Experimental Economics (2025), **28**, 758–781
doi:10.1017/eec.2025.10009

SPECIAL ISSUE ARTICLE

Positive reciprocity when motives are ambiguous

Johannes Müller-Trede[1] and Yuval Rottenstreich[2]

[1]IESE Business School, University of Navarra, Pamplona, Spain
[2]Rady School of Management, University of California, San Diego, CA, USA
Corresponding author: Johannes Müller-Trede; Email: jmuller@iese.edu

(Received 30 November 2023; revised 6 October 2024; accepted 28 April 2025)

Abstract

We present and test a model of reciprocity in which people are more likely to repay good treatment to the extent they judge it as motivated by true caring rather than tactical self-interest. The model's key contributions stem from how it handles ambiguously motivated behavior. It allows people to maintain divergent hypotheses: They can view behavior as driven by caring, self-interest, or a mix thereof. In contrast, previous analyses resolve rather than maintain ambiguity. They treat caring and self-interest as mutually exclusive hypotheses, and require that people commit to one and dismiss the other. By more realistically handling ambiguity, our model yields three benefits. First, it accommodates intuitive patterns of play that existing analyses do not and which we experimentally corroborate. These patterns reflect intermediate inclinations to reciprocate ambiguously motivated positive behavior. Second, it challenges conventional interpretations of long-studied phenomena, including unraveling in finitely iterated prisoners' dilemmas, substantial offers in ultimatum games, and gift exchange. Third, it highlights how diversity in perceptions – the same action can appear generous to one person and miserly to another – is empirically consequential. Under conventional interpretations and without accounting for diverse perceptions, the aforementioned phenomena have been viewed as inconsistent with a taste for repaying good treatment. Our model shows that they are entirely consistent with a nuanced form of this taste: a desire to repay good treatment that seems to largely reflect genuine caring.

Keywords: Attribution; Cooperation; Reciprocity; Social preferences

JEL Codes: A13; C70; D63

1. Introduction

Consider the following scenario. Elliott and Michelle are acquaintances who work as freelancers in the same field. Each of them is about to commence a valuable project. They cross paths. In talking, they realize that Elliott's project would be of greater value to Michelle and vice versa. They also realize that each could hand their project off to the other. 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. That is, after Elliott transfers his project to Michelle, she retains her project for herself. She does not reciprocate.

Much research stresses two complementary explanations for Michelle's behavior and for decisions concerning reciprocity more generally (Cabral et al., 2014). First, people often make tradeoffs between self-interest and other-regard (Fehr et al., 2002; Fehr & Gintis, 2007). Michelle may care about how she treats Elliot and may like to repay kindness with kindness, yet she decides not to transfer her project to him because she thinks it is too valuable to give up. Second, people sometimes pursue their self-interest tactically (Axelrod & Hamilton, 1981; Kreps et al., 1982). Michelle may not see a personal

© The Author(s), 2025. Published by Cambridge University Press on behalf of the Economic Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (http://creativecommons.org/licenses/by-nc-nd/4.0), which permits non-commercial re-use, distribution, and reproduction in any medium, provided that no alterations are made and the original article is properly cited. The written permission of Cambridge University Press must be obtained prior to any commercial use and/or adaptation of the article.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

advantage in transferring her project to Elliott. Had she calculated that transferring it to him would induce him to give her additional projects in the future, she might have done so.

By a third explanation, people assess the motivations underlying their counterpart's treatment of them (Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006; Orhun, 2018; Rabin, 1993; Sobel, 2005; Stanca et al., 2009). Just as Michelle could have reciprocated out of either caring or tactical self-interest, she might think about whether Elliott was driven by each of these factors. Did he give his project to her because he genuinely wanted to treat her nicely? Or to elicit and profit from her reciprocal cooperation? In essence, Michelle might ask whether Elliott was trying to help her or influence her. She might be more likely to reciprocate if she views him as trying to help her rather than influence her (McCabe, Rigdon, & Smith, 2003; McCabe & Smith, 2000; Woods & Servátka, 2016).

In this article, we present a game-theoretic model of reciprocity in which players consider each other's motivations. The model's key contribution lies in how it handles ambiguity regarding these motivations.[1] Specifically, it allows players to maintain divergent hypotheses: Good treatment can be perceived as caring, self-interested, or a mix thereof. Michelle, for instance, may not be willing to fully attribute Elliot's behavior to either tactical self-interest or genuine caring. In the model, players become increasingly inclined to reciprocate good treatment as they increasingly perceive good treatment to definitively reflect caring rather than self-regard.

Several existing game-theoretic analyses also consider motivations. But they handle ambiguity very differently. Rather than allowing people to maintain ambiguity about motives, they require that people resolve this ambiguity (Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006; Rabin, 1993; see also Sobel, 2005). In other words, past analyses treat the hypotheses of caring and tactical self-interest as mutually exclusive and require that people commit to one and dismiss the other.

This seems unrealistic – and makes it difficult to capture commonplace tendencies, such as intermediate inclinations to respond positively to ambiguously motivated behavior. For example, in Rabin's (1993) theory, anyone who stands to gain materially from their own actions is judged definitively self-interested and entirely undeserving of reciprocity, no matter what effects their actions have on others. That is, behavior that helps someone else but could potentially also help oneself – much like Elliott's – is always deemed unworthy of reciprocity. Rabin (1993, p. 1296) himself points out this issue. As we thoroughly detail later, similar issues arise in the theories of Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006).

By more realistically handling ambiguity, our model provides three benefits. First, it accommodates intuitive patterns of play which existing theories of reciprocity do not, and which we corroborate in a pair of experiments. These patterns reflect intermediate inclinations to reciprocate to ambiguously motivated good treatment.

Second, our model challenges conventional interpretations of several long-studied phenomena, including unraveling in finitely iterated prisoners' dilemmas and even-split or nearly even-split offers in ultimatum games. It thereby nuances influential work which has questioned whether people have much taste for responding to good treatment with good treatment (Charness & Rabin, 2002, 2005; Malmendier et al., 2014; Offerman, 2002; for less questioning views, see Cox et al., 2008; Falk et al., 2008). Under conventional interpretations, unraveling has been seen as inconsistent with this taste; even split and nearly even-split ultimatum offers have not been interpreted as evidence for it. However, our model shows that both phenomena readily follow from a motivation-based, ambiguity-influenced form of such a taste: People do like to repay good treatment, but more so when it reflects genuine caring rather than calculated self-interest.

Third, our model highlights that the same action can appear generous to one person and miserly to another. It also points to how this diversity in perceptions can be empirically consequential. In doing

[1] It is of course possible for ambiguity to emerge regarding many motivations, beyond the genuine caring and tactical self-interest on which we focus. Rapoport (1987), for instance, cites ambiguity about two drivers of defection in step-level social dilemmas: desire to free ride on others' contributions to the common pool and fear of wasting one's contribution to the pool.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

so, it again nuances work that has been skeptical about people's taste for repaying good treatment. Well-known results from gift exchange and trust games have been taken as evidence against such a taste. However, once diverse perceptions are accounted for, these results are entirely consistent with a desire to repay good treatment that seems to largely reflect genuine caring.

1.1. From here

The remainder of the article is organized as follows. We first present our model and further explore how it departs from existing game-theoretic analyses that explore people's taste for repaying good treatment. We illustrate the model's workings via a set of sequential games. Across this set, it predicts an intuitive pattern of play that includes the aforementioned intermediate inclination to reward ambiguously motivated prosocial behavior.

We corroborate this pattern in Experiment 1. Then, in Experiment 2, we consider a novel, sequential prisoners' dilemma with "endogenous sequencing." This game does not assign players the roles of first and second mover (cf. Murphy et al., 2006). Instead, either player can decide to make the first move or not. As the interaction unfolds, a player may first-move cooperate, first-move defect, or simply continue to wait and see what happens. Once a player has acted, the other player gets to respond. This setup is consistent with many social interactions outside the laboratory. Elliott initiating contact with Michelle was not a given, for example. She could have initiated contact with him instead. Our model makes the intuitive prediction that under endogenous sequencing, first-move cooperation is more likely to be reciprocated the faster it occurs. This prediction arises because as first-move cooperation is more and more delayed, it becomes more and more ambiguously motivated. Our results corroborate this prediction.

In assessing the results of our experiments, we discuss a theory proposed by Levine (1998), which applies to many of the same settings as our model. But whereas we are concerned with reciprocity and a preference for repaying good treatment, Levine focuses on altruism and a preference for rewarding altruistic people. Because of this subtle yet fundamental difference between the two accounts, Levine's theory cannot account for the results of our experiments.

To conclude, we discuss findings concerning gift exchange, trust games, repeated prisoners' dilemmas, and ultimatum games. Research on these settings has downplayed people's desire for reciprocating good treatment. We show how our model's more realistic handling of ambiguity indicates that a nuanced form of this desire, a preference for reciprocating good treatment that seems to largely reflect caring, may be important in each setting.

2. Model

Our model builds on classic psychological research concerning attributions and discounting (Einhorn & Hogarth, 1986; Fishbein & Ajzen, 1975; Jones, 1979; Kelley, 1973; Nisbett & Ross, 1980). It presumes 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.

Mathematically, we build on a framework from Segal and Sobel (2007). We include a formal statement of the model in the Appendix. Here, we explain its workings via the games in Fig. 1, which together spotlight responses to ambiguously motivated other-regarding 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

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

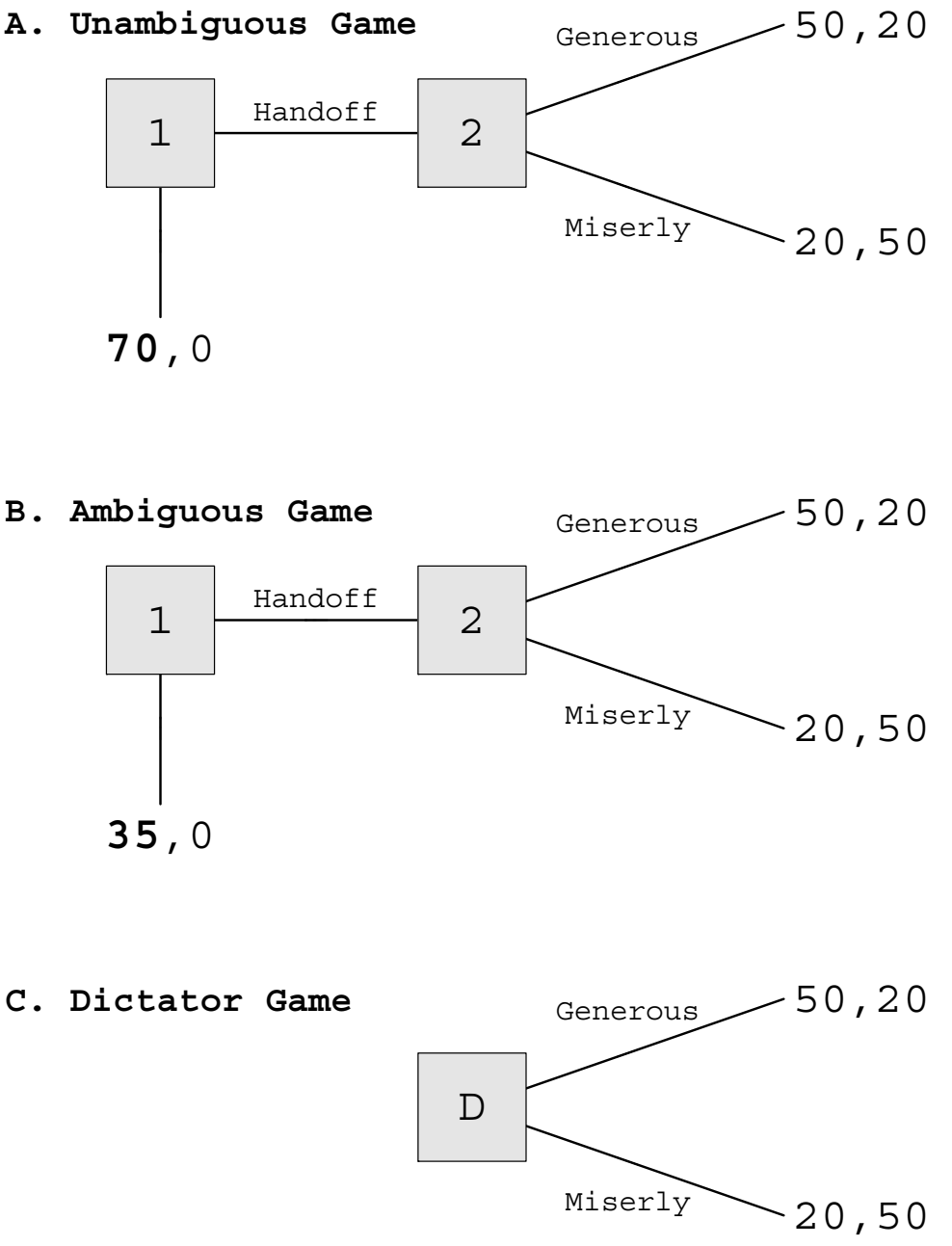


Fig. 1 Handoff games in which a first mover's handoff is either unambiguously motivated (panel A) or ambiguously motivated (panel B), and a parallel dictator game (panel C)

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.

Fig 1B is the best place to start. In the game there, the first mover can unilaterally garner 35 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 70 between

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

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

them. She can act generously, granting 50 to the first mover and 20 to herself, or miserly, granting the reverse payoffs.

In Fig. 1A, the first mover's unilateral option is worth 70 to him; all else remains as in Fig. 1B. Meanwhile, Fig. 1C is a simplified dictator game (Camerer & Thaler, 1995; Dana et al., 2006; Kahneman et al., 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 70.

2.1. Motivations and motivation scores

Suppose the individuals in Fig. 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?

Our 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 Fig. 1B, the second mover would have to act miserly in any such strategy. 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 50 and would receive less – 20 – if the second mover were miserly. Furthermore, relative to the alternative, the second mover's strategy materially "hurts" the second mover herself. She receives 20 but would receive 50 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 stresses that in Fig. 1B, this strategy can be viewed questioningly. First, the only alternative has him act unilaterally. Second, handing off control yields the players 50 and 20, respectively, rather than the 35 and 0 they would receive if the first mover acted unilaterally. So, while the first mover materially "helps" the second mover ($20 > 0$), he also materially "helps" himself ($50 > 35$). 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 $0 \leq \theta_i \leq 1$. This parameter 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.[2]

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

[2]In the game of Fig. 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. We discuss this aspect of the model in more detail in the context of Experiment 2, where it plays a key role in generating our behavioral predictions.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

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 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.[3]

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 is his nonmaterial, socioemotional payoff. It 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. $\lambda_i \geq 0$ 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. 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 sufficiently large and her θ is sufficiently small, then the game of Fig. 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 forms an equilibrium, for any values of λ and θ.

Note that when players are either mutually kind or mutually unkind, our model has 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 the positive reciprocity on which we focus. It is not compelling in the case of negative reciprocity. Modeling negative reciprocity when motives are ambiguous would require a different approach. For instance, the basic logic underlying our motivation scores could be added to the models by Rabin (1993) or Dufwenberg and Kirchsteiger (2004).

Before we derive our model's predictions for the games in Fig. 1, we note that our model reduces extensive form games to their normal form. Unlike Dufwenberg and Kirchsteiger's (2004) theory of reciprocity, which is fully history-dependent, it thus cannot address issues concerning how players might revise assessments of motivations off the equilibrium path. But as suggested by the above discussion of motivation scores, and as we explain in more detail below, anticipated actions off the path of play nevertheless factor into the assessment of motivations.

2.2. Altering the unilateral action impacts attributions

How does increasing the first mover's incentive for unilateral action from 35 to 70 (Fig. 1B vs. 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 20. But the first mover's social motivation is now signaled more clearly: By handing off control, he ends up with 50 rather than 70. 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

[3] As we detail 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).

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

must instead reflect genuine caring. As long as $\theta > 0$, the second mover should therefore be more likely to reciprocate with generosity in Fig. 1A than Fig. 1B.

There are many additional other-regarding motivations beyond a desire for reciprocity. With that in mind, consider the simplified dictator game in Fig. 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 additional other-regarding motivations. 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 (Falk et al., 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 et al., 2006). In Fig. 1C, the erstwhile second mover's choice between generous and miserly behavior may be seen as tapping a baseline level of other-regard from these two sources. The interest in reciprocity captured by our model will add to this baseline.

Our model therefore makes the intuitive, overall prediction that from Fig. 1C to 1B to 1A, as first movers forego an increasingly attractive unilateral option, second movers should become increasingly likely to respond generously. In Fig. 1C, the erstwhile first mover cannot impact the second mover. He is therefore judged neutrally and as not meriting reciprocity. In Fig. 1A, by handing off, the first mover helps the second mover and hurts himself. His behavior is thus judged as unquestionably kind and fully deserving of reciprocity. Fig 1B engenders a middle ground. By handing off in this game, the first mover helps the second mover but also helps himself. His motives are therefore ambiguous. Consequently, his kindness is discounted, and his behavior is viewed as intermediately deserving of reciprocity.

2.3. Comparisons with prior accounts of reciprocating good treatment

As we have mentioned, existing game-theoretic analyses have trouble with middle ground. By 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. In our handoff games, this approach implies that a first mover forgoing 35 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 Fig. 1C and 1B.

Dufwenberg and Kirchsteiger's (2004) theory views a person as fully deserving of positive reciprocity if he helps his counterpart, and his chosen actions expose him to a potential material loss relative to his alternative actions. It views a person as fully undeserving of positive reciprocity if he helps his counterpart but his actions do not leave him exposed. It thus casts both a first mover who foregoes 35 and a first mover who foregoes 70 as fully deserving of positive reciprocity. By giving up either payoff, the first mover helps the second mover and leaves himself vulnerable to receiving only 20. In sum, Dufwenberg and Kichsteiger's (2004) analysis implies that second movers will behave equivalently across Fig 1A and 1B.[4]

[4]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, and Experiment 1 concerns the first limitation. Our closing discussion of positive reciprocity in the ultimatum game speaks to the second limitation.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

Table 2 Motivation scores in our model and associated predictions about reciprocity, for both Experiments 1 and 2. The motivation scores reflect the strategy profiles and equilibria detailed in the text and formally derived in an online supplement available at https://osf.io/7k5n9/

	Motivation Scores		*Reciprocity*
Experiment 1	1st-Mover Handoff	2nd-Mover Generosity	Predicted 2nd-Mover Generosity
Unambiguous Handoff Game (Fig. 1A)	$+1$	$+1$	Highest
Ambiguous Handoff Game (Fig. 1B)	$1-\theta_i$	$+1$	Intermediate
Dictator Game (Fig. 1C)	N/A	$+1$	Lowest
Experiment 2	1st-Mover Cooperation	2nd-Mover Cooperation	Predicted 2nd-Mover Cooperation
Sequential Prisoners' Dilemma	$1-\theta_i$	$+1$	Lower
Endogenous-Sequencing Prisoners' Dilemma, "Fast"	$+1$	$+1$	Higher

Falk and Fischbacher's (2006) theory of reciprocity and social comparisons is concerned with considerations of equity. It holds that 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 setup distinguishes between Fig. 1C and the two handoff games because the erstwhile first mover in Fig. 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 Fig. 1A and 1B.

In sum, regarding second-mover behavior, Rabin's (1993) theory equates Fig. 1B and 1C, while Dufwenberg and Kirchsteiger's (2004) and Falk and Fischbacher's (2006) theories equate Fig. 1A and 1B. Among accounts of reciprocity, only our model predicts that the frequency of second-mover generosity in Fig. 1B will fall in between that of Fig. 1A and 1C. We will test this prediction in Experiment 1.

3. Open practices statement

The data, analysis code, and study materials for both experiments reported here are publicly available at the project's Open Science Framework (OSF) website: https://osf.io/7k5n9/. Experiment 1 was preregistered at https://doi.org/10.17605/OSF.IO/GN6H8.

4. Experiment 1: Handoff games

4.1. Method

4.1.1. Preregistration.

We preregistered our experimental design, data collection, exclusion criteria, and statistical methodology at https://doi.org/10.17605/OSF.IO/GN6H8. In line with our model's predictions (Table 2), we preregistered that second movers in the unambiguous game (Fig. 1A) would be more generous than second movers in the ambiguous game (Fig. 1B), who would in turn be more generous than dictators in the dictator game (Fig. 1C).

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

Participants. We recruited a total of 7,792 US adults on the online platforms Connect and Amazon Mechanical Turk (51.8% identified as female, 47% as male, and 1.2% as nonbinary or other gender; $M_{age} = 41.1$). For recruitment on Mechanical Turk, we used CloudResearch (Litman et al., 2017). This sample size reflects a preregistered target that we explain below. Each participant received $.60 plus a bonus payment based on the outcome of their game.

As preregistered, we excluded 809 or 10.4% of participants for failing at least one of two attention checks. The exclusion rate differed significantly across conditions, $\chi^2(1, 7792) = 41.1$, $p < .0001$. But importantly, it did not differ significantly across the three critical conditions, $p = .9$, and was lower in these conditions (7.5%) than in most other conditions. After exclusions, our final sample included 6,983 participants. All results and their statistical significance remain qualitatively unchanged in analyses that include all participants.

4.1.2. Procedure

We implemented the games in Fig. 1 with payoffs in US cents. Data collection proceeded in three waves, each of which consisted of several distinct tasks on the online platforms (i.e., subsamples). Each wave was run on a single day, and the three waves were run on consecutive days. In Wave 1, we randomly assigned participants to be either the first mover in the unambiguous handoff game in Fig. 1A, the first mover in the ambiguous handoff game in Fig. 1B, or the passive counterpart to a dictator in Fig. 1C. Wave 2 featured the experiments' three critical conditions: Participants were randomly assigned to be either the second mover counterpart to a first mover who opted in to the unambiguous handoff game (Fig. 1A), the second mover counterpart to a first mover who opted in to the ambiguous handoff game (Fig. 1B), or the dictator (Fig. 1C). Our model's predictions concern participants in these critical conditions. Finally, in Wave 3 we randomly assigned participants to be the second mover counterpart to either a first mover who opted out for 70 cents (Fig. 1A) or a first mover who opted out for $.35 (Fig. 1B). These participants could not affect the outcome of their game. Note that this scheme deviates slightly from our preregistered strategy. We had planned for the second wave to include the passive second movers along with the active second movers and the dictators. Ultimately, we decided to collect the critical data from the latter prior to collecting the data from the passive second movers because we worried about the data quality as we went deeper into the participant pool.

We preregistered to terminate data collection once we reached a target of 700 first movers opting in to each of the handoff games and 350 dictators. Because of operational issues arising from the sheer size of the experiment, we ended up with slightly larger subsamples in the two handoff games. Our results and their statistical significance remain qualitatively unchanged if we restrict our analyses only to the first 700 handoff games involving first movers who opt in and the first 350 dictator games.

4.1.3. Materials

The experiment consisted of Qualtrics web forms (). Dictators, first movers, and second movers responding to first movers who had opted in were informed that they would make a decision that would affect both themselves and another participant. On a single screen and in everyday language, they were then presented with instructions explaining (one of) the games in Fig. 1. The first mover's unilateral action was described as "opting out" and "receiving $.70 [.35]," while the second mover would "not receive anything." The first mover's engagement was described as "opting in" so that the second mover could decide between "two ways of splitting $.70." In the critical conditions involving dictators and second movers responding to first-mover opt-ins, the order in which the different actions were presented (left vs. right) was randomized. Passive participants in the dictator games and second movers matched with first movers who had opted out were presented with the same game descriptions. They were informed of the decision made by their counterpart but were not themselves asked to make any decision.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

Table 3 Proportions of first-mover opt-ins and second-mover generosity in Experiment 1, for participants who passed both attention checks

	First-Movers Opting In	Dictator and Second-Mover Generosity
Unambiguous Game (Fig. 1A)	32.6% (623/1,911)	59.2% (388/655)
Ambiguous Game (Fig. 1B)	48.8% (600/1,229)	50.7% (330/651)
Dictator Game (Fig. 1C)	N/A	19.8% (64/323)

All participants then answered two simple attention checks. They also provided demographic information and indicated what kind of device (laptop vs. tablet vs. phone vs. other) they were using. The complete sequence of instructions from every condition is available at https://osf.io/7k5n9/.

4.2. Results

In the ambiguous game, in which first movers could opt out for $.35, 48.8% of first movers opted in (Table 3). In the unambiguous game, where first movers could opt out for $.70, only 32.6% of first movers opted in. The difference across conditions was statistically significant, $\chi^2(1, 3140) = 82.8$, $p < .0001$.

More critically, consider second movers and dictators. As predicted by our model, and in line with our preregistration, these players were increasingly generous as we move from Fig. 1C to 1B and on to 1A. In Fig. 1C, only 19.8% of dictators acted generously (Table 3). In Fig. 1B, where a first mover's motives for opting in were ambiguous, 50.7% of second movers responded generously. The increase in generosity is significant $\chi^2(1, 974) = 85.4$, $p < .0001$. Finally, in Fig. 1A, where a first mover's opting in was unambiguously prosocial, 59.2% of second movers responded generously. This additional increase in generosity is also significant, $\chi^2(1, 1306) = 9.63$, $p < .01$. In sum, second-mover behavior reflects sensitivity to ambiguous motives. Relative to second movers in Fig. 1A and 1C, second movers in Fig. 1B show an intermediate tendency to reciprocate ambiguously motivated positive behavior.

4.3. Discussion

Experiment 1 documents an intermediate tendency to reciprocate ambiguously motivated positive treatment. It also speaks to influential work that demonstrates that some well-studied games confound interest in reciprocity with distributional concerns like inequality aversion and efficiency-seeking and is therefore skeptical about whether instances of positive reciprocity are prevalent (Charness & Rabin, 2002, 2005). For instance, in a sequential prisoners' dilemma, a second mover who cooperates in response to a first mover's cooperation may not be driven by a preference for helping those who help her. She may instead have a preference for equality; that is, she may be inequality-averse (Bolton et al., 1998; Bolton & Ockenfels, 2000; Charness & Haruvy, 2002). Or she may not want to be "wasteful"and thus strive to maximize the parties' joint return; that is, she may be efficiency-seeking (Engelmann & Strobel, 2004). Such social preferences are surely widespread, but Experiment 1 is not straightforwardly susceptible to them. In the games of Fig. 1, second movers select between unequal distributions that hold constant the sum of player's payoffs. Equal treatment is not possible, nor is waste.

4.3.1. Reciprocating good treatment versus rewarding good people

Levine (1998) proposed a theory that applies to many of the same settings as our model. But whereas we are concerned with reciprocity and a preference for repaying good treatment, Levine focuses on

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

altruism and a preference for rewarding altruistic people. Because of this subtle yet fundamental difference between the two accounts, Levine's theory cannot account for the results of our experiments. To explicate, we next briefly revisit our model, then fully introduce and detail Levine's theory, and finally apply each account to Experiment 1.

The desire to reciprocate good treatment that is at the heart of our model concerns behavior within a specific game. Players consider whether their counterpart is being nice to them. Given a counterpart's nice strategy, they may be inclined to respond with a nice strategy. This may be viewed as a preference for closing a circle of good deeds. As an example, in our initial scenario, Michelle might wish to be nice to Elliott because he has been nice to her.

Levine's (1998) theory is not about closing a circle. Instead of focusing exclusively on behavior within a specific game, it concerns proclivities across games. Players consider whether their counterpart is *in general* nice, not whether their counterpart is being nice to them; whether their counterpart is being nice to them is merely a signal of that individual's general tendencies. Players are inclined to treat nicely those people who in general tend to be nice. To illustrate, Michelle might wish to be kind to Elliott because she likes treating good people well, and the positive behavior he has exhibited convinces her that he is indeed a good person. Rather than closing a circle, this is picking one's spots for one's good deeds. In sum, whereas our model is about reciprocating good treatment, Levine's theory is about rewarding good people.

In Levine's theory, the socioemotional payoffs from rewarding good people depend on both a player's own "type" and their beliefs about counterparts' types. First, each player's type lies somewhere between altruistic and spiteful. As a player is increasingly altruistic (spiteful), they increasingly prefer strategies that yield counterparts more (less) material utility. Simply put, altruists enjoy treating people nicely, whereas spiteful individuals enjoy treating people negatively. Second, players make Bayesian inferences about others' types based on observed behavior and, irrespective of their own type, prefer strategies that yield more altruistic (spiteful) counterparts with greater (lesser) material utility. Simply put, both altruists and spiteful individuals enjoy treating nice people relatively nicely and mean people relatively meanly.[5]

Levine imposes reasonable limits on players' altruism and their beliefs about others' altruism. He restricts types to allow for people who enjoy benefiting others to some extent but not for saints who simply prioritize others' material welfare over their own. To operationalize this restriction, Levine requires that absent downstream strategic consequences, every player always prefers to keep a unit of material utility over giving it away for socioemotional payoff (within the mathematical formulation of the theory, this requirement follows from types that lie in the interval from -1 to $+1$).

In Experiment 1, Levine's theory naturally yields second-mover beliefs that parallel the logic of our model. Let Fig. 1C be the baseline. In Fig. 1B, the first mover's foregoing of a middling unilateral option signals his potential altruism, and the second mover may accordingly update her beliefs about his type. In Fig. 1A, his foregoing a highly valuable unilateral option is an even stronger signal that should lead her to a more pronounced updating.

Such belief updating is perfectly in line with our results. Crucially, however, within Levine's model, it does not yield an increasing inclination for second-mover generosity across Fig. 1A, 1B, and 1C. Indeed, Levine's theory does not accommodate any degree of generosity in any of those games. Because of its no-saints restriction on types and because there are no strategic consequences that hinge on the second mover's decision between generosity and miserliness, the theory holds that second movers must always prefer to keep 50 rather than 20 for themselves.

To appreciate the logic of Levine's theory, it is worth applying the theory to Experiment 1 a bit more step-by-step. Once the first mover has acted and the second mover is deciding how to respond, the

[5]In a given game, this inclination may or may not be observable. For instance, if a player is sufficiently spiteful, even the relatively nice treatment they give to a counterpart who they believe to be a nice person can be quite mean. Conversely, if a player is sufficiently altruistic, even the relatively mean treatment that they will give to a counterpart who they believe to be mean can be quite nice.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

second mover's beliefs about the first mover's type are fixed. The valence of the socioemotional payoffs of any two strategies available to her would then differ only if one strategy granted the first mover positive material utility and the other consigned him to negative material utility. But in Experiment 1, miserliness and generosity both provide the first mover with positive material utility and thus generate socioemotional payoffs of positive valence (of different magnitudes). The overall utility of each strategy is a weighted sum of positive material utility from money the second mover assigns to herself (more from miserliness, less from generosity) and positive socioemotional utility from money she assigns to the first mover (less from miserliness, more from generosity). As a result, the second mover can choose generosity only if she prioritizes the first mover's material welfare over her own, which is ruled out.

Our model functions very differently precisely because it concerns a desire to reciprocate positive behavior, not to reward good people. Under our model, the second mover garners positive socioemotional utility from responding generously to the first mover's nice behavior and negative socioemotional utility from responding with miserliness to his nice behavior. Critically, the gap between these utilities can be greater than the material utility loss she sustains from generously parting with money (even for a player with $\lambda < 1$). In other words, what drives generosity is not that the second mover wishes to give sufficiently altruistic first movers money. It is that she feels good if she repays nice treatment and feels bad if she doesn't, and the gap between these feelings can prod her to give. Lastly, our model allows for increasing generosity across Fig. 1A, 1B, and 1C because both the good and bad feelings, and hence the gap, are more pronounced absent ambiguity about the first mover's motivations in Fig. 1A than given ambiguity about them in Sig. 1B.

5. Experiment 2: Who makes the first move?

In seminal sociological work, Gouldner (1960) notes that decisions about whether to make the first move are intimately intertwined with issues of reciprocity. He considers two people who each hold an item prized by the other and could conduct an exchange:

> 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.

Against the backdrop of Gouldner's observations, we now experimentally contrast a standard, sequential prisoners' dilemma with an "endogenous sequencing" prisoners' dilemma. In the standard game, the first mover can "defect" by claiming \$2 for him or "cooperate" by generating \$4 for the second mover; then, the second mover can similarly claim \$2 for herself or generate \$4 for the first mover. In the endogenous sequencing game, each player can also "defect" by claiming \$2 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.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

If neither player moves before the countdown expires, they are placed in a simultaneous prisoners' dilemma.

5.1. Less ambiguous first-move cooperation under endogenous sequencing

The ex-ante symmetry across players in our endogenous-sequencing game is reminiscent of the real-time trust game (Murphy et al., 2006; Rapoport & Murphy, 2012). But in a departure from that setting, first moves in our game can be cooperative. Crucially, our model suggests that the kindness of a first-move cooperator will seem less ambiguous under endogenous sequencing than in the standard game and, more importantly, that this effect will be magnified the more rapidly the person cooperates.

To introduce the relevant intuition, suppose that in each game the players expect that both first-move defection and first-move cooperation will be reciprocated. The model then succinctly characterizes the ambiguity of standard game, first-move cooperation. This strategy yields each player \$4, while the first mover's only pure strategy alternative, defection, yields each player just \$2. First-move cooperation thereby "helps" the second mover but also "helps" the first mover. It could potentially reflect kindness or self-interest or some combination thereof. It is perceived as ambiguous and receives a motivation score of $+ 1 - \theta_i$.[6]

Next, turn to the endogenous sequencing game. In this setting, first-move cooperation can be perceived as less ambiguously motivated. The potential change in perceptions arises because there are now additional alternatives to first-move cooperation, beyond first-move defection. In particular, a player could pursue the maximal payoff by waiting, hoping his counterpart first move cooperates, and if she does, then defect on her. This path of play would yield him \$6, while she ends up with nothing. In comparison to this path of play, engaging in first-move cooperation, which is expected to yield each player \$4, "helps" the second mover while "hurting" him. Note that the comparison is not ambiguous: By first-move cooperating rather than pursuing the maximal payoff, a player eschews his own self-interest and benefits his counterpart. His motivation score is thus $+ 1$. In essence, while a standard-game first-move cooperator does not bear a cost for helping the counterpart, an endogenous sequencing first-move cooperator does. Bearing this cost lends credence to the possibility that the person is genuinely other-regarding rather than tactically self-interested.

Finally, consider the speed with which a player 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 some ambiguity about the player's motives – maybe he would have defected if his counterpart had already cooperated.

The foregoing analysis suggests two empirical predictions. First, many endogenous sequencing 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 renders their kindness credible and thus best allows them to pursue positive reciprocity. Second, rapid first-move cooperation will be reciprocated at a higher rate than first-move cooperation in the standard game.

5.1.1. A reciprocating equilibrium with fast first-move cooperation

Table 2 summarizes our empirical predictions, which we formally derive in an online supplement available at https://osf.io/7k5n9/. We do so by analyzing a set of endogenous sequencing strategy profiles that capture the aforementioned intuition. We suppose that both players anticipate that the

[6]This calculation of motivation scores resembles the calculation for the game in Fig. 1B. But it better highlights that conditioning players' judgments on an entire anticipated strategy profile is a crucial feature of our 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.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

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 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 $+1$ rather than 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 approaches $+1$. A similar analysis holds for the second mover, whose overall motivation score also approaches $+1$ as the first move comes earlier.

While our model permits other equilibria, the formal analysis suggests the two empirical predictions presented above: Many endogenous sequencing games will feature rapid, first-move cooperation, and such rapid first-move cooperation will be reciprocated at a higher rate than first-move cooperation in the standard game.

5.1.2. Extant game-theoretic analyses

Levine's (1998) theory can in principle accommodate our predictions. In his framework, altruistic types may convincingly signal their other-regard via rapid cooperative first moves, which would then be reciprocated at higher rates than slower cooperative first moves. Because the endogenous-sequencing game's strategy space is relatively large, however, the specific distributions of types and beliefs required by the relevant equilibria may be quite complex.

Other theories do not as readily accommodate our predictions. Rabin (1993) and Dufwenberg and Kirchsteiger (2004) judge rapid 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. But in the class of equilibria we have outlined, both theories view all cooperative first moves as equally kind – regardless of their timing. So, they do not make the key prediction that quicker first-move cooperation is more likely to be reciprocated than slower first-move cooperation.

We also note that our model, as well as Rabin's (1993) and Dufwenberg and Kirchsteiger's (2004), can 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. Finally, Falk and Fischbacher (2006) do not make any such predictions. The distinction between endogenous sequencing and the standard game is not germane to their theory.

5.2. 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 years and 129 of them (54.9%) were female. As in Experiment 1, 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 of the 5 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.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

Table 4 First- and second-mover cooperation and defection in the standard, sequential and endogenous-sequencing games in Experiment 2

	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 (85.7%)	N/A (15.0% // 24.8%)
Endogenous-Sequencing, 20 Seconds	16/20 (80%)	9/13 (69.2%)	4/5 (80%)	0/2 (0%)	6 // 8.2 seconds (30% // 41.1%)
Endogenous-Sequencing, 60 Seconds	21/29 (72.4%)	17/20 (85%)	6/8 (75%)	1/5 (20%)	9 // 10.8 seconds (15.0% // 18.1%)
Endogenous-Sequencing, 120 Seconds	16/20 (80%)	13/17 (76.5%)	4/4 (100%)	N/A	13 // 24 seconds (11% // 19.9%)

Procedure. We conducted sessions of between 6 and 20 participants in the same room and using the same methods as in Experiment 1. Each session was randomly assigned to either the sequential condition or an endogenous-sequencing condition with a countdown length of 20, 60, and 120 seconds (which we varied for robustness). In the sessions featuring the standard, sequential game, participants were then randomly assigned the role of first or second mover. In all 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.

Materials. The experiment was programmed in z-Tree (Fischbacher, 2007). Screenshots of the complete instructions, which did not use the terms defection or cooperation, are available on the OSF at https://osf.io/7k5n9/.

Each player was initially credited with $2. At any moment during the countdown, each player could continue to "wait," choose to "keep" his $2, or choose to "give" his money to his counterpart, with the proviso that if a player chose to give then we as the experimenters would add $2 to the amount conveyed, so that the counterpart would receive a total of $4. Keeping constitutes defection, while giving constitutes cooperation. If either player chose to "keep" or "give," the countdown immediately stopped, and the other player was informed of the move and provided with the opportunity to respond.[7] Finally, if neither player chose to "keep" or "give" before the countdown expired, the players were placed in a simultaneous prisoners' dilemma (and this was common knowledge).

5.3. Results

As Table 4 shows, in the standard game, 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 4 confirms that many games were indeed over very rapidly, in a matter of seconds. Collapsing across the various timer lengths, the median time elapsed before a first move

[7] The z-Tree program made sure there were no "nearly simultaneous" moves, by which a second mover keeps or gives slightly later than an initial mover keeps or gives but does not yet know how the initial player acted. As soon as a player made the first move, the software notified the other player and requested an explicit response to the first mover's keeping or giving.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

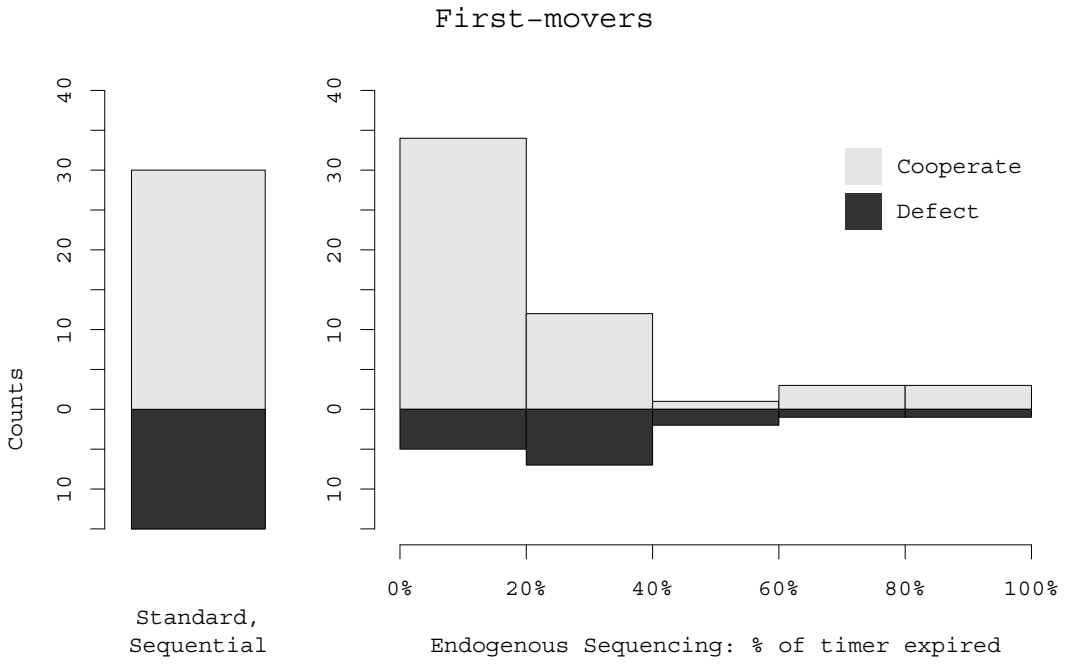


Fig. 2 First movers' behavior in the standard, sequential (left) and endogenous-sequencing (right) games in Experiment 2

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 seconds corresponds to 30% of the countdown. In what follows, we employ 20% of the countdown as a conservative cutoff defining fast versus slow first moves, and we show that our conclusions are qualitatively unchanged as the cutoff is moved earlier.

Figure 2 provides initial evidence that games typically ended quickly because one of the players elected to first-move cooperate: 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 shows, 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%).

Recall the second critical element of our prediction: Quick first-move cooperation under endogenous sequencing will be reciprocated at a higher rate than first-move cooperation in the standard game. Figure 3 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, $p = .03$ by two-sided Fisher's exact test).

Table 5 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.

Importantly, second movers reacting to quick first moves under endogenous sequencing are largely unaffected by selection bias. 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. In contrast, second movers who faced slow first movers chose not to make an early first move and thus form a biased sample.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

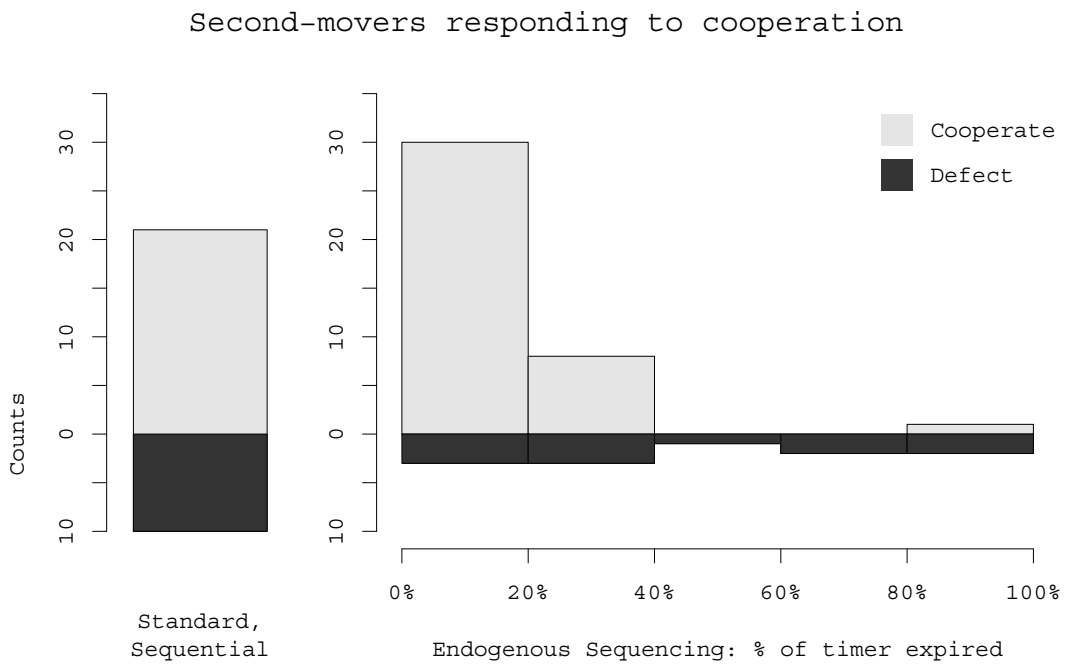


Fig. 3 Second movers' responses to first-mover cooperation in the standard, sequential (left) and endogenous-sequencing (right) games in Experiment 2

Table 5 Positive reciprocity in the standard, sequential game and in games of various speeds under endogenous sequencing in Experiment 2

	1st-Mover *C*	2nd-Mover Reciprocity of*C* with *C*
Standard Sequential	30/45 (66.7%)	21/31 (67.7%)
1st Move in $\leq$ 5% of Countdown	8/11 (72.7%)	9/9 * (100%)
1st Move in $\leq$ 10% of Countdown	22/25 (88%)	21/22 ** (95.5%)
1st Move in $\leq$ 15% of Countdown	29/34 (85.3%)	26/29 * (89.7%)
1st Move in $\leq$ 20% of Countdown	34/39 (87.2%)	30/33 ** (90.9%)
1st Move after $>$ 20% of Countdown	19/30 (63.3%)	9/17 (52.9%)

Note. * means $p < .1$ and ** means $p < .05$, vs. the standard sequential game, Fisher's exact test.

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.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

5.4. Discussion

Experiment 2 indicates that fast first-move cooperation in one-shot endogenous sequencing prisoners' dilemmas sends a powerful signal about a cooperator's good intentions. This finding complements the unraveling of cooperation observed in repeated real-time trust games, in which sequencing is also endogenous, but first moves cannot be cooperative (Murphy et al., 2006; Rapoport & Murphy, 2012). It also speaks to influential work that is skeptical about whether instances of positive reciprocity are prevalent (Charness & Rabin, 2002, 2005). To reiterate, this work was catalyzed by papers that showed that much behavior ascribed to positive reciprocity may reflect other social preferences, like inequality aversion or efficiency-seeking (Bolton et al., 1998; Bolton & Ockenfels, 2000; Charness & Haruvy, 2002; Engelmann & Strobel, 2004). But the pronounced tendency to reciprocate rapid first-move cooperation under endogenous sequencing cannot be explained by such preferences. Under endogenous sequencing, second movers' ability to avoid inequality or strive for efficiency is not impacted by whether the first mover acted relatively rapidly or slowly. Nor is it impacted by whether the game being played is an endogenous sequencing interaction or a standard sequential interaction.

6. General discussion

We have introduced an attributional model of reciprocity in which people judge whether good treatment they receive is motivated by genuine kindness, tactical self-interest, or some combination thereof. The model's key contribution lies in how it handles ambiguity. It allows players to maintain divergent hypotheses: Good treatment can be perceived as partially caring and partially self-interested, and players become increasingly inclined to reciprocate as they increasingly perceive good treatment to reflect caring rather than self-regard. Previous analyses did not handle ambiguity in this realistic manner. They resolved rather than maintained ambiguity. They treated the hypotheses of caring and tactical self-interest as mutually exclusive and required that people commit to one and dismiss the other. In turn, they required that any positive behavior be viewed as either fully meriting reciprocity or entirely unworthy of it.

By highlighting the notion of positive reciprocity given ambiguously motivated treatment, our models yield three benefits. The first concerns the data we collect. Our model accommodates intuitive patterns of play, which previous analyses did not and which we experimentally corroborate. These patterns involve intermediate tendencies to reciprocate ambiguously motivated behavior.

Applying our model to settings in which past research suggested that positive reciprocity does not matter much yields two additional benefits. Because it handles ambiguity more realistically than previous theories, our model is able to uncover positive reciprocity where it had previously been overlooked. Specifically, it highlights that the same action can appear generous (and meriting reciprocity) to one person and miserly (and not meriting reciprocity) to another. Appreciating such diversity in perceptions helps explain well-known results from gift exchange and trust games. In addition, our model challenges the conventional view that unraveling in finitely iterated prisoners' dilemmas and generous offers in ultimatum games are inconsistent with positive reciprocity. It shows that both phenomena readily fit with a preference for repaying ambiguously motivated good treatment.

While our findings suggest that positive reciprocity may be more common than previously thought, we note that the relatively high reciprocal cooperation rates we observe could be particular to our participant samples, namely US online participants and undergraduates. To take just one relevant example, it is known that some patterns of cooperation are more prevalent among Americans than Japanese (Yamagishi et al., 2005, 2007).

6.1. Diverse perceptions of others' motives

In gift exchange and trust games, a first mover decides what portion of an endowment to transfer to the second mover and what portion to keep, and a second mover then does likewise (for reviews, see

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

Gilchrist et al., 2016; Johnson & Mislin, 2011). In these games, in which players choose among several levels of generosity versus miserliness, our model emphasizes that the same action may be viewed as positive, negative, or neutral, depending on a person's perceptions of self-interest and caring. For instance, is transferring a small portion of the endowment generous or miserly? To someone who believes that self-interest predominates, even a minimal transfer can seem generous and constitute positive treatment. On the other hand, to somebody who believes that kindness predominates, minimal transfers can seem miserly and constitute negative treatment.

This diversity in perceptions can lead to diversity in second movers' reactions – which suggests that the relationship between first and second moves in gift exchange and trust games may be noisy. Indeed, correlations between first- and second-mover transfers are typically modest (Johnson & Mislin, 2011; also Cox, 2004) and often not statistically significant (e.g., Pillutla, Malhotra, & Murnighan, 2003). The low correlations and their lack of robustness are frequently interpreted as a lack of positive reciprocity. In contrast, our model suggests that they reflect diversity in perceptions of motives and that they are entirely consistent with positive reciprocity.

Stanca et al. (2009)'s contrast of standard and "decoupled" gift exchange illustrates the point. In these authors' experiment, first movers were initially endowed with 20 experimental tokens and could transfer any amount between 0 and 20 to a second mover. But while first movers in the standard treatment knew that the second mover would subsequently have a chance to reciprocate (or not), first movers in the decoupled treatment acted before being informed that the second mover would have a chance to respond.

Stanca et al. (2009) reasoned that tactical motives could only underlie first-mover behavior in the standard but not decoupled game. Only in the former could first movers anticipate that they could influence second movers. Accordingly, Stanca et al. (2009) predicted greater first-mover generosity in the standard game. This prediction, which is in line with our model, was strongly corroborated.

On the other hand, Stanca et al. (2009) also predicted greater second-mover generosity in the decoupled game because any generosity in that game must be genuine rather than tactical. Yet collapsing across all possible first-mover gifts, the average second-mover return gift was only modestly greater under decoupling (their Tables 2 and 7). This finding appears to beg the question of whether second movers distinguished between different first mover motivations. By our model, however, it simply highlights the importance of diversity in perceptions of motivations.

Consider a decoupled first mover who gives 5 of his 20 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, so that the motivation score is $-1 + \theta$. Compared to less generous gifts, the first mover has helped the second mover and hurt himself, so that the motivation score is $+1$. Averaging over the 15 more generous and 5 less generous gifts yields a total motivation score of $[15(-1 + \theta) + 5(+1)]/20 = -\frac{1}{2} + \frac{3}{4}\theta$, which can be negative or positive for different values of θ. Whether the first mover has been kind or unkind is thus open to interpretation.

The same first-mover gift may then elicit disparate return gifts, depending on whether a second-mover discounts a lot or a little. The relationship between initial and return gifts will in turn be noisy in both the standard and decoupled games, rendering it difficult to isolate the impact of specific motivations. In this way, rather than revealing a limited impact of distinct first-mover motives on second-mover reciprocity decisions, Stanca et al.'s (2009) gift exchange data reveal how diverse perceptions attenuate the correlation between initial and return gifts.

6.2. Positive reciprocity in repeated prisoners' dilemmas

In finitely iterated prisoners' dilemmas, cooperation rates are initially high but then "unravel," so that defection becomes increasingly common as the final period approaches (see, e.g., Andreoni & Miller, 1993). Many authors suggest that unraveling is inconsistent with a preference for reciprocating good treatment. They argue that given this preference, a history of mutual cooperation should

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

engender subsequent mutual cooperation (Selten & Stoecker, 1986; see also Axelrod & Hamilton, 1981; Kreps et al., 1982). In contrast, our model implies that unraveling is entirely consistent with a preference for repaying good treatment – once motivational ambiguity is accounted for. Players will note that in early and middle periods, their counterpart's cooperation may have been genuine. Or it could have been tactical, designed to facilitate a stretch of materially profitable, reciprocal cooperation (Axelrod & Hamilton, 1981; Kreps et al., 1982). But then even long histories of mutual cooperation may be attributed largely to self-interest rather than caring and deemed not to merit late-period reciprocity.

6.3. Positive reciprocity in ultimatum games

By more realistically handling ambiguity, our model also points to the role of positive reciprocity in ultimatum games. In the canonical instantiation of this game, two players divide a pot of money. The "proposer" offers a specific split, which the "responder" can accept or reject. If she accepts the split, it is implemented. If she rejects it, neither player receives anything. A vast experimental literature reveals two stylized facts: many proposers are generous, offering an even or nearly even split, and many responders reject miserly offers of less than about a quarter of the pot (Camerer & Thaler, 1995; Güth & Tietz, 1990; Oosterbeek et al., 2004; Tisserand, 2014).

To explain these stylized facts, researchers have focused on "threatening" equilibria and negative reciprocity. In such equilibria, the proposer is generous because he anticipates that the receiver will reject a miserly split. That is, he is compelled to act generously because of his counterpart's power to reject miserly offers. The stick of negative reciprocity is key (e.g., Falk et al., 2003; Levine, 1998).

Our model also offers this type of explanation. But unlike existing theories, it offers an additional, complementary explanation in terms of 'nice' equilibria and positive reciprocity.[8] In a nice equilibrium, the proposer's generosity is not a reaction to threat. It reflects other-regard. The proposer acts nicely in pursuit of the good feelings that ensue when the responder accepts an equitable outcome and in avoidance of bad feelings that would ensue were he to consign the responder to a meager outcome. The carrot of positive reciprocity plays an important role.[9]

We suspect that many investigations of the ultimatum game betray conceptualizations of it akin to threatening rather than nice dynamics. Hoffman et al. (1994), for example, considered altruism as a source of generous offers and argued against it (see also Harrison & McCabe, 1992; Marlowe, 2004, p. 186; Ruffle, 1998). 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

[8] A generous offer that is accepted cannot generate feelings of positive reciprocity in existing accounts. Under Rabin's (1993) theory, a responder who accepts an offer cannot be deemed kind because this action materially benefits her. Granted, accepting an offer helps the proposer, but the responder might be motivated by self-interest. Under Dufwenberg and Kirchsteiger's theory (2004), an accepting responder cannot be deemed kind for essentially the same reason (accepting does not open up the possibility of the responder being materially hurt). In Falk and Fischbacher's (2006) theory, a proposer who offers an even split or anything less is not kind because he is not granting the proposer a greater payoff than he himself receives.

[9] By our model, good feelings of positive reciprocity emerge when a generous offer is accepted, and bad feelings can emerge when a meager offer is accepted. Consider the canonical version of the game that allows for any integer split of a payoff of 10 and assume the respondent will accept any offer (so that there is no threat). Because acceptance of any offer helps both players, the respondent's motivation score is 1-θ. Meanwhile, the proposer's total motivation score for an offer of n is equal to $[(10 - n)(-1 + \theta) + n(+1)]/10$. For $n = 5$ this motivation score is positive. Hence, the product of the players' motivation scores is positive, meaning that good feelings are generated. For $n < 5$ and sufficiently small θ, the proposer's motivation score is negative. In other words, when players do not discount too much, they take smaller offers more or less at face value and view them negatively. This leads to the product of motivation scores being negative, so that bad feelings emerge. However, diversity of perceptions matters. For $n < 5$ and sufficiently large θ, the proposer's total motivation score is positive. That is, when players discount a lot, they are keenly aware of self-interest as a driver of behavior and view even meager offers positively, which allows good feelings to emerge despite the proposer's miserliness. In sum, our model highlights the special status of even splits; they always generate good feelings amongst the players, and it shows that meager offers, while susceptible to diverse perceptions, will often generate bad feelings among the players.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

should influence 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 fear that meager offers will be rejected, that is, fear of the negative reciprocity underlying threatening dynamics. But nice dynamics that implicate positive reciprocity fit equally well with data presented by Hoffman et al. (1994). As we emphasized with the game of Fig. 1C, because reciprocity is contingent on a counterpart's behavior, it is relevant to proposers but not dictators: Proposers might make generous offers because it feels good when the responder accepts, but this motivation is not relevant to dictators, who cannot tap such feelings.

Levine (1998) also explains even-split or nearly even-split ultimatum offers in terms of threatening dynamics. Seminal data, however, are at odds with his explanation and instead in line with the nice dynamics our model uncovers. Levine emphasizes that proposers (accurately) believe that a nonnegligible fraction of receivers are spiteful and will thus reject meager offers. The possibility of spiteful rejection is the threat that pushes proposers to be even-handed. Furthermore, by Levine's theory, altruistic players are relatively likely both to make generous offers as proposers and to accept meager offers as responders, whereas spiteful players are relatively likely to both make miserly offers and reject meager offers. In studies in which each participant plays both roles, Levine's model thus predicts a negative correlation between the offer a participant proposes and the minimum offer they are willing to accept. But in the original study of the ultimatum game by Güth et al. (1982), the offers that participants propose are positively correlated with their rejection thresholds, $r = 0.30$ ($p = .07$; their Table 7). Blanco et al. (2011) later replicated this finding, reporting $r = 0.40$ ($p < .01$; their Table 3).

Such positive correlations are consistent with our model's nice dynamics. By our account, people who place greater weight on socioemotional payoffs are more likely to both seek out the good feelings that emerge when a generous offer is accepted and avoid the bad feelings that emerge when a meager offer is accepted. To do so, they must be generous as proposers and demanding as receivers.

To be clear, we do not claim that threatening dynamics and negative reciprocity are rare in the ultimatum game. They are surely common. We merely claim that nice dynamics involving positive reciprocity may be common too. Indeed, it is notable that several other positive social motivations have been cited for generous offers, including altruism (Camerer & Thaler, 1995), fairness (Nowak et al., 2000), and image maintenance (Dana et al., 2006), while positive reciprocity has not. In our view, the possibility that good feelings are generated when proposer generosity is met with a positive response is just as plausible as the possibility that, say, image maintenance guides proposer generosity. Yet this possibility appears to have been overlooked, probably because existing work does not allow for ambiguous attributions.

Replication Packages. The replication material for the study is available at https://osf.io/7k5n9/.

References

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

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

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, *72*(2), 321–338.

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.

Cabral, L., Ozbay, E. Y., & Schotter, A. (2014). Intrinsic and instrumental reciprocity: An experimental study. *Games and Economic Behavior*, *87*, 100–121.

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

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

Charness, G., & Haruvy, E. (2002). Altruism, equity, and reciprocity in a gift-exchange experiment: An encompassing approach. *Games and Economic Behavior*, *40*(2), 203–231.
Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.
Charness, G., & Rabin, M. (2005). Expressed preferences and behavior in experimental games. *Games and Economic Behavior*, *53*(2), 151–169.
Cox, J. C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior*, *46*(2), 260–281.
Cox, J. C., Sadiraj, K., & Sadiraj, V. (2008). Implications of trust, fear, and reciprocity for modeling economic behavior. *Experimental Economics*, *11*(1), 1–24.
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.
Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298.
Einhorn, H., & Hogarth, R. (1986). Judging probable cause. *Psychological Bulletin*, 99(1), 3–19.
Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, *94*(4), 857–869.
Falk, A., Fehr, E., & Fischbacher, U. (2003). On the nature of fair behavior. *Economic Inquiry*, *41*(1), 20–26.
Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness – Intentions matter. *Games and Economic Behavior*, *62*(1), 287–303.
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*(1), 43–64.
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*. Addison-Wesley.
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.
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.
Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review*, *25*(2), 161–178.
Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, *35*(1), 1–15.
Güth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, *3*(4), 367–388.
Güth, W., & Tietz, R. (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology*, *11*(3), 417–449.
Harrison, G. W., & McCabe, K. (1992). Testing noncooperative bargaining theory in experiments. *Research in Experimental Economics*, *5*, 137–169.
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.
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.
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.
Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*(2), 107.
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.
Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, *1*(3), 593–622.
Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z
Malmendier, U., te Velde, V. L., & Weber, R. A. (2014). Rethinking reciprocity. *Annual Review of Economics*, *6*(1), 849–874.

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

Marlowe, F. W. (2004). Dictators and ultimatums in an egalitarian society of hunter-gatherers: The Hadza of Tanzania. In J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, & H. Gintis, (Eds.), *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies* (168–193). Oxford University Press.

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.

Murphy, R. O., Rapoport, A., & Parco, J. E. (2006). The breakdown of cooperation in iterative real-time trust dilemmas. *Experimental Economics*, *9*(2), 147–166.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Prentice Hall.

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

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.

Orhun, A. Y. (2018). Perceived motives and reciprocity. *Games and Economic Behavior*, *109*, 436–451.

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.

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

Rapoport, A. (1987). Research paradigms and expected utility models for the provision of step-level public goods. *Psychological Review*, *94*(1), 74–83.

Rapoport, A., & Murphy, R. (2012). Evolution and breakdown of trust in continuous time. In R. Croson, & G. E. Bolton (Eds.), *The Oxford* h*andbook of economic conflict resolution* (199–215). Oxford University Press.

Ruffle, B. J. (1998). More is better, but fair is fair: Tipping in dictator and ultimatum games. *Games and Economic Behavior*, *23*(2), 247–265.

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.

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

Woods, D., & Servátka, M. (2016). Nice to you, nicer to me: Does self-serving generosity diminish the reciprocal response? *Experimental Economics*, *19*(1), 1–24.

Yamagishi, T., Kanazawa, S., Mashima, R., & Terai, S. (2005). Separating trust from cooperation in a dynamic relationship: Prisoner's dilemma with variable dependence. *Rationality and Society*, *17*(3), 275–308.

Yamagishi, T., Terai, S., Kiyonari, T., Mifune, N., & Kanazawa, S. (2007). The social exchange heuristic: Managing errors in social exchange. *Rationality and Society*, *19*(3), 259–291.

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 $\Delta(X_i)$.

Segal and Sobel (2007) study players who, conditional on an anticipated strategy profile $\sigma^\star = (\sigma_i^\star, \sigma_j^\star)$, have preferences $\succeq_{i,\sigma^*}$ defined over their own strategies Σ_i. Segal and Sobel axiomatically characterize the conditions for representing such preferences as linear combinations of a player's and his counterparts' expected utilities (for material outcomes). We situate our model within this framework.

Let $u_i(\sigma_i, \sigma_j^\star)$ denote player i's expected utility for his material payoff, and let $u_j(\sigma_i, \sigma_j^\star)$ denote his counterpart's expected utility for *her* material payoff. Let $\lambda_i \geqslant 0$ index player i's other-regard, and let M_{i,σ^*} and M_{ij,σ^*} denote player i's assessments of his and his counterpart's motivations, conditional on the anticipated strategy profile $\sigma^\star$(defined below). 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 $\sigma^\star$:

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

The 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

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press

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

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

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

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

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 $(\sigma_i, \sigma_j{}^*)$ 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 all 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

$$m_{ij,\sigma^*}(s_i, s_j) = \frac{\sum\limits_{s'_j \in \boldsymbol{A_{j,\sigma^*}}(s_i, s_j)} b_{ij,\sigma^*}(s'_j)}{\left|\boldsymbol{A_{j,\sigma^*}}(s_i, s_j)\right|}.$$

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 $\sigma_i{}^*$ and $\sigma_j{}^*$ 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.

Cite this article: Müller-Trede, J., & Rottenstreich, Y. (2025). Positive reciprocity when motives are ambiguous. *Experimental Economics*, 28(3), 758–781. https://doi.org/10.1017/eec.2025.10009

https://doi.org/10.1017/eec.2025.10009 Published online by Cambridge University Press