., Sul, S., & Fehr, E. (2016). The brain's functional network architecture reveals human motives. *Science*, 351(6277), 1074-1078.

Hilton, J. L., Fein, S., & Miller, D. T. (1993). Suspicion and dispositional inference. *Personality and Social Psychology Bulletin*, 19(5), 501-512.

Hoffman, M. L. (1981). Is altruism part of human nature?. *Journal of Personality and Social Psychology*, 40(1), 121-137.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7(3), 346-380.

Inesi, M. E., Gruenfeld, D. H., & Galinsky, A. D. (2012). How power corrupts relationships: Cynical attributions for others' generous acts. *Journal of Experimental Social Psychology, 48*(4), 795-803.

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, 32(5), 865-889.

Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist*, 34(2), 107-117.

Jones, E. E., Davis, K. E., & Gergen, K. J. (1961). Role playing variations and their informational value for person perception. *The Journal of Abnormal and Social Psychology*, 63(2), 302-310.

Jordan, J. J., Hoffman, M., Nowak, M. A., & Rand, D. G. (2016). Uncalculating cooperation is used to signal trustworthiness. *Proceedings of the National Academy of Sciences*, 113(31), 8658-8663.

Kahneman, D., Knetsch, J. L., & Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review*, 76(4), 728-741.

Kameda, T., & Nakanishi, D. (2002). Cost–benefit analysis of social/cultural learning in a nonstationary uncertain environment: An evolutionary simulation and an experiment with human subjects. *Evolution and Human Behavior*, 23(5), 373-393.

Kameda, T., Takezawa, M., & Hastie, R. (2003). The logic of social sharing: An evolutionary game analysis of adaptive norm development. *Personality and Social Psychology Review*, 7(1), 2-19.

Kameda, T., Tsukasaki, T., Hastie, R., & Berg, N. (2011). Democracy under Uncertainty: The Wisdom of Crowds and the Free-Rider Problem in Group Decision Making. *Psychological Review*, 118(1), 76-96.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107.

Kelley, H. H., & Stahelski, A. J. (1970). The inference of intentions from moves in the Prisoner's Dilemma game. *Journal of Experimental Social Psychology*, 6(4), 401-419.

Kennedy, K. A., & Pronin, E. (2008). When disagreement gets ugly: Perceptions of bias and the escalation of conflict. *Personality and Social Psychology Bulletin*, 34(6), 833-848.

Kerr, N. L. (1983). Motivation losses in task-performing groups: A social dilemma analysis. *Journal of Personality and Social Psychology*, 45(4), 819-828.

Kessler, J. B., & Meier, S. (2014). Learning from (failed) replications: Cognitive load manipulations and charitable giving. *Journal of Economic Behavior & Organization*, 102, 10-13.

Keysar, B., Converse, B. A., Wang, J., & Epley, N. (2008). Reciprocity Is Not Give and Take Asymmetric Reciprocity to Positive and Negative Acts. *Psychological Science*, *19*(12), 1280-1286.

Kiesler, S. B. (1966). The effect of perceived role requirements on reactions to favor-doing. *Journal of Experimental Social Psychology*, 2(2), 198-210.

Kiyonari, T., Tanida, S., & Yamagishi, T. (2000). Social exchange and reciprocity: Confusion or a heuristic? *Evolution and Human Behavior*, *21*, 411-427

Komorita, S., Parks, C., & Hulbert, L. (1992). Reciprocity and the induction of cooperation in social dilemmas. *Journal of Personality and Social Psychology*, *62*(4), 607-617.

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

Krebs, D. L. (1970). Altruism: An examination of the concept and a review of the literature. *Psychological Bulletin*, 73(4), 258-302.

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, 27(2), 245-252.

Kruger, J., & Gilovich, T. (1999). " Naive cynicism" in everyday theories of responsibility assessment: On biased assumptions of bias. *Journal of Personality and Social Psychology*, 76(5), 743-753.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing." *Psychological Review*, 103, 263 - 283,

Kube, S., Maréchal, M. A., & Puppe, C. (2012). The currency of reciprocity: Gift exchange in the workplace. *American Economic Review*, 102(4), 1644-1662.

Lerner, M. J. (1982). The justice motive in human relations and the economic model of man: A radical analysis of facts and fictions. In *Cooperation and helping behavior* (pp. 249-278).

Liberman, V., Samuels, S. M., & Ross, L. (2004). The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and Social Psychology Bulletin*, 30(9), 1175-1185.

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, 58, 259-289.

Lieberman, M. D., Gaunt, R., Gilbert, D. T., & Trope, Y. (2002). Reflexion and reflection: A social cognitive neuroscience approach to attributional inference. *Advances in Experimental Social Psychology*, 34(2), 199-249.

Malmendier, U., te Velde, V., & Weber, R. (2014). Rethinking reciprocity. *Annual Review of Economics*, *6*(1), 849-874.

Marlowe, F. W. (2004). Dictators and ultimatums in an egalitarian society of hunter-gatherers: The Hadza of Tanzania. *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*, 168-93.

Markle, A. (2011). Dysfunctional learning in decision processes: The case of employee reciprocity. *Strategic Management Journal*, 32: 1411-1425

McCabe, K. A., Rigdon, M. L., & Smith, V. L. (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior & Organization*, *52*(2), 267-275.

McCabe, K. A., & Smith, V. L. (2001). Goodwill accounting and the process of exchange. *Bounded rationality: The adaptive toolbox*, 319-40.

Miller, D. (1999). The norm of self-interest. *American Psychologist, 54,* 1053 – 1060.

Miller, D., & Ratner, R. (1998). The disparity between the actual and assumed power of self-interest. *Journal of Personality and Social Psychology*, *74*(1), 53 – 62.

Morehous, L. G. (1966). One-play, two-play, five-play, and ten-play runs of Prisoner's Dilemma 1. *Journal of Conflict Resolution*, 10(3), 354-362.

Murphy, R., & Ackermann, K. (2014). Social Value Orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, 18(1), 13-41.

Nisbett, R. E. & Ross, L. (1980). *Human Inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, N.J.: Prentice Hall

Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, 314(5805), 1560-1563.

Nowak, M. A., Page, K. M., & Sigmund, K. (2000). Fairness versus reason in the ultimatum game. *Science*, *289*(5485), 1773-1775.

Ochs, J., & Roth, A. E. (1989). An experimental study of sequential bargaining. *The American Economic Review*, 355-384.

Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, 46(8), 1423-1437.

Oosterbeek, H., Sloof, R., & Van De Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2), 171-188.

Orbell, J., & Dawes, R. (1981). Social dilemmas. *Progress in Applied Social Psychology*, 1, 37-66.

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial Behavior: Multilevel Perspectives. *Annual Review of Psychology*, 56, 365-392.

Pierce, J. R., Kilduff, G. J., Galinsky, A. D., & Sivanathan, N. (2013). From glue to gasoline: How competition turns perspective takers unethical. *Psychological Science*, 24(10), 1986-1994.

Pillutla, M. M., Malhotra, D., & Murnighan, J. K. (2003). Attributions of trust and the calculus of reciprocity. *Journal of Experimental Social Psychology*, 39(5), 448-455.

Porter, L. W., Pearce, J. L., Tripoli, A. M., & Lewis, K. M. (1998). Differential perceptions of employers' inducements: Implications for psychological contracts. *Journal of Organizational Behavior*, 769-782.

Pruitt, D. G. (1968). Reciprocity and Credit Building in a Laboratory Dyad. *Journal of Personality and Social Psychology*, 8(2), 143-147.

Quattrone, G. A. (1982). Overattribution and unit formation: When behavior engulfs the person. *Journal of Personality and Social Psychology*, 42(4), 593-607.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, 83(5), 1281 - 1302.

Rand, D. G. (2016). Cooperation, fast and slow: Meta-analytic evidence for a theory of social heuristics and self-interested deliberation. *Psychological Science*, 27(9), 1192-1206.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 486(7416), 427-431.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences*, 17(8), 413-425.

Rand, D., Peysakhovich, A., Kraft-Todd, G., Newman, G., Wurzbacher, O., Nowak, M., & Greene, J. (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, *5*:3677. doi:10.1038/ncomms4677

Rapaport, A. & Chammah, A. (1965). *Prisoner's dilemma: A study in conflict and cooperation*. Ann Arbor: University of Michigan Press.

Ratner, R., & Miller, D. (2001). The norm of self-interest and its effects on social action. *Journal of Personality and Social Psychology*, *81*(1), 5 – 16.

Regan, D. T. (1971). Effects of a favor and liking on compliance. *Journal of Experimental Social Psychology*, 7(6), 627-639.

Roberts, G. (1998). Competitive altruism: from reciprocity to the handicap principle. *Proceedings of the Royal Society of London B: Biological Sciences*, 265(1394), 427-431.

Sahlins, M. (2008). *The Western illusion of human nature*. Chicago: Prickly Paradigm Press.

Sahlins, M. (2013). *Culture and practical reason.* Chicago: University of Chicago Press.

Saito, K. (2015). Impure altruism and impure selfishness. *Journal of Economic Theory*, *158*, 336-370.

Schopler, J., & Thompson, V. D. (1968). Role of attribution processes in mediating amount of reciprocity for a favor. *Journal of Personality and Social Psychology*, 10(3), 243-250.

Schotter, A., & Sopher, B. (2006). Trust and trustworthiness in games: An experimental study of intergenerational advice. *Experimental Economics*, 9(2), 123-145.

Segal, U., & Sobel, J. (2007). Tit for tat: Foundations of preferences for reciprocity in strategic settings. *Journal of Economic Theory*, *136*(1), 197-216.

Selten, R., & Stoecker, R. (1986). End behavior in sequences of finite Prisoner's Dilemma supergames A learning theory approach. *Journal of Economic Behavior & Organization*, 7(1), 47-70.

Sobel, J. (2005). Interdependent preferences and reciprocity. *Journal of Economic Literature*, 43(2), 392-436.

Stanca, L., Bruni, L., & Corazzini, L. (2009). Testing theories of reciprocity: Do motivations matter?. *Journal of Economic Behavior & Organization*, *71*(2), 233-245.

Suleiman, R. (1996). Expectations and fairness in a modified ultimatum game. *Journal of Economic Psychology*, 17(5), 531-554.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations*, 33, 33-47.

Thibaut, J. W., & Walker, L. (1975). *Procedural justice: A psychological analysis*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Tinghög, G., Andersson, D., Bonn, C., Böttiger, H., Josephson, C., Lundgren, G., Daniel Vastfjall, D., Kirchler, M., & Johannesson, M. (2013). Intuition and cooperation reconsidered. *Nature*, 498 (7452), E1-E2.

Tisserand, J. C. (2014, October). Ultimatum game: A meta-analysis of the past three decades of experimental research. In *Proceedings of International Academic Conferences* (No. 0802032). International Institute of Social and Economic Sciences.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology*, 46(1), 35-57.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, 93(3), 239-257.

Trope, Y., & Alfieri, T. (1997). Effortfulness and flexibility of dispositional judgment processes. *Journal of Personality and Social Psychology*, 73(4), 662-674.

Trope, Y., & Gaunt, R. (1999). A dual-process model of overconfident attributional inferences. In: *Dual-process Theories in Social Psychology* (pp. 161-178), New York, NY: Guilford Press.

Trope, Y., & Gaunt, R. (2000). Processing alternative explanations of behavior: Correction or integration?. *Journal of Personality and Social Psychology*, 79(3), 344-354.

Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science*, 3(5), 305-309.

Tyler, T. R. (1994). Psychological models of the justice motive: Antecedents of distributive and procedural justice. *Journal of Personality and Social Psychology*, 67(5), 850-863.

Tyler, T. R. (2013). The psychology of cooperation. In E. Shafir (Ed.) *The behavioral foundations of policy*. Princeton: Princeton University Press.

Tyler, T. R., & Blader, S. L. (2001). Identity and cooperative behavior in groups. *Group Processes & Intergroup Relations*, 4(3), 207-226.

Tyler, T. & Blader, S. (2003). Procedural justice, social identity, and cooperative behavior. *Personality and Social Psychology Review*, 7, 349-361

Van de Calseyde, P., Keren, G., & Zeelenberg, M. (2014). Decision time as information in judgment and choice. *Organizational Behavior and Human Decision Processes*, 125(2), 113-122.

Vohs, K. D., Baumeister, R. F., & Chin, J. (2007) Feeling duped: Emotional, motivational, and cognitive aspects of being exploited by others. *Review of General Psychology*, 11(2), 127-141.

Woods, D., & Servátka, M. (2016). Nice to you, nicer to me: Does self-serving generosity diminish the reciprocal response?. *Experimental Economics*, 1-24.

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H. & Simunovic, D. (2012). Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity. *Proceedings of the National Academy of Sciences*, 109(50), 20364-20368.

Yamagishi, T., & Sato, K. (1986). Motivational bases of the public goods problem. *Journal of Personality and Social Psychology*, *50*(1), 67.

Yamagishi, T., Takagishi, H., Fermin, A. D. S. R., Kanai, R., Li, Y., & Matsumoto, Y. (2016). Cortical thickness of the dorsolateral prefrontal cortex predicts strategic choices in economic games. *Proceedings of the National Academy of Sciences*, 113(20), 5582-5587.

Yamagishi, T., Li, Y., Takagishi, H., Matsumoto, Y., & Kiyonari, T. (2014). In Search of Homo economicus. *Psychological Science*, *25*(9), 1699-1711.

Appendix

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 $\boldsymbol{s_i} = \{s_i^1, s_i^2, \ldots, s_i^{Ni}\}$. Specify a material payoff function O: $\boldsymbol{s_1} \times \boldsymbol{s_2} \rightarrow X_1 \times X_2$. Let $\Delta(X_i)$ denote the space of lotteries over X_i. Let Σ_i denote the space of mixed strategies of player i, and extend O to be from mixed strategies to lotteries accordingly. Finally, let each player have preferences $\succeq_i^O$ over $\Delta(X_i)$.

Our model builds on the 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 Σ_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 we examine players whose preferences give rise to a particular functional form for the total utility, $v_{i,\sigma^*}(\sigma_i)$, that player i receives by playing strategy σ_i in the context of the anticipated strategy profile σ^*:

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

A player's total utility is the sum of his material and psychological payoffs. For instance, 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 λ_i, the more player i cares about reciprocating kindness with kindness and unkindness with unkindness. Second, as we next detail, M_{i,σ^*} and M_{ij,σ^*} are player i's assessments of his and his counterpart's motivations. Combining motivation scores multiplicatively allows for reciprocity.

The motivation scores M_{i,σ^*} and M_{ij,σ^*} 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, $M_{i,\sigma^*}(\sigma_i)$. Denote by $\sigma_i(s_i)$ the probability that a strategy σ_i assigns to each pure strategy s_i, and let $supp(\sigma_i)$ denote its support, i.e., the set of all s_i for which $\sigma_i(s_i) > 0$. For each pure strategy profile (s_i, s_j) in $supp(\sigma_i) \times supp(\sigma_j^*)$, let $\boldsymbol{A_{i,\sigma^*}}(s_i, s_j)$ denote the set of all alternative pure strategies s_i' in $\boldsymbol{s_i}$ 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 $\boldsymbol{A_{i,\sigma^*}}(s_i, s_j)$, associate a basic motivation score $b_{i,\sigma^*}(s_i', s_j)$ given by

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

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 $\left|\boldsymbol{A}_{i,\sigma^*}(s_i, s_j)\right| > 0$, let the partial motivation score $m_{i,\sigma^*}(s_i, s_j)$ be given by

$$m_{i,\sigma^*}(s_i, s_j) = \frac{\sum_{s_i' \in \boldsymbol{A}_{i,\sigma^*}(s_i, s_j)} b_{i,\sigma^*}(s_i')}{\left|\boldsymbol{A}_{i,\sigma^*}(s_i, s_j)\right|}.$$

If $\left|\boldsymbol{A}_{i,\sigma^*}(s_i, s_j)\right| = 0$, 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 (σ_i, σ_j^*) assigns to each pure strategy profile (s_i, s_j) in $supp(\sigma_i) \times supp(\sigma_j^*)$:

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

Note that motivation scores consider <u>all</u> alternative strategies. Even strategies that may appear implausible, such as cooperation in response to defection in a sequential prisoners' dilemma, can thus 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.

$M_{ij,\sigma^*}(\sigma_j^*)$ captures *i*'s assessment of *j*'s anticipated strategy and is calculated similarly. For each pure strategy profile (s_i, s_j) in $supp(\sigma_i^*) \times supp(\sigma_j^*)$, let $\boldsymbol{A_{j,\sigma^*}}(s_i, s_j)$ denote the set of all alternative pure strategies s_j' which yield 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 $\boldsymbol{A_{j,\sigma^*}}(s_i, s_j)$, 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 θ_i (not θ_j). Again, for $\left|\boldsymbol{A}_{j,\sigma^*}(s_i, s_j)\right| = 0$, let $m_{ij,\sigma^*}(\sigma_j^*) = 0$, and for $\left|\boldsymbol{A}_{j,\sigma^*}(s_i, s_j)\right| > 0$, let the partial motivation score $m_{ij,\sigma^*}(s_i, s_j)$ be given by

$$m_{ij,\sigma^*}(s_i, s_j) = \frac{\sum_{s_j' \in \boldsymbol{A}_{j,\sigma^*}(s_i, s_j)} b_{ij,\sigma^*}(s_j')}{\left|\boldsymbol{A}_{j,\sigma^*}(s_i, s_j)\right|}.$$

The motivation score $M_{ij,\sigma^*}(\sigma_j^*)$ that player *i* assigns to player *j*'s anticipated strategy is then given by

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

Segal and Sobel (2007) define Nash equilibrium as an anticipated strategy profile $\sigma^* = (\sigma_i^*, \sigma_j^*)$ in which σ_i^* and σ_j^* are the players' preferred strategies conditional on σ^*, so that neither player has an incentive to unilaterally deviate. Lemma 1 in Segal and Sobel (2007) can be used to show that when players' preferences can be represented via the total utilities in Equation 1, every two-player game has at least one such Nash equilibrium.

Table 1. Motivation scores for a player's strategy as a function of how it impacts her counterpart as well as the player herself.

	Help Self	Hurt Self
Help Other	$+1 - \theta_i$	+1
Hurt Other	$-1 + \theta_i$	–1

Table 2. Summary statistics for the standard, sequential and de-coupled games in Experiment 2.

	1st-mover *C*	2nd-mover reciprocity of *C* with *C*	2nd-mover reciprocity of *D* with *D*
Standard Sequential	16/27 (59.3%)	20/32 (62.5%)	21/28 (75.0%)
De-Coupled	23/62 (37.1%)	38/45 (84.4%)	77/103 (74.8%)

Table 3. Summary statistics for the standard and endogenous-sequencing games in Experiment 3.

	1st-mover *C*	2nd-mover reciprocity of *C* with *C*	2nd-mover reciprocity of *D* with *D*	Simultaneous game if timer expired, *C*	Median // Mean time elapsed before 1st-move (*% of total time*)
Standard Sequential	30/45 (66.7%)	21/31 (67.7%)	15/16 (93.8%)	N/A	N/A
Endogenous-Sequencing, Aggregated	53/69 (76.8%)	39/50 (78.0%)	14/17 (82.4%)	1/7 (14.3%)	N/A (*15.0%* // *24.8%*)
Endogenous-Sequencing, 20 Seconds	16/20 (80.0%)	9/13 (69.2%)	4/5 (80.0%)	0/2 (0.0%)	6.0 // 8.2 seconds (*30.0%* // *41.1%*)
Endogenous-Sequencing, 60 Seconds	21/29 (72.4%)	17/20 (85.0%)	6/8 (75.0%)	1/5 (20.0%)	9.0 // 10.8 seconds (*15.0%* // *18.1%*)
Endogenous-Sequencing, 120 Seconds	16/20 (80.0%)	13/17 (76.5%)	4/4 (100%)	N/A	13.0 // 24.0 seconds (*11.0%* // *19.9%*)

Table 4. Positive reciprocity in the standard 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 3. The statistical tests in the right-hand column compare the positive reciprocity rate under endogenous sequencing with that of the standard game.

		2nd-movers responding to 1st-mover *C*:	
	1st-mover *C*	2nd-mover reciprocity of *C* with *C*	Fisher's Exact *p*-value vs. seq. game
Standard Sequential	30/45 (66.7%)	21/31 (67.7%)	
≤ 5% of Countdown Elapsed Prior to 1st-Move	8/11 (72.7%)	9/9 (100%)	.081
≤ 10% of Countdown Elapsed Prior to 1st-Move	22/25 (88.0%)	21/22 (95.5%)	.017
≤ 15% of Countdown Elapsed Prior to 1st-Move	29/34 (85.3%)	26/29 (89.7%)	.061
≤ 20% of Countdown Elapsed Prior to 1st-Move	34/39 (87.2%)	30/33 (90.9%)	.029
> 20% of Countdown Elapsed Prior to 1st-Move	19/30 (63.3%)	9/17 (52.9%)	N/A

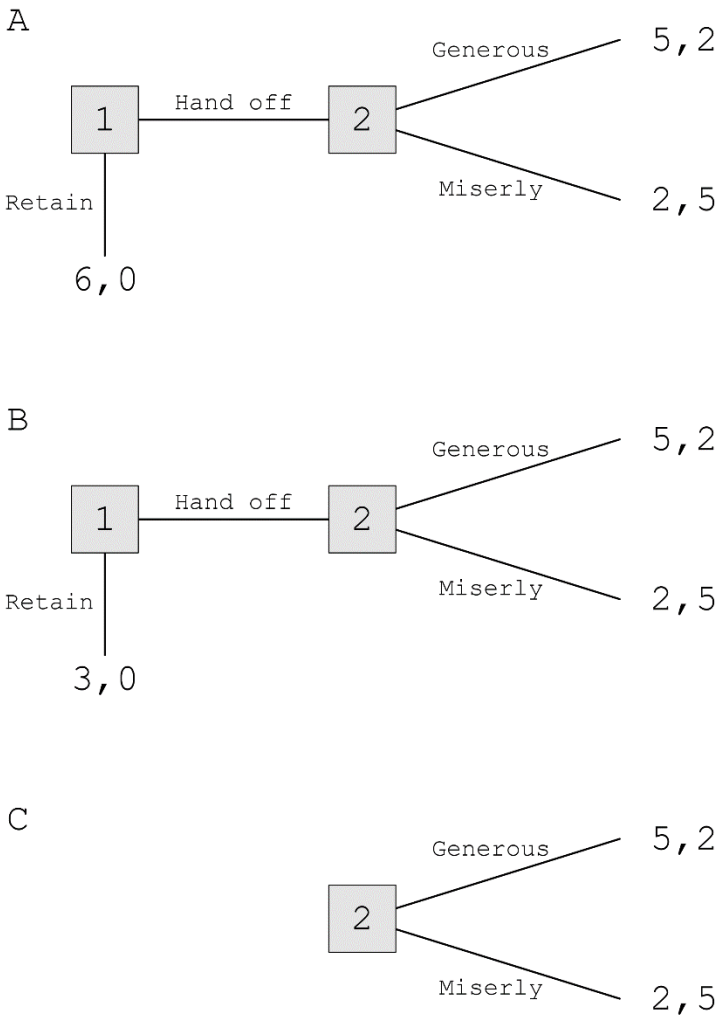

Figure 1. Handoff games in which a first-mover's handoff is either unambiguously socially motivated (Panel A) or ambiguously motivated (Panel B), and an analogous simplified dictator game (Panel C).

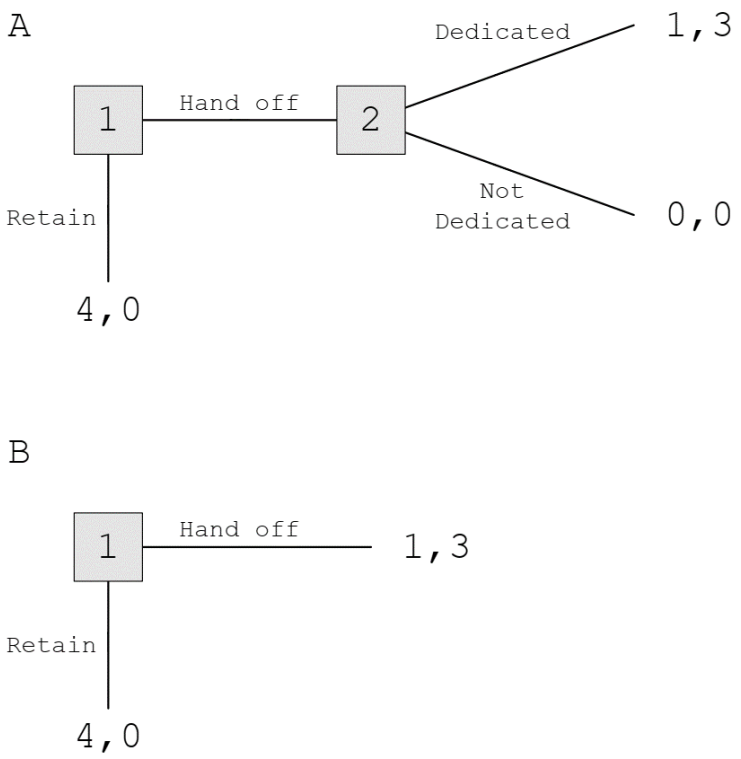

Figure 2. Handoff game in which a second-mover's acceptance of her counterpart's handoff is ambiguously motivated (Panel A), and an analogous simplified dictator game (Panel B).

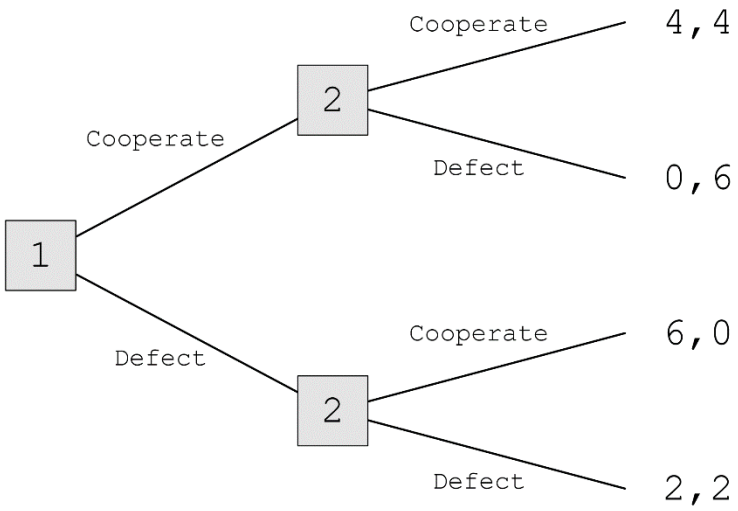

Figure 3. A sequential prisoners' dilemma.

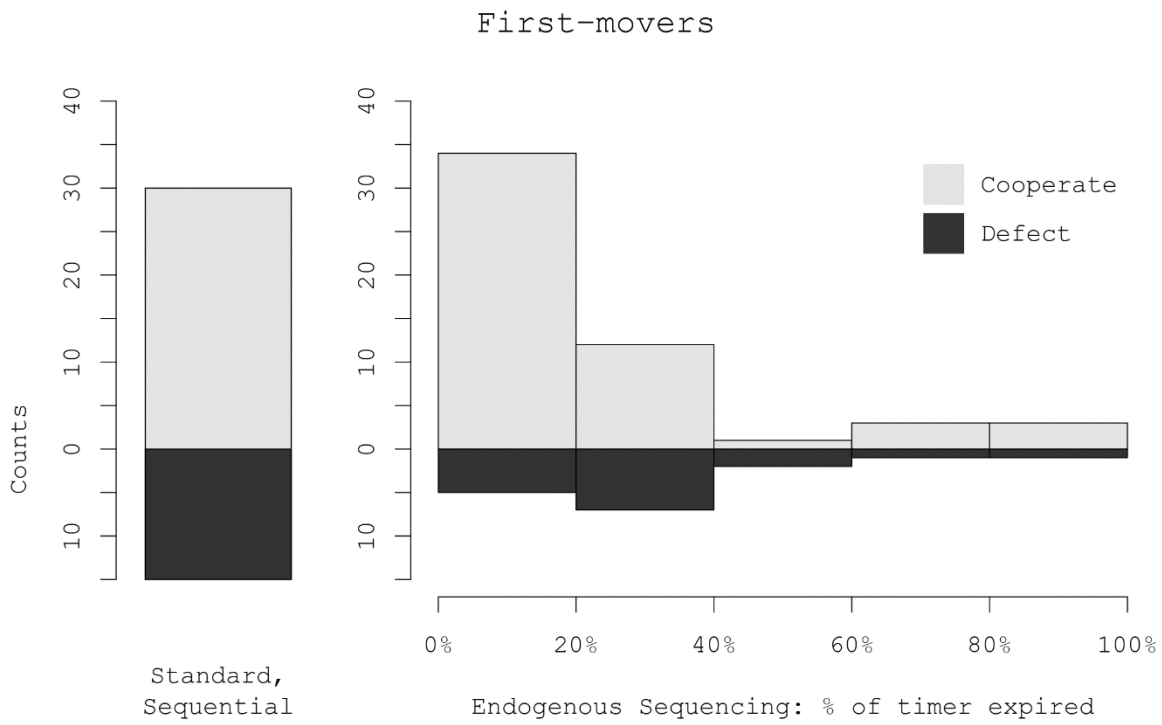

Figure 4. First-movers' behavior in the standard, sequential (left) and endogenous-sequencing (right) games in Experiment 3.

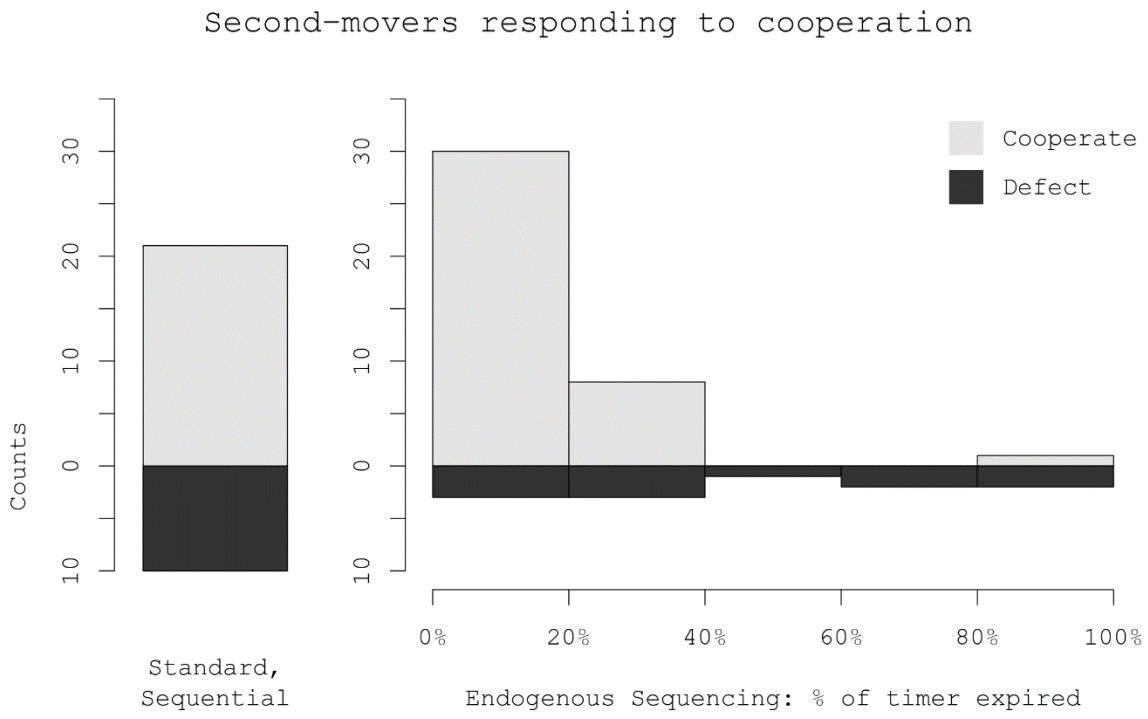

Figure 5. Second-movers' responses to first-move cooperation, in the standard, sequential (left) and endogenous-sequencing (right) games in Experiment 2.

	Left (Generous)	Right (Miserly)
Top (Hand off)	2,2	0,3
Bottom (Retain)	1,1	1,1

Figure 6. Game in Guerra and Zizzo (2004) that resembles a simultaneous handoff game.