Perceived Motives and Reciprocity*

A. Yeşim Orhun

Ross School of Business, University of Michigan

March 30, 2016

Abstract

Both genuine kindness and self-interested material gain may motivate actions that benefit others in situations where a reciprocal reaction is plausible. Perceived motives have been suggested to influence the kindness judgments and positive reciprocity towards such beneficial actions. We present results from two experiments that distinguish the role of perceived motives in reciprocal decision-making from the role of outcomes or perceived intentions. Our results suggest that the degree of positive reciprocity triggered by the same helpful action is much lower when the strategic incentives the first-mover has for taking that action are stronger. Moreover, the decline in the degree of positive reciprocity is associated with the deterioration of the degree of altruism inferred regarding the helpful first-movers. We show that the results cannot be captured by intention-based reciprocity theories, and discuss theoretical implications.

Keywords: Motives, Beliefs, Reciprocity, Intentions, Social Preferences.

*We thank Gary Bolton, Jonathan Carmel, Bogachan Celen, Yan Chen, James Cox, Emel Filiz-Ozbay, Aradhna Krishna, Steve Leider, Yusufcan Masatlioglu, Neslihan Uler, Axel Ockenfels, Erkut Ozbay, Tanya Rosenblat, Andrew Schotter, Katharina Schüssler, Severine Toussaert, Peter Werner, and seminar participants at Erasmus University Rotterdam, George Mason University (ICES), New York University (CESS), University of Cologne, University of Michigan and University of Texas at Dallas (CLBOE) for their comments and suggestions. Lillian Chen, Michael Payne, Arun Varghese, Roshni Kalbavi, Hannah Lee, Valerie Laird and Catherine Dolan provided excellent support in conducting experimental sessions.

1 Introduction

All reciprocal relations feature rewards for helpful behavior and/or punishment for harmful behavior. Therefore, helpful actions in these relations may be driven by either an intrinsic motive of benefiting others, or a strategic motive to secure future gains and/or avoid future losses.[1] Importantly, actions that are intended to benefit others can be driven entirely by strategic motives. For example, a manufacturer cleaning up the environmental waste generated by its factory may be driven mainly by fear of being sued by the community rather than a genuine concern for its neighbors. Alternatively, a retailer running a campaign to benefit poor children may be mainly driven by curtailing its losses to competitors who are more socially responsible. Are these helpful actions considered kind? This question is central to reciprocal decision-making. Rabin (1998, p.22) observes: "a crucial feature of the psychology of reciprocity is that people determine their dispositions toward others according to motives attributed to these others... If you think somebody has been generous to you solely to get a bigger favor from you in the future, then you do not view his generosity to be as pure as if he had expected no reciprocity from you." Rabin's insight would suggest that the same helpful action would trigger more positive reciprocity when it is perceived to have been motivated by instrinsic rather than strategic motives. Interestingly, the impact of perceived motives on reciprocal decision-making has been largely unexplored, possibly due to challenges in distinguishing it from the role of perceived intentions.

In this paper, we provide novel experimental evidence regarding the impact of perceived motives on reciprocal decision-making as distinct from the impact of outcomes or perceived intentions. Before we detail our experimental design, let us distinguish the potential role of perceived motives from the role of percieved intentions. Individuals may evaluate both the motives and the intentions behind an action in their assessment of its kindness and in determining the appropriate response to it (Heider, 1958; Kelley, 1973; Ross and Fletcher 1985). The importance and distinction of these constructs is immediately apparent in criminal law. Consider the trial of an alleged murderer. The court must investigate the defendant's intent behind the action that resulted in murder: Did the defendant shoot the victim by mistake, or on purpose? The court must also establish the defendant's motive: Was the action in self-defense, or fueled by revenge? More generally, the intention of a person is often judged by the consequence the actor expected his action to have on others. On the

[1]We refer the interested reader to Cabral, Ozbay and Schotter (2014), Dreber, Fudenberg and Rand (2014), Gneezy, Güth and Verboven (2000), Reuben and Suetens (2012), Segal and Sobel, (2007, 2008) and Sobel (2005) for documentation of the relative role of intrinsic and strategic motives in sequential games with a reciprocal nature.

other hand, the motive of a person relates to the reason why he wanted to achieve this consequence.

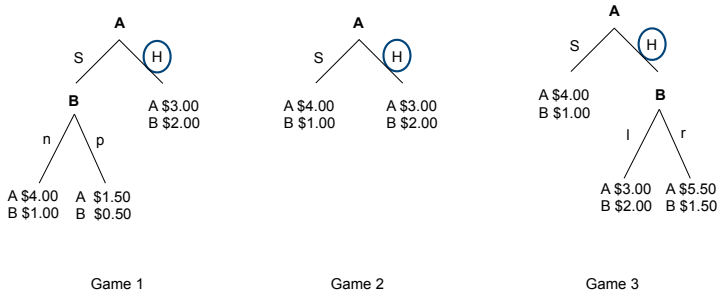
Befittingly, intention-based reciprocity theories proposed by Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) extend outcome-based reciprocity models (e.g. Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) by conceptualizing the kindness of the first-mover's intentions based on beliefs regarding the consequence the first-mover expected his action to have on the second-mover's payoffs. An action is considered kind if the second-movers' expected payoff is larger than a fair benchmark, where this benchmark is defined in the context of possible payoffs achieveable in the interaction.[2] Similarly, the entire body of experimental research investigating the role of perceived intentions manipulates perceptions regarding the consequence the first-mover expected his action to produce for the second-mover. Some experiments compare the reciprocity towards the same action based on whether the action was intentionally chosen by the first-mover or chosen by an external process, such as a computer or a third-person.[3] Other experiments manipulate whether a different choice by the first-mover could have produced a better outcome for the second-mover.[4] Overall, the experimental evidence on the role of intentions show that reciprocity is sensitive to the perceived volition of the first-mover and to what is considered a fair benchmark for the second-mover's payoffs. Clearly, a consideration of the second-mover's payoffs is insufficient to glean the motive of the first-mover without considering what the first-mover could have gained or lost if he acted differently. However, these results are silent on whether perceptions regarding the degree to which the first-mover's choice was motivated by strategic considerations versus a genuine care for the second-mover matters for reciprocity.

In some situations, the impact of perceived intentions and perceived motives may go hand in hand. Yet in others, kind intentions can be driven entirely by self-interest, highlighting the need to consider perceived motives in reciprocal decision-making.

[2]The benchmark in the Dufwenberg and Kirchsteiger (2004) model depends on expectations regarding what the second-mover could have obtained had the first-mover behaved differently. The benchmark in the Falk and Fischbacher (2006) model is the expected payoffs of the first-mover. In addition, Falk and Fischbacher (2006) propose the degree of agency the first-mover had to amplify kindness perceptions. Rabin (1993) introduced the concept of intention-based reciprocity and defined kindness based on whether the first-mover expects his action to actualize an equitable payoff for the second-mover. Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006), among others, extended this idea to sequential reciprocal interactions, which our investigation focuses on.

[3]This literature shows that the same helpful action triggers more positive reciprocity if the first-mover had a higher degree of control over the outcome (Offerman, 2002; Charness and Haruvy, 2002; Charness, 2004; Charness and Levine, 2007; Falk, Fehr and Fischbacher, 2008; Klempt, 2012),

[4]Experimental evidence suggests that the same helpful action triggers more positive reciprocity if the first-mover chose the option that favored the second-mover the most among a set of reasonable choice alternatives (Brandts and Sola, 2001; Nelson, 2002; McCabe, Rigdon and Smith, 2003).



To illustrate when the impact of perceived intentions and perceived motives on kindness inferences may differ and when they may be aligned in a simple manner, let us evaluate the kindness of choosing (H) across the three games illustrated above from both a perceived motives and from a perceived intentions perspective. The proposed impact of perceived motives is strightforward: if the second-mover believes that some first-movers are motivated to choose (H) merely due to a fear of punishment in Game 1, and merely due to a hope of rewards in Game 3, then the second-mover would perceive the choice of (H) to be more likely motivated by genuinely altruistic motives in Game 2 than in Game 1 or Game 3. Choosing (H) in Game 2 is also associated with kinder intentions than choosing (H) in Game 3, because expecting (costly) positive reciprocation is synonymous with lower expected payoffs for the second-mover. Therefore, differences in perceived kindness of choosing (H) between Game 2 and Game 3 can be driven by the differences in perceived intentions or motives, rendering the distinct impact of perceived motives, even if present, impossible to identify from the impact of perceived intentions. Consequently, evidence regarding the distinct role of perceived motives has been elusive in such contexts (e.g. Stanca, Bruni, and Corazzini, 2009). Importantly, however, a comparison of the perceived kindness of choosing (H) across Game 1 and Game 2 allows us to distinguish the impact of perceived motives. An account based on perceived intentions would not predict (H) to be associated with kinder intentions in Game 2 compared to Game 1, because the choice of (H) does not lead to lower expected payoffs for the second-mover in Game 1 even if some negative reciprocation is expected when the first-mover chooses (S).[5] The comparison between Game 1 and Game 2 serves as a simple example for a much wider array of economic interactions where actions driven by self-interest need not be associated with unkind intentions.

[5]In particular, the reciprocity model proposed by Falk and Fischbacher (2006) would not predict a difference in the perceived kindness of choosing (H) across Game 1 and 2, because the payoff the first-mover intended (H) to produce for the second-mover ($2) compared to what he intended it to produce for himself ($3) is kept constant. On the other hand, the Dufwenberg and Kirchsteiger (2004) model may predict the perceived kindness of (H) to be higher in Game 1 if the second-mover believes that the first-mover expected negative reciprocation if he chose (S), because while the payoff the first-mover expected (H) to produce is the same across the two games, the second-mover could have attained a lower outcome if the first-mover chose (S) in Game 1 compared to in Game 2.

The pertinent question, then, is whether motives matter above and beyond the consequence of an action and the intention behind it. A thorough understanding of reciprocal decision-making is incomplete without a clear understanding of the role of perceived motives, because all reciprocal interactions, albeit to differing degrees, present strategic incentives for desired behavior. This paper explores whether positive reciprocity towards the same helpful action is lower when the action is more likely to be motivated by strategic considerations, even though it produces the same consequences and is associated with kind intentions. Conceptually, our design leverages the insights gained by comparing Game 1 and Game 2. The experimental design gives the first-mover a choice between a selfish and a helpful action. It keeps the second-mover's strategy space in the sub-game that follows a helpful action constant, and exogenously shifts beliefs regarding the likelihood of strategic motives by varying the existence of a punishment option in the sub-game that follows a selfish action.[6] Elicited beliefs demonstrate that manipulating the access of the second-mover to a punishment option appropriately shifts first- and second-order beliefs regarding behavior and inferences regarding the altruism of a helpful first-mover. This design allows us to manipulate beliefs regarding motives without relying on any surprises about the true nature of the interaction and without generating confounding movements in perceived intentions. We present results from two experiments that feature variations on this common structure. The first experiment presents novel evidence for the distinct role of perceived motives based on a between-subjects design. The second experiment builds on this evidence by linking within-person changes in reciprocal reactions to the same helpful action to within-person changes in the inferences regarding the inherent altruism of a helpful first-mover. We also present robustness studies to show that our results do not suffer from menu effects or contamination across different elicitation tasks.

Overall, our results corroborate Rabin's early foresight. Reciprocity towards a helpful action is not just a function of the beneficiary's perception as to whether the benefactor intended to make a sacrifice, but also hinges crucially on whether she believes that the benefactor made the sacrifice due to strategic motives or out of genuine care for others. Therefore, strong incentives may damage kindness perceptions and hurt reciprocal relations even though they are effective in generating beneficial behavior. This insight suggests important future avenues of inquiry, such as how to design contracts that trade-off the impact of incentives on behavior and attributions, or how a

[6]Costa-Gomes, Huck and Weizsäcker (2014) also rely on random disturbances to the second-movers' responses to create exogenous variation in first-movers' beliefs regarding their payoffs. Our approach differs in two ways. First, we do not replace the second-mover's choice by a random move, but rather change his action space outside of the sub-game of interest. Second, our objective is to manipulate not only the first-mover's expectations of what he may earn in the event he chooses (S), but also the second-movers beliefs regarding these expectations.

profit maximizing firm should choose and communicate its socially responsible activities. The results also may offer a new perspective to existing empirical regularities. How can we reconcile the impact of perceived motives with the evidence on guilt aversion that shows that people reciprocate more towards those who have high expectations of reciprocation? Can perceived motives be an important explanatory factor for the positive reciprocity puzzle, which demostrates that propensity to punish harmful behavior is stronger than the propensity to reward friendly behavior? We comment on the implication of our results for several future research avenues, including these issues, in the Discussion section.

Analysis of theoretical models that account for the role of perceived motives would greatly help in understanding how a consideration of motives may change the role of reciprocal decision-making across a wide array of economic decisions. As we elaborate in the Discussion section, models formalizing reciprocity as a response to the altruism of the benefactor as revealed by his actions (Levine 1998, Cox, Friedmand and Sadiraj, 2008; Gül and Pesendorfer, 2010) have the highest explanatory power in the context of our experiment. The notion that people respond to the benefactor's type, rather than action, has also recently been used to distinguish between genuine and strategic kindness in the context of a gift-exchange game (Arbak and Kranich, 2005; Dur, 2009; Non, 2012). We hope that the experimental design and results presented in this article are useful for future theoretical developments that define and generalize the role of perceived motives in reciprocal decision-making.

2 Experimental Investigation

2.1 Common Design Elements Across Experiment 1 and Experiment 2 in the Context of the Previous Experimental Literature

Identifying the role of motives is not without its challenges. The experiment needs to jointly and exogenously vary the motives of the first-mover and the second-mover's perceptions about these motives, without generating confounding variations in perceived intentions. We present two experiments that manipulate perceived motives of the first-mover by varying the second-mover's access to different response options outside of the sub-game of interest that follows a helpful decision by the first-mover. In particular, the first-mover decides between an option that maximizes his payoffs (S), and a helpful action that transfers some of his payoffs to the second-mover (H). We manipulate perceived motives by changing the incentives of the first-mover to help: treatment 1

provides the option to punish the first-mover, and treatment 2 does not. Importantly, the action space of the second-mover in the sub-game following a choice of (H) by the first-mover is also kept constant across treatments: she chooses between keeping the money given to her by the first-mover (N), or reciprocating by sending a specific amount that gets multiplied and added to the first-mover's account (R). The first experiment features a between-subject design that compares positive reciprocity of second-movers in response to a helpful action when the first-mover could have been motivated to help by fear of punishment, with their positive reciprocity in response to the same helpful action when the punishment-avoidance motive was absent. The second experiment compares secondmovers' demand for rewards across three within-subject treatments where first-movers i) have no strategic incentives to help, or ii) may be motivated to help by the hope of rewards, or iii) may be motivated to help by the fear of punishment.

Previous exprimental literature shows that the same helpful action triggers more positive reciprocity when the first-mover has a higher degree of control or agency over the beneficial outcome (Blount, 1995; Offerman, 2002; Charness and Haruvy, 2002; Charness, 2004; Charness and Levine, 2007; Falk et al., 2008; Klempt, 2012), and if the first-mover chose the best he could among his choice alternatives (Brandts and Sola, 2001; Nelson, 2002; McCabe et al., 2003). In addition, recent work (Rand, Fudenberg and Dreber, 2013; Toussaert, 2014) vary the degree to which the first-mover's intended choices are observed in situations where first-movers' choices are implemented with noise. These articles show that the second-mover is sensitive to the degree of ambiguity regarding whether the first-mover had agency over the realized outcome. In our experiments, we keep agency and the choice set of the first-mover constant across treatments. In particular, the first-mover is given the same two choices, these choices are common information to all players, and his choice is directly implemented. Therefore, the design does not allow for any ambiguity regarding the choice the first-mover wanted to implement, and does not present any variation across treatments in what else he could have chosen.

Importantly, we rely on variations in the possibility of punishment to manipulate perceived motives for being helpful. This choice is driven by identification challenges that arise from relying only on variations in reward incentives, as demonstrated by the comparison of Game 2 and Game 3 in the Introduction. As a notable exception to the literature testing for the role of perceived intentions, Stanca et al. (2009) provide an experimental design that aims to test for the role of perceived motives. In treatment 1, the first-mover decides on a transfer that gets multiplied before being given to the second-mover, and the second-mover in return decides on a transfer that gets

multiplies before being given to the first-mover, much like the Game 3 presented in the Introduction. In treatment 2, the first-mover makes the same transfer decision as the first-stage of treatment 1 in the context of a modified dictator game, without knowing that there will be a second-stage, much like Game 2. This decision is followed by a surprise, where the second-mover makes a transfer decision in a modified dictator game featuring the same decision as the second-stage of treatment 1. The results from this experiment show that the second-movers are more responsive to transfers in treatment 2. Even though we are tempted to ascribe the role of perceived motives to explain their results, as the authors point out, the results can also be explained by an account of perceived intentions, such as the Falk and Fischbacher (2006) model, because the first-mover is perceived to have expected the second-mover to obtain a higher payoff as a result of not having the opportunity to exercise the costly reward option.

The current paper builds on the existing literature in several important ways. The experiments provide evidence for differences in the level of positive reciprocity towards the same helpful action across treatments that vary incentives for the first-movers to choose the helpful action, in a manner that isolates the role of motives from the role of perceived intentions, or the impact of outcomes. Importantly, the experiments elicit higher order beliefs, allowing us to expose the mental models of subjects regarding others, to test the explanatory power of different reciprocity models in the context of our experiments, and to provide evidence for a relationship between the inferred inherent altruism of a helpful first-mover and the degree of positive-reciprocation his helpful action triggers. Finally, our design circumvents the need to mislead subjects about the nature of the interaction.

To achieve these contributions, both experiments feature four parts. The focal aim of both experiments is to provide evidence for differences in the reciprocal behavior in the second-stage of the reciprocal interaction presented as we manipulate the strength of strategic incentives to help in the first-stage. The reciprocal interaction is presented in part 3. Part 1 elicits other-regarding preferences in the absence of strategic considerations, and part 2 elicits expectations of other-regarding preferences in the given session. Inclusion of parts 1 and 2 allow us to cleanly identify reciprocity versus altruism (Cox, 2004), and beliefs regarding reciprocity versus altruism. In particular, part 1 provides a baseline for helpful behavior in part 3, and part 2 provides a baseline for part 4, which elicits beliefs regarding behavior in the sequential reciprocity game. Participants complete all four parts without receiving feedback about their performance or others' behaviors until the very end of the experiment.

Eliciting beliefs regarding equilibrium play is crucial when exploring the impact of notions that

depend on the second-mover's inferences regarding the first-mover, such as his intentions or his motives. Beliefs elicited in part 4 regarding behavior in the reciprocal interaction address the concern that differences in behaviors may be confounded with other strategic issues (Bolton et al., 1998), expose players' mental models of other players and crucially add to our ability to discern between theories or test their explanatory power by contrasting their predicted beliefs with subjects' elicited beliefs (Costa-Gomes and Weizsacker, 2008). Belief elicitations usually do not generate large distortions in choice data (e.g., Costa-Gomes and Waizsacker, 2008), but some studies find contamination (Gächter and Renner, 2010; Rutström and Wilcox, 2009). In our experiments, the possibility of such contaminations appears small, because we elicit beliefs regarding others' choices after choices are made. That said, we recognize that elicitation of behavior of interest might be cleaner if we had not elicited beliefs or other choices outside of the main decisions of interest due to a cross-task contamination that may arise when more than one risky payoff is simultaneously unrealized until the end of the experiment (Cox et al., 2015). Therefore, we also confirm our results with a robustness study (Study 2a) that presents the subjects only with the main interaction of interest. In addition, we present robustness studies that check for confounds that may arise from menu effects and self-serving beliefs (Studies 1a and 1b). Our results are robust to a variety of design choices. In the remainder of this section, we present each experiment, starting with the reciprocal interaction of interest, followed by our hypotheses, the protocol, and the results.

2.2 Experiment 1

The reciprocal interaction

Experiment 1 presents a two-stage reciprocity game (Game Γ_1) depicted in Figure 1. In the first-stage, the first-mover (player A) makes a choice between (S) and (H). If he chooses (S), he receives \$4 and the second-mover (player B) receives \$1. If he chooses (H), he sacrifices \$1 in order to increase the payoff of player B by \$1.

In the second-stage, having observed the decision of player A, player B in turn makes a decision that affects player A's material payoffs. If player A chooses (H), player B chooses between (l), whereby she leaves the distribution that player A delivered unchanged, and a costly reward option (r), whereby she can increase player A's payoffs by \$2.50 by sacrificing \$0.50 from her own payoffs. The outcome (H, l) yields (\$3, \$2) and the outcome (H, r) yields (\$5.50, \$1.50). If player A chooses (S), player B chooses between (l), whereby she leaves the distribution that player A delivered

unchanged, and a costly option whereby she sacrifices \$0.50 from her payoffs to change the payoff of player A to the amount m. Game Γ_1 takes on very different natures depending on the value of m.

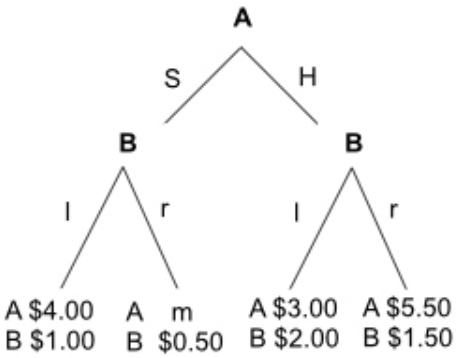


Figure 1: Game Γ_1

The experiment manipulates the first-mover's motives in a between-subject design by setting m=\$1.50 in treatment 1 and m=\$6.50 in treatment 2. The treatments are symmetric: treatment 1 (2) gives player B a costly punishment (reward) option if player A chooses (S), whereby she can decide to sacrifice \$0.50 in order to decrease (increase) player A's payoff by \$2.50. In treatment 2, Game Γ_1 is a simple trust game, offering first-movers the potential of being rewarded for choosing (H). Therefore, among the first-movers who choose (H), there could be a mix of people motivated by other-regarding preferences (altruistic motive) and/or hope of rewards (reward-seeking motive). In treatment 1, Game Γ_1 is a judgment game that also offers the potential of being punished for not choosing (H)[7]. Clearly, treatment 1 gives stronger strategic incentives for the first-mover to choose (H): the first-movers who choose (H) could be motivated by altruism, hope of rewards and/or fear of punishment (punishment-avoidance motive). Given that previous literature showed that sanctions in combination with rewards are more motivating than rewards alone (Andreoni et al. 2003), the manipulation in Experiment 1 should motivate more player As to choose (H) in treatment 1, and do so for a fear of punishment.

Note that the material payoffs of player B are the same across the two treatments. The design also keeps the action space of the first-mover, and that of the second-mover conditional on the first-

[7]Co-existence of rewards and punishments in the second-stage of a reciprocal interaction are not very common in the literature. Abbink et al. (2000) presented a "moonlighting" game where kind and unkind actions were available to both the first- and second-movers. Offerman (2002) presented a "hot response" game where rewarding and punishing were available options for the second-mover, regardless of first-mover's choice. Experiment 1 design is closer to Al-Ubaydli and Lee (2012) in presenting a "judgment" game where a reward option is available if the first-mover has been helpful, and a punishment option is available if the first-mover has chosen selfishly.

mover choosing (H) constant. The two treatments only vary in the action space of the second-mover conditional on the first-mover choosing (S). The payoffs of the second-mover are never higher than that of the first-mover, which guarantees that 1) players' relative standing is kept stable, and 2) punishing player A always decreases the magnitude of the inequality of material payoffs between player A and player B, and rewarding him always increases it.

Hypotheses: identifying the distinct role of motives

The central hypothesis of Experiment 1 is

H1: In response to player A choosing (H), a higher proportion of player Bs will choose (r) in treatment 2.

Would the consideration of perceived intentions would produce the same hypothesis? Let us consider the Dufwenberg and Kirchsteiger (2004) (DK) model and the Falk and Fischbacher (2006) (FF) models. Our experimental design limits the situations under which the FF and the DK models can predict this hypothesis to an undesirable set: a consideration of intentions could only predict the main hypothsis if the second-order expectations of player Bs are misaligned with equilibrium play or with first-order expectations. We defer the technicalities to the Appendix, and provide an intuitive explanation here.

The FF model defines the relative kindness of (H) from the perspective of player B based on the difference between player B's beliefs about player A's intended outcome for player B versus player A's intended outcome for himself as a result of choosing (H). A strict interpretation of the FF model would not predict any positive reciprocity to (H) in either of the treatments because player B's payoffs are always lower than those of player A. Since the central element of the FF model is the kindness term, we focus on discussing under what conditions the FF model would predict (H) to look less unkind in treatment 2 than in treatment 1. Experiment 1 restricts payoffs in the final nodes of Game Γ_1 such that rewarding A increases the inequality of material payoffs between player A and player B. Therefore, the only way (H) can be perceived as kinder in treatment 2 than in treatment 1 is if player Bs thought player As expected fewer player Bs to positively reciprocate in treatment 2. Clearly, this condition requires a misalignment between second-order expectations and the predicted equilibrium behavior.

The DK model defines the relative kindness of (H) from the perspective of player B based on the difference between player B's expected payoffs in the sub-game reached after (H) to the average of her expected payoffs across both sub-games. Since Experiment 1 keeps all of player B's payoffs

constant across the two treatments, the difference in the perceived kindness of (H) can only arise from the differences in player B's second-order beliefs. In order for the DK model to predict a higher degree of positive reciprocity to (H) in treatment 2, either the second-order expectations of reward in response to the helpful action, i.e. (H, r), should be higher in treatment 1, which is inconsistent with the predicted behavior itself, or second-order expectations regarding (S, r) should be higher in treatment 2. Importantly, if the second-order expectations are consistent with expected equilibrium play, the DK model predicts the opposite of our main hypothesis: the choice of (H) is associated with kinder intentions, thus commands higher reciprocity in treatment 1 than in treatment 2, mainly because the benchmark payoff the second-mover could have received is lower if a larger proportion of player Bs would pay to punish (S) rather than pay to reward it.

The discussion of perceived intentions as an alternative account highlights the importance of eliciting higher order beliefs. If equilibrium behaviors and beliefs align, the central hypothesis cannot be predicted by a consideration of intentions alone. Therefore, in addition to the main hypothesis of this experiment, we conjecture that

H2a: Player Bs expect more player As to choose (H) in treatment 1 than in treatment 2.

H2b: Expectations of player As regarding the likelihood of player Bs choosing (r) in response to (S) are higher in treatment 1 than in treatment 2.

H2c: Expectations of player As regarding the likelihood of player Bs choosing (r) in response to (H) are higher in treatment 2 than in treatment 1.

H2d: Player Bs think that the expectations of player As regarding the likelihood of player Bs choosing (r) in response to (S) are higher in treatment 1 than in treatment 2.

H2e: Player Bs think that the expectations of player As regarding the likelihood of player Bs choosing (r) in response to (H) are higher in treatment 2 than in treatment 1.

Testing these ancillary hypotheses allow us to see whether second-movers are sophisticated about the incentives first-movers face and if the first-movers understand how second-movers feel. Importantly, these hypotheses help distinguish the role of perceived motives from alternative explanations based on an account of perceived intentions.

The Protocol

A total of 258 subjects above the age of eighteen were recruited through ORSEE to participate in eighteen 60-minute sessions at the [blinded for review] Lab during November 2014. In each session, an even number of participants (10 to 20 participants per session) interacted using the software

Z-Tree (Fischbacher, 2007) in a double-blind payoff protocol. Only one treatment was implemented for all subjects in a session, and subjects could only participate in one session. The participants were told that the session would last 60 minutes and had 4 parts. Each part was introduced with its own set of instructions to all subjects at the same time. Subjects were informed that their payments from each part were independent of their choices in the future or previous parts of the experiment. All identities and choices were kept anonymous throughout the experiment.

Subjects earned a fixed participation fee of \$5. They also earned additional payments from each of the four parts. If the parts included more than one task, one task was selected at random from each part to determine additional payments. Subjects learned the randomly selected tasks and their earnings at the end of the study. The average total earnings were \$15.14. Experimental instructions, questions and detailed protocol are included in the Experimental Instructions Appendix.

In part 1, half the participants were randomly and anonymously assigned the role of player A and the rest were assigned the role of player B. The participants kept these roles throughout the experiment. Player As made decisions in six binary modified dictator games, while player Bs waited. Player As were asked to choose between (\$4.50, \$1.50) and (\$4, \$4); (\$2.50, \$0) and (\$2, \$1.50); (\$4, \$1) and (\$3, \$2); (\$5, \$2) and (\$4, \$4); (\$1, \$4) and (\$0.50, \$6.50); (\$2, \$3) and (\$1.50, \$5.50) where the first amount denotes the payoff of player A and the second denotes that of player B. Note that these choices included some of the same binary options player A and player B would choose between later in the context of Game Γ_1. All participants were told that one game from part 1 would be randomly chosen, and player A's choices in that game would determine payments for that player A and a randomly matched player B at the end of the experiment.

In part 2, four of the modified dictator games from part 1 were presented to all the participants. Participants were incentivized to predict the percentage of player As in that session who had chosen each option in the following decision tasks presented as modified dictator games: (\$2.50, \$0) and (\$2, \$1.50); (\$4, \$1) and (\$3, \$2); (\$1, \$4) and (\$0.50, \$6.50); (\$2, \$3) and (\$1.50, \$5.50). They earned \$4 if they guessed the proportion of player As who picked each option correctly, and their earnings declined quadratically as a function of their inaccuracy.[8] They were informed that one question would be chosen at random at the end of the experiment to determine their earnings from part 2.

In part 3, all participants in a given session made decisions in either treatment 1 or treatment 2

[8]Quadratic scoring rule is commonly used in other studies. Please see Armantier and Treich (2013), Schotter and Trevino (2014), and Trautmann and van de Kuilen (2015) for in depth discussions of different methods of belief elicitation in the lab.

versions of Game Γ_1. All the details of the game were explained to all participants at the same time. The program matched player As and player Bs randomly and handled communication of choices anonymously.

In part 4, player A's first-order beliefs about player B's responses, player B's first-order beliefs about player A's choices were elicited. For example, we asked player A's "What percentage of Person B's chose each option (r or l) in response to S?," and prompted them to make sure that the percentage of choices indicated added up to 100%. We also elicited player B's second-order beliefs (expectations regarding player A's first-order beliefs). For example, we asked "We asked Person A's "What percentage of Person B's chose each option (r or l) in response to S?" What do you think was the average of their predictions?." Player B's answered by filling in the blanks in two statements of the form: "On average, Person A's expected ___% of Person B's to choose r (or l) in response to S." The participants were again incentivized for accuracy using a quadratic scoring rule, and were informed that one question would be chosen at random at the end of the experiment to determine their accuracy payments from part 4.

At the end of the experiment, the program displayed the earnings to each participant, and explained how these earnings were achieved by going over their decisions in the tasks that were selected from each part. Each participant was paid privately.

We briefly explain our motivation for including all four parts in the experimental design. Asking Player As to make choices in modified dictator games in part 1 allows us to learn about their other-regarding preferences (Charness and Rabin, 2002)[9]. Note that part 1 included a game that presented the same choice options as (S) and (H) in Game Γ_1, as well as games that presented the same choice options that player Bs in the two sub-games of Game Γ_1 would face. Knowing their choices in a situation where player Bs cannot respond provides information regarding how much of the helpful behavior in Game Γ_1 results from strategic considerations. The predictions elicited in part 2 serve as baseline beliefs about the degree of altruism in the population of participants in a given session. This information is useful in determining whether the beliefs elicited in part 4 reflect an understanding of strategic considerations on the part of the first-movers, as well as an understanding of the mental model of the second-movers. Therefore, results from the first two parts

[9]One may be concerned that telling the subjects that there will be another decision task following the dictator games may make the dictators more generous. If such a bias existed, we would underestimate the degree of strategic considerations in the reciprocal interaction in both treatments. Cox, Sadiraj and Sadiraj (2008) investigated this methodology question. They found that tests for trust, fear and reciprocity using data from a within-person experiment that involves the moonlighting game as well as dictator games imply the same conclusions as tests using data from across-subjects experiments where different groups of people play these games.

of the experiment can provide baseline of behavior and expectations when reciprocal or strategic considerations are absent. Finally, the beliefs elicited in part 4 are useful in establishing internal validity of the experiment, exposing the mental models of players regarding the game and the other player, identifying beliefs regarding reciprocity separately from beliefs regarding distributional preferences, and for ruling out alternative explanations based on second-order beliefs.

Results

The identification of the impact of perceived motives relies on exogenously shifting beliefs of second-movers regarding why the first-mover might have chosen (H). Therefore, we begin with an investigation of beliefs elicited in part 4. Table 1 summarizes these beliefs, and Figure 2 illustrates them. The leftmost panel reports the second-movers' first-order expectations (B FOE). These expectations are indicated in gray in Figure 2. Across the two treatments, player Bs expected meaningful differences in the extent to which player As were willing to choose (H) in treatment 1 (62%) versus in treatment 2 (41%) (Two-sample Wilcoxon rank-sum (Mann-Whitney) test, $z = 5.78$, $p = 0.000$). This result supports hypothesis *H2a*, suggesting that second-movers understand the incentives each treatment presents to the first-movers, which is a prerequisite for contemplating the first-mover's motives.

The middle panel of Table 1 reports the first-movers' first-order expectations (A FOE), and Figure 2 depicts them in red. First, let's consider expectations of behavior in the sub-game that follows (S). We see that fear of punishment is likely to have motivated player As to be more helpful in treatment 1, as player As reported a relatively high potential punishment expectation (41% on average) in this treatment 1 (one sample t-test, $t = 36.52$, $p = 0.000$). On average, player A's expect 19% of player B's to help in response to (S) in treatment 2.[10] Therefore, in support of hypothesis *H2b*, the expectations of player As regarding the likelihood of player Bs choosing (r) in response to (S) is higher in treatment 1. Also, in support of hypothesis *H2c*, expectations of player As regarding the likelihood of player Bs choosing (r) in response to (H) are 40% in treatment 2 and 30% in treatment 1 (Two-sample Wilcoxon rank-sum (Mann-Whitney) test, $z = -1.74$, $p = 0.082$). Expectations of positive reciprocity can be identified by comparing how likely a player A thought

[10]This expectation may seem optimistic given that only 1 out of 24 player Bs who faced a choice in this subgame chose ($0.50, $6.50) over ($1, $4), however, the number of observations in the sub-game are too low to conclude that the belief is biased. Interestingly, player B's also expected a similar (16%) fraction of other player Bs do so. Elicited beliefs may be displaying some conservatism bias. For reference, 31% of subjects who faced a choice between the same options picked ($0.50, $6.50) over ($1, $4) when these options were presented in a binary modified dictator game context in part 1.

the general population of participants were to choose (r) over (l) in a modified dictator game that presented the same options (elicited in part 2) to how likely they thought this choice was when the person was responding to (H) in Game Γ_1 (elicited in part 4). Player As reported an average expectation of only 24% of people making the same choice in part 1. Within-person differences in expectations reveal that player As expected positive reciprocity (over and beyond an expectation of altruism) in treatment 2, (Wilcoxon sign-ranked test, $z = -4.25$, $p = 0.000$), but not in treatment 1 (Wilcoxon sign-ranked test, $z = -1.01$, $p = 0.311$).

Table 1: Beliefs about Game Γ_1 play across two conditions

	N^{11}	B FOE		A FOE				B SOE			
		S	H	l|S	r|S	l|H	r|H	l|S	r|S	l|H	r|H
Treatment 1	59	28%	62%	59%	41%	70%	30%	46%	54%	73%	27%
Treatment 2	70	59%	41%	81%	19%	60%	40%	84%	16%	57%	43%

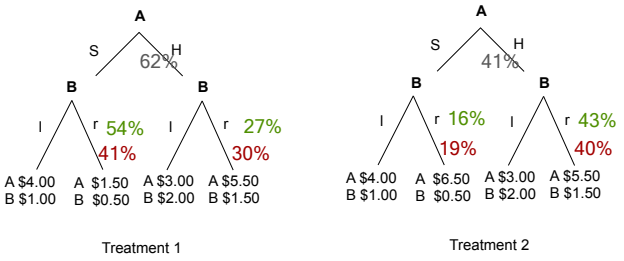


Figure 2: Beliefs about Game Γ_1

Player B's second-order beliefs are reported in the rightmost panel of Table 1, and depicted in green in Figure 2. There is a strikingly close alignment of Player As' FOEs and Player Bs' SOEs, in support of hypotheses *H2d* and *H2e*. Moreover, as we demonstrate below, these expectations are also reflective of actual behavior. On average, player B's believe that player As expect an average of 43% of player Bs to reward (H) in treatment 2, and an average of 27% of player Bs to reward (H) in treatment 1 (Two-sample Wilcoxon rank-sum (Mann-Whitney) test, $p = 0.006$). Similarly, player B's believe that player As expect an average of 16% of player Bs to reward (S) in treatment 2, and an average of 54% of player Bs to punish (S) in treatment 1 (Two-sample Wilcoxon rank-sum (Mann-Whitney) test,). In sum, the elicited beliefs suggest a high level of sophistication among subjects about each others' behavior and expectations.

Table 2: Behavior in Game Γ_1

	# sessions	N	A Choice		B Response			
			S	H	l|S	r|S	l|H	r|H
Treatment 1	18	118 (59 pairs)	4	55	3	1	36	19
Treatment 2	10	140 (70 pairs)	24	46	23	1	20	26

We now turn to the behavior in the sequential reciprocity game. Note that the first-stage of Game Γ_1 presents a choice between {\$4 for player A, \$1 for player B} and {\$3 for player A, \$2 for player B} to player As. When player As were asked to choose between the same options in part 1 where player Bs could not respond in any way, only 29% of them chose to transfer \$1 from their payment to the other subject, and 71% chose to keep all \$4 to themselves.[12] We expect a larger fraction of player As to transfer \$1 in the first-stage of Game Γ_1 than they did in part 1, as Game Γ_1 offers strategic incentives for doing so in both treatments, and player As expectations regarding punishment and rewards reflect an understanding of these incentives. Table 2 summarizes the choices observed in Game Γ_1 across the two conditions. We see that player As are more willing to sacrifice \$1 to help player B in the first-stage of Game Γ_1 than in part 1, in both treatment 1 (93% vs. 29%, McNemar test, $\chi^2(1) = 39$, $p = 0.000$) and treatment 2 (66% vs. 29%, McNemar test, $\chi^2(1) = 23.15$, $p = 0.000$). Furthermore, more player As choose (H) in treatment 2 than in treatment 1 (93% vs. 66%, Chi-square test, $\chi^2(1) = 14.25$, $p = 0.000$).[13] Thus, the manipulation of the incentives across the treatments achieved its objective: the proportion of player As who are strategically motivated within the set of player As who choose (H) is higher in treatment 1.

In support of the main hypothesis of the experiment that the choice of (H) will trigger a higher degree of positive reciprocity in treatment 2 than in treatment 1, only 20 out of 56 (36%) of player Bs rewarded (H) in treatment 1, whereas 26 out of 45 (58%) of player Bs rewarded (H) in treatment 2 (Chi-square test, $\chi^2(1) = 4.90$, $p = 0.027$). This result suggests that second-movers are less likely to positively reciprocate to the same helpful action when the reciprocal interaction provides stronger strategic incentives for the first-movers to be helpful. This evidence demonstrates the distinct role of perceived motives in reciprocal decision making. Because second-order beliefs of the player Bs are consistent with equilibrium behavior and first-order beliefs of player As, the data cannot be

[12] Detailed results from part 1 and part 2 are presented in Table 5 of the Appendix. In line with early findings of Charness and Rabin (2002), the results from part 1 suggest that dictators are more likely to sacrifice their own payoffs to help another person when their payoffs are higher than the other person, and when the sacrifice produces a larger gain on the part of the other person. Beliefs elicited in part 2 properly reflect the choice ordering elicited in part 1, even though we find evidence for projection and conservatism bias as in other studies.

[13] Clearly, player Bs' estimations of the proportion of player As who would choose (H) are conversative, i.e. biased towards the uniform, but their beliefs correctly reflect the ordering across treatments.

rationalized by the intention-based reciprocity theories proposed by Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006). In Section 3, we discuss the explanatory power of of models that consider inferences of kindness regarding the first-mover (Cox, Friedman and Sadiraj, 2008; Gül and Pesendorfer, 2010).

Robustness Checks and Additional Evidence We also conducted two separate studies in order to check the robustness of our results to the design features of Experiment 1. We include the protocol and detailed results of these studies in the Experimental Appendix. Study 1A reports third-party beliefs to provide support that the beliefs elicited from participants are not driven by self serving biases. Study 1B checks whether second-movers' choices in the sub-game reached after (H) could be driven by the mere existence of different alternatives in the sub-game reached after (S). For both studies, participants were recruited among those who had not participated in Experiment 1.

In Study 1A, we asked the same belief questions that Player As and Player Bs responded to in part 4 to 121 participants who had not participated in Experiment 1 or Experiment 2. The participants were incentivized based on the accuracy of their answers. We found that the beliefs of third-parties are in line with the beliefs of the players in the game. They predicted more player As to choose (H) in treatment 1 (60%) than in treatment 2 (37%), and this difference was highly significant (Wilcoxon sign-ranked test, $z = 8.05$, $p = 0.000$). Conditional on player A choosing (H), third-parties expected 30% of player Bs to reward player in treatment 1, and 34% of them to do so in treatment 2 (Wilcoxon signed-rank test, $z = 2.93$, $p = 0.003$). Conditional on player A choosing (S), third-parties expected on average 47% of player Bs to punish player in treatment 1, and 18% of player Bs to reward player A.

In addition, Study 1A also asked these third-parties to predict the proportion of player As who would have helped in the absence of any strategic considerations among those who chose to be helpful in each of the reciprocal interactions. In particular, third-parties were asked two questions. First, they were asked to predict the proportion of Person As who chose (\$3, \$2) over (\$4, \$1) when choosing between these options in part 1, where Person Bs could not respond. They predicted 25% of the Player As to make this choice. Then, they were asked to predict the proportion of Person As who made the same choice among the Person As who chose (H) over (S) in part 3. On average, third-parties predicted that only 43% of helpful first-movers in treatment 1 also chose (\$3, \$2) over (\$4, \$1) when choosing between these options in part 1. In other words, they predicted that the

remaining 57% of the helpful player As were motivated by strategic considerations in treatment 1. In comparison, they predicted 49% of the helpful first-movers in treatment 2 to be motivated by strategic considerations (Wilcoxon signed-rank test, $z = 3.79$, $p = 0.0001$). These beliefs reveal that third-parties are able to recognize the difference in the mix of motives between treatment 1 and treatment 2. Overall, the data presented by Study 1A suggests that the beliefs reported by second-movers in Experiment 1 are not driven by self-serving biases, and reflect a clear understanding of differences in the potential motives if the first-mover across the two treatments.

In study 1B, we checked whether the differences in choices in the sub-game reached after (H) across our treatments could be driven simply by the existence of different alternatives that the sub-game reached after (S). In particular, we modified the game such that player As had no choice to make. We informed player Bs that the computer picked H, and asked them to choose between (\$3, \$2) and (\$5.50, \$1.50) under two treatments: either when their choice would have been between (\$4, \$1) and (\$1.50, \$0.50) if the computer would have picked S (treatment 1), or when their choice would have been between (\$4, \$1) and (\$6.50, \$0.50) if the computer would have picked S (treatment 2). Basically, Study 1B replicates Game Γ_1, but player A has no agency over whether (S) or (H) is chosen. We find no differences in player B's choices across the two treatments in Study 1B. Out of 63 player Bs, 38% of them chose (\$5.50, \$1.50) over (\$3, \$2) in treatment 1, and 39% of them made the same choice in treatment 2. Therefore, the differences in positive reciprocity across our original treatments are not driven by the mere existence of different options in the sub-game that followed (S), but hinge on Person A deliberately considering how Person B could respond if he chose (S).

Summary

In sum, Experiment 1 compares the level of positive reciprocity second-movers display towards a helpful first-mover in a situation where the first-mover could have been motivated to help because he feared punishment (treatment 1), with the level of positive reciprocity to the same helpful action in a situation where the punishment option in the second-stage was absent (treatment 2). The results demonstrate that in treatment 1: i) more player As choose (H) due to a punishment-avoidance motive, ii) player Bs expect this to be the case, and iii) the choice of (H) triggers lower reciprocity from player Bs. Importantly, player As expect this withdrawal of concern, and player Bs expect player As to expect it. We present additional evidence that third parties also share these beliefs, and the difference in the second-mover's degree of positive reciprocation is not driven by menu effects. Overall, the results provide strong support for a distinct role of perceived motives in reciprocal-

decision making.

2.3 Experiment 2

Experiment 2 extends Experiment 1 in several aspects. First, it explores the mechanism by which the existence of strategic incentives may influence reciprocity. We expect there to be a close relationship between perceptions of motives and perceptions of altruism regarding a helpful person. Therefore, Experiment 2 elicits beliefs about the altruism of the person taking the helpful action in a within-subjects design. Second, Experiment 2 allows for comparing the role of perceived reward-seeking motives and the role of perceived punishment-avoidance motives separately to the no-incentive benchmark, while Experiment 1 features a potential for the first-mover to be motivated by reward-seeking across all treatments.

The reciprocal interaction

Experiment 2 investigates positive reciprocity in the context of a probabilistic sequential game where the same helpful action could be motivated by punishment-avoidance, reward-seeking, and/or altruism. Consider Game Γ_2 depicted in Figure 2. First, Player A chooses between (S), which pays him \$4.50 and player B \$2.50, and (H), which pays both players \$4. Therefore, in the first-stage player A decides whether to take the option that pays him more, or the option where he sacrifices \$0.50 to increase player B's earnings by \$1.50. Then, nature chooses either 1, 2 or 3. If Nature chooses 2, the game ends, and the option player A chose determines both players' final payments. If nature chooses 1, the game ends if player A chose (H). But if Player A chose (S), then player B chooses between not altering the allocation player A choice (N) or paying \$0.50 to *decrease* player A's earnings by \$1.50 (P). The outcome of (H,2) yields (\$4, \$4), outcome of (H,2,N) yields (\$4.50, \$2.50) and the the outcome of (S,2,P) yields (\$3, \$2). If nature chooses 3, the game ends if player A chose (S). But if player A chose (H), then player B gets to choose between (N) and (R). The choice of (N) leaves the allocation player A chose unaltered. The choice of (R) costs player B \$0.50 and *increases* player A's earnings by \$1.50. The outcome of (S,1) yields (\$4.50, \$2.50), the outcome of (H,1,N) yields (\$4, \$4) and the the outcome of (H,1,R) yields (\$5.50, \$3.50).

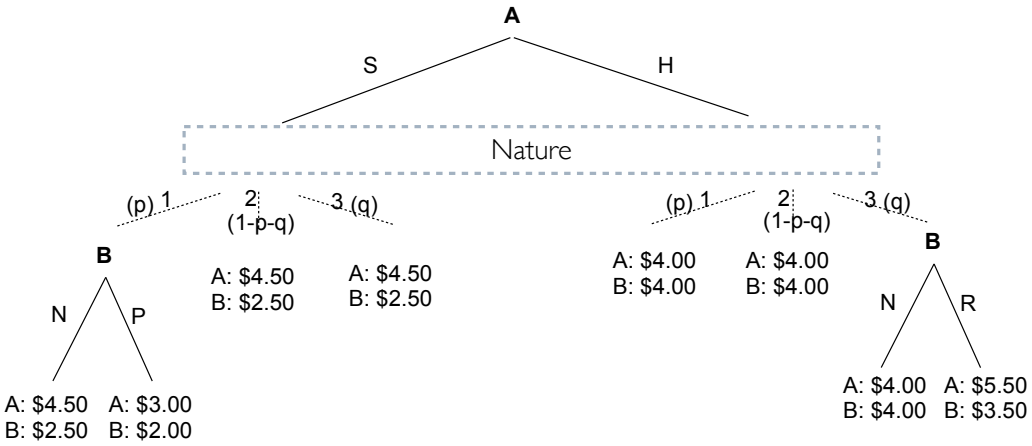


Figure 3: Game Γ_2

Let p be the probability that nature chooses 1, and q be the probability that nature chooses 3. Consider how changing p and q may affect player A and player B behavior. First, consider q approaching 1. Then if player A chooses (S), player A will get \$4.50. But if player A chooses (H), Person B may choose (R), giving player \$5.50. Depending on his belief about how likely player B is to choose (R), player A may be inclined to choose (H) even if he puts no weight on player B's well-being. Thus, as q gets larger, there will be more player As who are primarily motivated by reward-seeking motives among those who choose (H). Similarly, consider p approaching 1. Then if player A chooses (H), player A will earn \$4. But if player A chooses (S), player B may choose (P), giving player A only \$3. Therefore, player A may be inclined to choose (H) in order to avoid potential punishment. Thus, as p gets larger, there will be more player As who are primarily motivated by punishment-avoidance among player As who choose (H). In contrast, when p+q is very close to zero, player A would mainly choose (H) if he genuinely prefers the more equitable allocation (\$4, \$4) to the more profitable allocation (\$4.50, \$2.50).

Experiment 2 features three treatments. In treatment 1, $p = .98$, and $q = .01$, therefore the first-stage player expects the second-stage player (almost always) to have the option to punish a selfish action. In treatment 2, where $p = .01$, and $q = .01$, and therefore the first-stage player expects the second-stage player to be a passive recipient most the time. In treatment 3, $p = .01$, and $q = .98$, therefore the first-stage player expects the second-stage player (almost always) to have

the option to reward a helpful action. In treatment 2, player A is mainly motivated by altruism, however in treatments 1 and 3, he can also be motivated by punishment-avoidance and reward-seeking respectively.

We want to compare how likely player Bs are to reward player A for choosing (H) across the three treatments. The probabilistic design allows us to elicit reward demand from player Bs even in cases where player Bs were expected to be able to reward player As, without misleading participants about the nature of the interaction. In each of the three treatments, player B is asked to designate her response for each contingency using the strategy method.[14] In particular, player Bs are asked to indicate, in each treatment, whether they would choose (R) or (N) if player A chose (H) and Nature chose 1. They are also asked to indicate whether they would choose (P) or (N) if player A chose (S) and Nature chose 2 in each treatment. For example, in treatment 2, player Bs will not get the chance to act 98% of the time, and therefore player As are mostly motivated by altruism, and player Bs know that. We ask player Bs whether they would choose (N) or (R) if they faced with these options. In this manner, we learn about the reciprocal preferences across all treatments.

Is player B more likely to want to reward (H) when player A expected her not to be able to respond or when he expected her to be able to punish? Note that unlike Experiment 1, Experiment 2 focuses on one strategic motivation in each treatment, minimizing any expectations of the alternative strategic motivation. As a result, it permits comparing reciprocity to a helpful action in a case where a punishment-avoidance motive is likely to drive helpfulness to reciprocity to the same action in a case where strategic motivations are not likely to drive the first-mover's choice. Based on an account of perceived motives, we hypothesize that player B is more likely to want to reward (H) in treatment 2 compared to treatment 1. Comparing reciprocity differentials between treatments 2 and 3 is interesting, however, as we discussed in the Introduction, such differences can also be driven by intention-based reciprocity.[15] Therefore, while we also analyze and present the response differences across treatments 2 and 3, we rely on the comparison of reciprocal differences across treatments 1

[14]Investigating differences between the direct-response and the strategy methods, Brandts and Charness (2011) showed that any treatment effect demonstrated using the direct-response method could also be demonstrated using the strategy method. Also, Charness and Levine (2007) noted that the strategy method should be innocuous if it does not interact with the treatment status and tests changes in the rate of positive responses, rather than the level of the rate.

[15]Moreoever, note that Nature is equally unlikely to choose 1 in treatments 1 and 2, but is very likely to choose 1 in treatment 3. We may worry about comparing responses elicited across sub-games that have drastically different probabilities of being carried out. For example, when the probability of implementation is low, the participants may have other objectives, such as seeming nice to the experimenter. We tried to eliminate such concerns by making all actions and all pairings anonymous. More importantly, any bias generated by the low probability of the event should be common to both treatments 1 and 2. As a result, we do not expect such potential biases to impact the difference in reciprocal responses between treatments 1 and 3 - the main comparison of interest in Experiment 2.

and 2 to provide evidence for the distinct role of motives. For reasons similar to those discussed in Experiment 1, intention-based accounts do not predict this outcome. While the FF model does not predict a difference in the perceived kindness of choosing (H), the DK model would predict (at least weakly) lower level of positive reciprocity in treatment 2 than in treatment 1. In the context of the DK model, the perceived kindenss of (H) is higher in treatment 1 than in treatment 2 because choosing (S) may lead to player B sacrificing her payoffs to punish player A in treatment 1. The technical details are included in the Appendix.

The Protocol

A total of 176 participants above the age of 18 were selected from the undergraduate and graduate student population to participate at [blinded for review] University participated in eleven 45-minute sessions conducted at the [blinded for review] Lab on September 2013. Each session had 12-20 subjects. Subjects earned a $5 participation fee and up to $5.50 in additional earnings.

In part 1, half the participants were randomly and anonymously assigned the role of player A and the rest were assigned the role of player B. The participants kept these roles throughout the experiment. In part 1, all participants made decisions in eight modified dictator games, each of which presented two options where the payoffs were denoted in tokens.[16] The conversion rate was 200 tokens = $1. In particular, all participants were asked to choose between (800, 800) and (700, 1100); (800, 200) and (600, 400); (900, 500) and (800, 800); (500, 900) and (400, 1200); (500, 0) and (400, 300); (900, 0) and (800, 200); (400, 600) and (300, 1100); (500, 900) and (400, 600).

The choices in part 1 elicited other-regarding preferences from player As and player Bs.[17] Knowing each person's other-regarding preferences allows us to separately identify transfers resulting from reciprocity and the role of strategic incentives in Game Γ_2. Elicitation of preferences in these games also allows us to incentivize reporting of truthful expectations regarding the genuine kindness of helpful first-movers in Game Γ_2, as we explain below.Subjects were presented with games that offered the same choice options as (S) and (H) that player As would face in the first stage of Game Γ_2. Similarly, subjects also made choices among the same choice options that player Bs would face in the different possible sub-games of Game Γ_2.

[16]In Experiment 2, we denote payment in tokens to confirm with the norms of the lab at the time we collected data. We have no reason to suspect that the token conversion would generate biases that would interact with the differences across treatments. Moreover, the robustness Study 2A, which tests treatments 1 and 2 in a between-subject design, uses dollar amounts.

[17]At the end of the experiment, if a dictator game was chosen to determine the payments, pairs of subjects and their roles were randomly assigned. Please see the Experimental Appendix for further details.

In part 2, participants were asked to predict the percentage of participants in that session who chose each option across four of the modified dictator games from part 1. The participants were incentivized for accuracy of their reported expectations.[18] These predictions served as baseline beliefs about the degree of genuine kindness in the population of participants in a given session.

Part 3 presented the subjects three within-person treatments of Game Γ_2.[19] The payoffs were denoted in tokens, where 200 tokens=\$1. Participants were randomly assigned to the role of player A and player B. Each player A was randomly and anonymously matched with one player B for each treatment. As player As made a choice between (S) and (H) in each treatment, player Bs were asked to indicate their preferred choices for each contingency using the strategy method in that treatment.

In part 4, player A's first order beliefs about player B's responses, and player B's first order beliefs about player A's choices for each treatment were elicited using accuracy incentives.[20] Part 4 also elicited player Bs' altruism inferences regarding player As who were helpful in each treatment. We wanted to know what proportion of the helpful player As player Bs thought would have behaved similarly if it weren't for the reciprocal nature of each treatment of Game Γ_2. In particular, player Bs were asked "Only consider the group of player As who chose H in (a given treatment). Among these player As, what percentage chose each of the following options presented to them in Part 1 of the study? Option 1. 500 tokens for him/herself, 0 for the other participant _____% , Option 2. 400 tokens for him/herself, 300 for the other participant _____%." Note that both in the first-stage of Game Γ_2 and in this modified dictator game, player As decide whether they want to sacrifice 100 tokens (\$0.50) in order to increase the payoff of player B by 300 tokens (\$1.50). Therefore, player Bs' beliefs regarding helpful player As' choices in this modified dictator game gives us an idea of their beliefs regarding how they would choose in the first-stage of Game Γ_2 if it were not for strategic considerations. We employ a within-subjects design in Experiment 2 in order to test whether the heterogeneity in these altruism inferences across subjects explain the heterogeneity they display in their reciprocal responses.[21] At the end of the experiment, one question was chosen at random to

[18] In this experiment, we used a linear scoring rule for simplicity.

[19] The order of the three treatments were counterbalanced among the following three treatment sequences (1-2-3, 2-3-1, 3-2-1). The ordering of the treatments did not affect any of the results.

[20] Note that because these beliefs are incentivized with respect to the actual proportions in the session, we could only ask for the FOE's of player As concerning the reaction of player Bs in the sub-game that was implemented with 98% chance.

[21] Charness, Gneezy and Kuhn (2012) discuss the advantages and disadvantages of within and between subject designs. This article establishes the influence of perceived motives on reciprocity with both designs. In both designs, we minimize experimenter demand effects by keeping choices anonymous. In Experiment 2, we vary the order of treatments to control for anchoring, but do not find any order effects. Moreover, we also provide results a between

determine payments of all participants. Further details of the instructions, questions and protocol of Experiment 2 are included in the Experimental Instructions Appendix.

Results

If player As are motivated by strategic considerations above and beyond altruism, and if punishment is a stronger motivator than rewards, we expect the percentage of player As choosing (H) is the greatest in treatment 1, followed by in treatment 3 and the least in treatment 2. In part 1, when faced with the same two options in the first-stage of Game Γ_2, 44% of player As chose the option {900 tokens for me, 500 tokens for another participant} over the option {800 tokens for me, 800 tokens for another participant}, and player Bs on average expected 46% of player As to do so.[22] Table 3 presents the behavior and expectations in Game Γ_2 across the three treatments. A total of 81% of player As chose (H) in treatment 1, followed by 72% in treatment 2 and 48% in treatment 3 (matched-pairs sign test, 2<3: $p = 0.000$; 2<1: $p = 0.000$; and 3<1: $p = 0.048$). This data shows that both the possibility of rewards and the possibility of punishment are successful motivators, with the possibility of punishment being the stronger of the two. Clearly, player As would not have been motivated by the mere existence of reward and punishment options in player B's disposal, if they did not think that player Bs were likely to use them when they had access to these options. Indeed, player As on average expected 40% of player Bs to choose R in treatment 3 if they chose (H), and they on average expected 43% of player Bs to choose P in treatment 1 if they chose (S).

Table 3: Observed Behavior and First-Order Beliefs in Game Γ_2

Treatment	N	% As	B's FOE	% Bs	A's FOE	% Bs	A's FOE
		choosing H	of H	choosing R | H	of R | H	choosing P | S	of P | S
Treatment 1	88	81%	75%	42%		50%	43%
Treatment 2	88	48%	51%	63%		43%	
Treatment 3	88	72%	67%	52%	40%	42%	

If player Bs understand the differences in the incentives, we expect them to predict meaningful differences in willingness to choose the helpful action across treatments. In accordance with player As' choices, player Bs expected the highest proportion of player As (75%) to choose (H) in treatment 1, followed by player As in treatment 3 (67%), and the lowest proportion of player As in treatment 2 (51%) (matched-pairs sign test, 2<3: $p = 0.000$; 2<1: $p = 0.000$; and 3<1: $p = 0.008$). This

subject design of treatments 1 and 2 below as a robustness check.

[22]The detailed results from Part 1 and part 2 are presented in the Appedix, Table 6.

result suggests that player Bs understood the differences in motivations across treatments regading why player As were willing to choose (H).

If player Bs care about why player A was helpful, above and beyond that he was helpful, we expect the following central hypothesis of this experiment to hold:

H2: The intended positive reciprocity in response to (H) is higher in treatment 2 than in treatment 1.

The percentage player Bs choosing R in response to (H) in treatment 2 is 63%. Given that player Bs have to move away from an equal distribution of 800 tokens for each player to 700 tokens for themselves and 1100 tokens for player A in order to reward the choice of (H), and only that 20% did so in part 1 in the context of a modified dictator game, 63% is a substantially positive reciprocal response. As hypothesized, player Bs were less reciprocal in treatment 1. Only a total of 42% of player B's indicated that they would choose R if player A chose (H) in treatment 1 (matched-pairs sign test, $p = 0.000$). This result replicates the main finding of Experiment 1 in a very different context: player Bs reciprocate less to the same helpful action when it could have been motivated by a strategic consideration of punishment avoidance. Therefore, Experiment 2 provides additional support for the distinct role of perceived motives on reciprocal decision-making.

Even though comparison of treatments 2 and 3 treatments are not our main focus, we would like to comment on the results in light of the previous experimental literature. The comparison of positive reciprocity across treatments 2 and 3 resemble the between-subject variation tested by Stanca et al. (2009). We find that only 52% of player B's indicated that they would choose R if player A chose (H) in treatment 3, which is somewhat smaller than the reciprocation rate in treatment 2 ($p = 0.047$). Therefore, our results seem to support the findings of Stanca et al. (2009), and could indicate that the reciprocity towards the same helpful action is lower when the action is more likely to be motivated by reward-seeking. Strassmair (2009) also compares two treatments similar to the 2 and 3 treatments in Experiment 2 by varying the probability with which the second-mover in a trust game could reciprocate to the first-mover's choice, but does not find any differences in the second-mover's reciprocal choices to a given level of transfer. Experiment 2's design differs from Strassmair (2009) in multiple ways, making direct comparisons of results difficult.

In addition to the first- and second-order beliefs, Experiment 2 also elicited a novel belief construct in part 4 to proxy for perceived motives. Player Bs reported the percentage of player As among those who chose (H) over (S) who would have made equally altruistic choice of choosing {400 tokens for player A, 300 tokens for another participant} over {500 tokens for player A, 0 to-

kens for another participant} in part 1, where they had no strategic incentives of doing so. Clearly, the perceived motive behind a helpful action is closely tied to the perceived altruism of the person taking the action. Therefore, the elicitation of this belief aims to proxy for their expectations regarding the fraction of intrinsically motivated player As among the helpful ones in each treatment. If player B's are sophisticated about the selection of player As induced by strategic motivations, we would expect them to report higher expectations of altruistic player As among H-choosers in treatments where strategic incentives are weaker.

Player Bs predicted on average 73% of player As among those who chose (H) in treatment 2 to choose {400 tokens for player A, 300 tokens for another participant} over {500 tokens for player A, 0 tokens for another participant}. However, they predicted only 54% of player As who chose (H) in treatment 1 to make the same choice (matched-pairs sign test, $p = 0.000$).[23] These results suggest that player Bs believed that strategic incentives to avoid punishment lead to a lower proportion of truly generous people among those who choose the helpful action than reward incentives do.[24] Interestingly, player Bs also correctly inferred that the H-choosers in treatment 1 are not kinder than the population of player As in general, since they had reported an expectation (elicited in part 2) of 56% of player As choosing {400 tokens for player A, 300 tokens for another participant} over {500 tokens for player A, 0 tokens for another participant}.

The within-person design of Experiment 2 also allows us to investigate whether differences in individual player B's altruism inferences across treatments are associated with changes in their responses. Although, overall, player Bs think that a lower proportion of the H-choosers in treatments 1 and 3 are altruistic than in treatment 2, they display considerable heterogeneity in the degree to which they the existence of each strategic motives to be implicating. For example, some players think that the H-choosers in treatment 1 are much less likely to be motivated by altruism than the H-choosers in treatment 2, whereas others do not infer such a big difference. The within-person design of Experiment 2 allows us to ask whether differences in individual player B's altruism inferences across treatments are associated with changes in their responses.

Importantly, we find that using a within-person differencing approach eliminates the potential

[23] In treatment 3, they predicted an average of 65% of player A's who chose (H) to choose {400 tokens for player A, 300 tokens for another participant} over {500 tokens for player A, 0 tokens for another participant}. This prediction is marginally lower than the prediction in treatment 2 ($p = 0.061$), but significantly larger than the prediction in treatment 1 ($p = 0.001$).

[24] Even though player B's are correct about the nature of type selection each treatment induces, they are pessimistic and conservative in their beliefs about the generosity of player As in this question. Looking at player A's actual choices in part 1 on this question, we see that 71% of all player As, 72% of those who chose (H) in treatment 1, 86% of those who chose (H) in treatment 3 and 93% of those who chose (H) in treatment 2 chose {400 tokens for player A, 300 tokens for another participant} over {500 tokens for player A, 0 tokens for another participant} in part 1.

confounds arising from possible correlation of preferences and beliefs. Player Bs who are more altruistic have higher expectations of altruism given helpful behavior.[25] If we simply test whether individuals with high kindness inferences are more likely to reciprocate to helpful behavior, we would be confounding the causal impact of kindness inferences with their baseline willingness to help in the given sub-game. Instead, we take an approach that controls for differences across individuals in order to infer a meaningful relationship between kindness inferences and concern withdrawal. In particular, we relate changes in kindness inference to changes in reciprocal behavior. The identifying assumption that player Bs who are more altruistic do not have lower degrees of deterioration in kindness inference across treatments. Our data supports this assumption.[26]

Among the 55 player B's with response (R | H) in treatment 2				
player B choice	N	Altruism belief		Difference
		(treatment 2)	(treatment 1)	
(R | H) in treatment 1	34	76.7%	60.1%	16.6%
(N | H) in treatment 1	21	84.7%	52.8%	31.9%
		(treatment 2)	(treatment 3)	
(R | H) in treatment 3	39	79.8%	70.2%	9.6%
(N | H) in treatment 3	16	79.6%	57.7%	21.9%

Table 4: Within-person changes in kindness inferences and reciprocity towards H

Table 2.3 presents the altruism inferences of the fifty-five player Bs who reward H-choosers in treatment 2. The first two rows in Table 2.3 split these player Bs based on whether they also reward H-choosers in treatment 1. The first column reports the percentage of H-choosers each group believes is altruistic in treatment 2. The second column reports their average beliefs concerning the percentage of altruistic H-choosers in treatment 1, and the last column reports the difference. We want to see whether those who withdrew rewards in treatment 1 differ in the change in their altruism inferences from those who continue to reward H-choosers in treatment 1. Player Bs who rewarded action (H) in treatment 2 but stopped rewarding it in treatment 1 perceive a larger difference in the

[25] Looking at player Bs' choices in Part 1 and their kindness inferences in Part 4, we find a significant correlation between choosing (\$3.50, \$5.50) over (\$4, \$4) in Part 1, and beliefs concerning the percentage of altruistic H-choosers in treatment 1 ($p = 0.033$) and in treatment 2 ($p = 0.087$). Similarly, we find a significant correlation between choosing (\$4, \$4) over (\$4.50, \$2.50) in Part 1, and beliefs concerning the percentage of altruistic H-choosers in treatment 1 ($p = 0.037$), in treatment 2 ($p = 0.000$), and in treatment 3 ($p = 0.000$). Previous literature has taken different approaches to deal with endogeneity concerns arising from projection bias (Costa-Gomes et al., 2014, Bellemare, Kröger and van Soest, 2008, 2011a; Bellemare, Sebald and Strobel; 2011).

[26] We do not find any siginificant correlation between choosing (\$3.50, \$5.50) over (\$4, \$4) in Part 1, and the degree to which beliefs concerning the percentage of altruistic H-choosers decline between treatment 2 and treatment 1 ($p = 0.539$), or between treatment 2 and 3 ($p = 0.426$). Similarly, we do not find any siginificant correlation between choosing (\$4, \$4) over (\$4.50, \$2.50) in Part 1, and the degree to which beliefs concerning the percentage of altruistic H-choosers decline between treatment 2 and treatment 1 ($p = 0.138$), or between treatment 2 and 3 ($p = 0.942$).

altruism of helpful player As, compared to those who continue to reward action (H) in treatment 1 (Wilcoxon rank-sum (Mann-Whitney) test: $z = 2.31$, $p = 0.017$).[27] The bottom panel of Table 2.3 split the player Bs who rewarded H-choosers in treatment 2 based on whether they also reward H-choosers in treatment 3. Again, the player Bs who withdraw rewards for helpful behavior show a larger decrease in their altruism inferences regarding the H-choosers in treatment 3 (Wilcoxon rank-sum (Mann-Whitney) test: $z = -2.12$, $p = 0.034$).[28] In sum, the results show that a within-person increase (deterioration) of kindness inference about helpful player As from one treatment to another is associated with an increase (decrease) in player B's propensity to reward player A for being helpful.

In sum, the results from Experiment 2 show that second-movers positively reciprocate more in response to a helpful action when the strategic incentives for the first-mover to choose that action are weaker. Results also reveal a direct association between inferences of altruism and reciprocal choices. We see that altruism inferences regarding player As who chose to be helpful decreases with the strength of the strategic motivation the game form presents to be helpful. Importantly, individual heterogeneity in kindness inferences explains individual differences in how much withdrawal of concern player Bs display when strategic motives to be helpful are present in the game form.

Robustness Check and Additional Evidence We conducted Study 2A among 146 subjects who had not participated in Experiment 2 to check the sensitivity of our results to the within-person and multi-task design of Experiment 2 and the choice of p and q. Study 2A presented either treatment 1 or treatment 2 of the reciprocal interaction to pairs of participants in the role of player A and player B. By eliminating other payoff-relevant tasks, Study 2A aims to assess whether the fact that more than one risky payoff is simultaneously unrealized until the end of the experiment

[27] We can also compare the changes in the kindness inferences of player Bs who did not reward H-choosers in treatment 1 based on whether they rewarded them in treatment 2. Among the fifty-one player Bs who did not reward H-choosers in treatment 1, thirty of them also did not reward H-choosers in treatment 2. These player Bs on average reported a 12.1% decline in the composition of genuinely kind player As among H-choosers (reported average beliefs of 60.5% of genuinely kind player As among H-choosers in treatment 2 and average beliefs of 48.3% of genuinely kind player As among H-choosers in treatment 1). Compared to the twenty-one player Bs who chose to reward H-choosers in treatment 2 even though they did not reward them in treatment 1, the average inference deterioration of these twenty-six player Bs is significantly lower (Wilcoxon rank-sum (Mann-Whitney) test: $z = 3.39$, $p = 0.0001$).

[28] Using a complementary data analysis, we can also look at the subset of player Bs who did not reward H-choosers in treatment 3 and test whether within-person differences in inferences can predict which ones are likely to reward H-choosers in treatment 2. Among the forty-two player Bs who did not reward H-choosers in treatment 3, twenty-six of them also did not reward H-choosers in treatment 2. These player Bs did not see any difference in the composition of genuinely kind player As among H-choosers (reported average beliefs of 56.1% of genuinely kind player As among H-choosers in treatment 2 and average beliefs of 56.3% of genuinely kind player As among H-choosers in treatment 3). Compared to the sixteen player Bs who chose to reward H-choosers in treatment 2 even though they did not reward them in treatment 3, the average inference deterioration of these twenty-six player Bs is significantly lower (Wilcoxon rank-sum (Mann-Whitney) test: $z = -3.43$, $p = 0.001$).

could have driven behavioral differences across treatments 1 and 2 in Experiment 2 (Cox et al., 2015). Study 2A also increased the chances that the second-mover's response options elicited by the strategy method would be implemented, by setting $p = .8$ and $q = .1$ in treatment 1 and $p = .1$ and $q = .1$ in treatment 2. The outcomes were kept the same as in Experiment 2.

In Study 2A, the results show that a greater chance of the potential of of punisment was indeed motivating: 96% of player A's chose (H) in treatment 1 compared to 67% in treatment 2 (two-sided Pearson Chi-square test, $\chi^2 = 8.97$, $p = 0.003$). In support of the main hypothesis that positive reciprocity to the same helpful action is lower when the interaction features stronger incentives for taking that helpful action, a total of 15% of player B's chose to reward (H) in treatment 1 compared to 46% in treatment 2 (two-sided Pearson Chi-square test, $\chi^2 = 5.51$, $p = 0.019$).[29] Further details of the study are included in the Experimental Instructions Appendix.

Summary

Experiment 2 features three treatments: 1) treatment 1, where the first-stage player therefore expects the second-stage player almost always to have the option to punish a selfish action, 2) treatment 2, where the first-stage player therefore expects the second-stage player to be a passive recipient most the time, and 3) treatment 3, where the first-stage player therefore expects the second-stage player almost always to have the option to reward a helpful action. The existence of the potential to reward a helpful action in this probabilistic game helps us compare demand for positive reciprocity across treatment 1 and treatment 2. Consistent with the notion that perceived motives matter for kindness judgments, we find that the same helpful triggers more positive reciprocity in treatment 2. Moreover, Experiment 2 elicits beliefs regarding altruism of first-movers who are helpful, and presents evidence that suggests that reciprocity is higher when the first-mover is inferred to be genuinely altruistic. The positive reciprocity differences are replicated in an experiment that makes three design adjustments to Experiment 2: 1) set p and q to 0.10 rather 0.01, 2) present treatment 1 and treatment 2 of Game Γ_2 in a between-subject design, and 3) as the only payoff relevant game. Overall, the results provide strong support for the notion that positive reciprocity to the same helpful action is lower when the strategic incentives for taking that helpful action are stronger.

[29]The degree of positive reciprocity is lower than in Experiment 2, possibly due to higher probability of being motivated by strategic considerations in Study 2A in treatment 2 due to non-negligible chances of punishment or rewards (10% each).

3 Discussion and Conclusion

3.1 Relation of results to reciprocity models

The reciprocity literature has proposed three main determinants of reciprocal decision-making. Outcome-based models of altruism and reciprocity (Fehr and Schmidt, 1998; Bolton and Ockenfels, 2000) model the positive relationship between a helpful action and a reaction based only on preferences over payoff allocations. Because our experiments compare differences in reciprocal reactions to the same helpful action, and keep the payoff structure in the subgame of interest constant, the results cannot be rationalized by outcome-based models.

A second class of models emphasize people's desire to punish hostile intentions and reward kind intentions (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). In these models, beliefs regarding what the first-mover intended for the second-mover are central in evaluating the kindness of an action: the second-mover considers an action to be relatively kind if she believes that the first-mover intended the second-mover's material payoff to be larger than a fair benchmark as a result of his action.[30] As discussed in the context of our experiments and detailed in the Appendix, these models also fail to explain the patterns demonstrated by our results, because they are designed to capture the role of intentions, aptly operationalized as relating only to the consequence the first-mover expected his action to have on the second-mover.

In order to explain the evidence presented in this paper, reciprocity models should either allow kindness judgments to also vary with what the first-mover could have gained or lost if he acted differently, or consider inferences regarding the underlying altruism of the first-mover directly as a decision-making input for the second-mover. A third class of reciprocity models have such features (Cox, Friedman and Sadiraj 2008; Gül and Pesendorfer, forthcoming). Levine (1998) introduced the idea that a person's concern for another person's well-being increases in relation to how altruistic the other person is. Building on this idea, the Gül and Pesendorfer (2010) (GP) model predicts higher rewards for the same helpful action when the person taking the action is perceived to have a higher degree of altruism. In a similar spirit, but without having to rely on beliefs, the Cox et al. (2008) (CFS) model considers what the first-mover stood to gain or lose by choosing an action. In

[30]The second-mover's beliefs regarding what the first-mover intended for the second-mover are central to this definition of kindness. Belief-driven intention-based models (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010) therefore rely on psychological game theory (see Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009 for an overview). Cox et al. (2007) capture similar drivers of intention-based reciprocity without relying on beliefs. They model reciprocal preferences as a function of gratitude and resentment emotions that are driven by the alternative action space of the first-mover and the maximum payoff that the second-mover can guarantee for herself.

particular, this model predicts higher rewards for the same helpful action if it helps the benefactor more than it helps the person who took the action.

The CFS model proposes that (H) would be perceived as a more generous action than (S) if (i) the maximum payoff player B can get if player A chooses (H) is greater than or equal to the maximum payoff player B can get if player A chooses (S), and (ii) the maximum payoff player A can get by choosing (H) minus what he can get by choosing (S) is (at least weakly) less than the maximum payoff player B can get if player A chooses (H) minus what she can get if player A chooses (S). In other words, (H) is more generous than (S) if it can help player B, and if it can help player B more than it can help player A.[31] This model is not immediately applicable to comparing the perceived generosity of (H) across the two treatments, because in each treatment, the payoffs (at least weakly) satisfy both (i) and (ii) and thus (H) is considered to be more generous than (S). Therefore, a strict interpretation of this model would not produce any differences in the degree of positive reciprocity in either experiment. However, because this model considers what player A can obtain, it can be particularly suitable for thinking about the role of motives. Let us imagine a simple extension that defines the generosity differential between (H) and (S) as the difference between how much choosing (H) over (S) helps player B minus how much it helps player A. Across the two treatments in Experiment 1, the payoffs of player B are fixed and the payoffs of player A are the same if player A chooses (H). In treatment 2, the maximum payoff player A can get if he chooses (S) is \$6.50 and in treatment 1, it is only \$4. Therefore, the extended Cox et al. (2008) model would predict that choosing (H) rather than (S) in treatment 2 looks more generous than choosing (H) rather than (S) in treatment 1, thus capturing the differences in positive reciprocity we document in Experiment 1. However, the extended CFS model cannot rationalize the results of Experiment 2. The maximum payoff player A can get if he chooses (H) in treatments 1 and 2 are both equal to \$4. Since choosing (H) over (S) also helps player B by the same amount in treatments 1 and 2, choosing (H) leads to the same generosity differential and thus reveals the same degree of generosity, and should be rewarded equally in treatments 1 and 2. This is not in line with the the prediction and findings of this paper.

Given the close relationship between perceptions of motives and perceptions of the altruism of the person taking the helpful action, our results can potentially be explained by the GP model. In

[31] We hold the default option across all treatments the same for player A, therefore the treatments do not present any differences in whether choosing (H) over (S) can be considered an omission or a commission. Thus, any reciprocity differences across treatments in these experiments can only be driven by the perceived generosity of choosing (H), as presented in Axiom R of Cox, Friedman, Sadiraj (2008).

Experiment 1, we do not have direct evidence of the second-mover's altruism inferences regarding the first-movers. However, presuming that the second-movers' altruism inferences regarding helpful first-movers are more positive in treatment 2 than in treatment 1 seems plausible for two reasons. First, second-movers expect a higher fraction of strategically motivated first-movers among those who are helpful in treatment 1. Second, Study 1A elicits altruism inferences regarding the first-movers and confirms that third-parties expect on averagelower altruism among the first-movers who were helpful in treatment 1 than in treatment 2. Therefore, we find it highly plausible that the evidence in Experiment 1 is consistent with the GP model. Importantly, in Experiment 2, we elicit direct evidence of altruism inferences. The data suggest that within-person changes in reciprocal reactions to the same helpful action are associated with within-person changes in the inferences regarding the the inherent altruism of a helpful first-mover. This result provides direct support for the reciprocity mechanism proposed in the GP model.

3.2 Summary, Implications and Future Directions

Making a relational investment often brings benefits in the future, since rewards or punishments are inherent to many professional and personal reciprocal relationships. These incentives are shown to motivate socially desirable, helpful actions (Andreoni et al., 2003) and can result in large efficiency gains by enforcing these actions (Fehr et al. 1997). By virtue of being successful, however, the existence of such incentives obscures the motives of people who act generously in these interactions.

This paper presents data from two experiments designed to isolate the role of perceived motives on reciprocal behavior. Evidence from both a between-subjects and a within-subject design show that positive reciprocity declines in the perceptions regarding the degree to which a helpful action is strategically motivated. These results suggest that people are quite sophisticated about others' mental models and contemplate their motives when deciding on the appropriate reciprocal response.

The finding that perceived motives play an important role in shaping reciprocal decisions paves the way for several future research directions. Our findings shed some light on the type of reciprocity model that can incorporate the role of perceived motives. The results suggest at least two directions. One possibility is to allow perceptions of what the first-mover expected to gain or lose as a result of his action to influence the perceived kindness of an action (Cox et al., 2008). Another possibility is to model reciprocity as a response to the revealed altruism of the first-mover, taking care to specify the equilibrium properties under which the first-mover's actions are informative regarding his altruism (Gül and Pesendorfer, 2010). Recent work has proposed models exploring reciprocal

behavior in gift-exchange games where i) individuals care about others to the extent that others are altruistic, and ii) altruism is private information (Arbak and Kranich, 2005; Dur, 2009; Non, 2012). We hope that the experimental design and data presented in this paper are useful for spurring an interest in future work in this area that considers sensitivity to outcomes, intentions and motives in explaining reciprocal decision-making.

The central hypothesis tested in this paper is related to a broader question that has been pivotal in the recent research on reciprocity: how to evaluate kindness. This question is important to answer across many domains that involve reciprocal considerations. It is our hope that the experiments and results presented in this article add to this discourse. In recent work, Celen, Blanco and Schotter (2014) offer a definition of kindness based on a notion of blame, similar to the notion of relative kindness of players in the GP model. Future experimental research can further this inquiry by testing different notions of kindness, and the relative importance of the consequences, intentions and motives on perceptions of kindness. In addition, it would be valuable to explore how perceived motives shape reciprocity towards a hurtful action.

The results presented in this paper may also have implications for the so-called positive reciprocity puzzle. There is an emerging consensus that the propensity to punish harmful behavior is stronger than the propensity to reward friendly behavior (for example, Fehr and Gächter, 2000; Cox and Deck, 2005; Charness and Rabin, 2002, 2005; Offerman, 2002). Offerman (2002) showed that subjects are 67% more likely to reciprocate to an intentional hurtful choice over an unintentional hurtful choice, but they are only 25% more likely to reciprocate to an intentional helpful choice over an unintentional helpful choice. Al-Ubaydli and Lee (2009) elicited second-order expectations and incorporated them into the Falk and Fischbacher (2006) model in order to tease out whether this asymmetry is a result of asymmetric intrinsic tendencies to reward or punish, or asymmetries in the extent to which rewards and punishments are objectively merited due to the differences in the perceived kindness of the first-mover, given the game form that Offerman (2002) used. In light of the evidence presented in this paper, the reader may also wonder whether the role of perceived motives contribute to this asymmetry. In the case of intentional hurtful actions in a reciprocal context, the motives of the first-mover are unambiguously unkind and therefore deserve retribution. However, the motives behind intentional helpful actions in a reciprocal context can be ambigous, as demonstrated by the examples in this paper. Therefore, the positive reciprocal response may not be as strong as it would have been if the helpful action were unambiguously driven by kindness. A closer look at the asymmetry between positive and negative reciprocity that disentangles these

possible explanations would be worthwhile.

Finally, the results urge us to deliberate on seemingly contradictory predictions stemming from the literature on guilt aversion (see Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007 for an overview of guilt aversion theory, and Al-Ubaydli and Lee, 2009 for a more specific discussion regarding this potential contradiction.). Consider the investment-game where the first-mover makes a risky investment by trusting the second-mover to reciprocate. The guilt aversion literature would predict that the higher the second-order expectations are of the second-mover regarding what the first-mover expected of her, the more likely she is to reciprocate. If we think that the likelihood of the first-mover being motivated by altruism is lower if his expectations of the second-mover are higher, we may conclude that the guilt aversion literature predicts the second-mover to reciprocate more positively towards the first-movers who are more strategically motivated. However, altruistic first-movers need not have lower expectations of the second-movers. In fact, altruistic players may expect more helpful behavior in general (Costa-Gomes et al., 2014, Bellemare, Kröger and van Soest, 2008, 2011a; Bellemare, Sebald and Strobel; 2011). Therefore, future research needs to isolate the second-mover's perceptions about the motives of the first-mover from her perceptions regarding his expectations from her.

REFERENCES

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner. 2000. “The moonlighting game. An experimental study on reciprocity and retribution.” Journal of Economic Behavior & Organization, 42, 265-277.

Al-Ubaydli, Omar, and Min Sok Lee. 2012. "Do you reward and punish in the way you think others expect you to?" The Journal of Socio-Economics 41.3: 336-343.

Al-Ubaydli, Omar and Min Sok Lee. 2009. “An experimental study of asymmetric reciprocity.” Journal of Economic Behavior & Organization, 72, 738-749.

Andreoni, James and John Miller. 2002. "Giving according to GARP: An experimental test of the consistency of preferences for altruism." Econometrica, 70, 737-753.

Andreoni, James, William Harbaugh, and Lise Vesterlund. 2003. "The carrot or the stick: Rewards, punishments, and cooperation." American Economic Review, 893-902.

Arbak, Emrah and Laurance Kranich. 2005. “Can Wages Signal Kindness?” Working paper, Groupe d’Analyse et de Theorie Economique, University of Lyon.

Armantier, Olivier, and Nicolas Treich. 2013. "Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging." European Economic Review 62: 17-40.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2007. "Guilt in games." The American Economic Review, 170-176.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2009. "Dynamic psychological games." Journal of Economic Theory, 144.1, 1-35.

(Bellemare, Kröger and van Soest, 2008, 2011a; Bellemare, Sebald and Strobel; 2011).

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, reciprocity, and social history." Games and Economic Behavior, 10, 122-142.

Blount, Sally. 1995. "When social outcomes aren´t fair: The effect of causal attributions on preferences." Organizational Behavior and Human Decision Processes, 63, 131-144.

Bolton, Gary E. and Axel Ockenfels. 1998. “Strategy and equity: An ERC-Analysis of the Güth-van Damme Game.” Journal of Mathematical Psychology, 42, 215-226.

Bolton, Gary E. and Axel Ockenfels. 2000. "ERC: A theory of equity, reciprocity, and competition." American Economic Review, 166-193.

Bolton, Gary E., Jordi Brandts, and Axel Ockenfels. 1998. “Measuring motivations for the reciprocal responses observed in a simple dilemma game.” Experimental Economics, 1, 207-219.

Brandts, Jordi and Carles Solà. 2001. “Reference points and negative reciprocity in simple sequential games.” Games and Economic Behavior, 36, 138-157.

Brandts, Jordi, and Gary Charness. 2011. "The strategy versus the direct-response method: a first survey of experimental comparisons." Experimental Economics 14.3: 375-398.

Cabral, L., Ozbay, E. Y., & Schotter, A. 2014. “Intrinsic and instrumental reciprocity: An experimental study.” Games and Economic Behavior, 87, 100-121.

Celen, Bogachan, Mariana Blanco and Andrew Schotter. 2014. “On blame and reciprocity: An experimental study.” Working paper.

Charness, Gary. 2004. "Attribution and reciprocity in an experimental labor market." Journal of Labor Economics, 22, 665-688.

Charness and Dufwenberg 2006. "Promises and partnership." Econometrica 74.6: 1579-1601.

Charness, Gary and David I. Levine. 2007. “Intention and stochastic outcomes: An experimental study.” The Economic Journal, 117, 1051-1072.

Charness, Gary and Ernan Haruvy. 2002. “Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach.” Games and Economic Behavior, 40, 203–231.

Charness, G., Gneezy, U., & Kuhn, M. A. 2012. “Experimental methods: Between-subject and within-subject design.” Journal of Economic Behavior & Organization, 81(1), 1-8.

Charness, Gary and Matthew Rabin. 2002. "Understanding social preferences with simple tests." Quarterly Journal of Economics, 817-869.

Charness, G., & Rabin, M. 2005. “Expressed preferences and behavior in experimental games.” Games and Economic Behavior, 53(2), 151-169.

Costa-Gomes, Miguel and Georg Weizsäcker. 2008. “Stated Beliefs and Play in Normal-Form Games,” Review of Economic Studies, 75, 729-762.

(Costa-Gomes et al., 2014),

Cox, James C. 2004. “How to identify trust and reciprocity.” Games and Economic Behavior, 46, 260-281.

Cox, James C. and Cary Deck. 2005. “On the Nature of Reciprocal Motives.” Economic Inquiry, Volume 43, Issue 3, pages 623–635, July 2005

Cox, James C., Daniel Friedman, and Steven Gjerstad. 2007 "A tractable model of reciprocity and fairness." Games and Economic Behavior 59.1: 17-45.

Cox, James C., Daniel Friedman, and Vjollca Sadiraj. 2008. "Revealed Altruism." Econometrica, 76, 31-69.

Cox, James C., Klarita Sadiraj, and Vjollca Sadiraj. 2008. "Implications of trust, fear, and reciprocity for modeling economic behavior." Experimental Economics, 11, 1-24.

Cox, James C., Vjollca Sadiraj, and Ulrich Schmidt. 2015. "Paradoxes and mechanisms for choice under risk." Experimental Economics 18.2: 215-250.

Dreber, Anna, Drew Fudenberg, and David G. Rand. 2014 "Who cooperates in repeated games: The role of altruism, inequity aversion, and demographics." Journal of Economic Behavior & Organization 98: 41-55.

Dufwenberg, Martin, and Uri Gneezy. 2000. "Measuring beliefs in an experimental lost wallet game." Games and Economic Behavior 30.2: 163-182.

Dufwenberg, Martin and Georg Kirchsteiger. 2004. "A theory of sequential reciprocity." Games and Economic Behavior, 47, 268-298.

Dur, Robert. 2009. "Gift Exchange in the Workplace: Money or Attention?" Journal of the European Economic Association. 7 (2-3): 550-560.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2009. "Homo Reciprocans: Survey Evidence on Behavioral Outcomes." Economic Journal, 119, 592–612.

Englmaier and Leider, 2010

Falk, Armin and Urs Fischbacher. 2006. "A theory of reciprocity." Games and Economic Behavior, 54, 293-315.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2008. "Testing theories of fairness—Intentions matter." Games and Economic Behavior, 62, 287-303.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. 1997. "Reciprocity as a contract enforcement device: Experimental evidence." Econometrica, Vol. 65, No. 4. p. 833-860.

Fehr, Ernst and Klaus M. Schmidt. 1998. "A theory of fairness, competition, and cooperation." Quarterly Journal of Economics, 817-868.

Fehr, Ernst and Simon Gächter. 2000. "Fairness and retaliation: The economics of reciprocity." The Journal of Economic Perspectives, 159-181.

Fischbacher, Urs. 2007. "z-Tree: Zurich toolbox for ready-made economic experiments." Experimental Economics, 10, 171-178.

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

Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological games and sequential rationality." Games and Economic Behavior 1.1: 60-79.

Gneezy, U., Güth, W., & Verboven, F. 2000. “Presents or investments? An experimental analysis.” Journal of Economic Psychology, 21(5), 481-493.

Gül, Faruk, and Wolfgang Pesendorfer. "Interdependent preference models as a theory of intentions." Conditionally accepted by: Journal of Economic Theory (2010).

Güth, Werner and Eric Van Damme. 1998. “Information, strategic behavior, and fairness in ultimatum bargaining: An experimental study.” Journal of Mathematical Psychology, 42, 227-247.

Heider, F. 1958. “The psychology of interpersonal relations.” Wiley, New York. Kelley, Harold H. 1967. "Attribution theory in social psychology." Nebraska symposium on motivation. University of Nebraska Press.

Huck, Steffen and Georg Weiszsäcker. 2002 “Do players correctly estimate what others do? Evidence of conservatism in beliefs.” Journal of Economic Behavior & Organization, Vol 47, 71-85.

Kelley, Harold H. 1973. "The processes of causal attribution." American psychologist 28.2.

Klempt, Charlotte. 2012 "Fairness, spite, and intentions: Testing different motives behind punishment in a prisoners' dilemma game." Economics Letters, 116/3: 429-431.

Levine, David K. 1998. "Modeling altruism and spitefulness in experiments." Review of Economic Dynamics, 1, 593-622.

McCabe, Kevin. A., Mary L. Rigdon, and Vernon L. Smith. 2003. “Positive reciprocity and intentions in trust games.” Journal of Economic Behavior & Organization, 52, 267-275.

Nelson Jr, William Robert. 2002. “Equity or intention: it is the thought that counts.” Journal of Economic Behavior & Organization, 48, 423-430.

Netzer, Nick, and Armin Schmutzler. 2014. "Explaining gift-exchange–the limits of good intentions." Journal of the European Economic Association, 12(6), 1586-1616.

Non, Arjan. 2012. “Gift-exchange, incentives, and heterogeneous workers.” Games and Economic Behavior, 75, 319-336.

Offerman, Theo. 2002. "Hurting hurts more than helping helps." European Economic Review, 46, 1423-1437.

Rabin, Matthew. 1993. "Incorporating fairness into game theory and economics." The American Economic Review, 1281-1302.

Rabin, Matthew. 1998. "Psychology and economics." Journal of economic literature, Vol XXXVI, March, 11-46.

Rand, David G., Drew Fudenberg, and Anna Dreber. 2013. “It's the thought that counts: The role of intentions in reciprocal altruism.” Working paper.

Ross, Michael, and Garth JO Fletcher. 1985. "Attribution and social perception." The handbook of social psychology 2: 73-114.

Rutström, E.E., Wilcox, N.T., 2009. "Stated beliefs versus inferred beliefs: a methodological inquiry and experimental test." Games and Economic Behavior, 67, 616–632.

Schotter, Andrew, and Isabel Trevino. 2014 "Belief elicitation in the laboratory." Annu. Rev. Econ. 6.1: 103-128.

Sebald, Alexander. 2010. "Attribution and reciprocity." Games and Economic Behavior 68.1: 339-352.

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

Segal, U., & Sobel, J. 2008. "A characterization of intrinsic reciprocity." International Journal of Game Theory, 36(3-4), 571-585.

(Smith, 2013)

Sobel, J. 2005. "Interdependent preferences and reciprocity." Journal of Economic Literature, 392-43

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

Strassmair, Christina. 2009. "Can intentions spoil the kindness of a gift? An experimental sudy." Working paper, University of Munich, Munich Discussion Paper No. 2009-4.

Toussaert, Séverine. 2014. "Intention-Based Reciprocity and Signalling of Intentions" Working paper, NYU.

Trautmann, Stefan T., and Gijs Kuilen. 2015 "Belief elicitation: A horse race among truth serums." The Economic Journal: 2116-2135.

Appendix

Predictions

We discuss the predictions of the intention-based reciprocity models proposed by Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) regarding Experiment 1 and Experiment 2. We simplify the discussion by considering slightly modified versions of the treatments in Experiment 2 by assuming that $p = 1$ in treatment 1, $p = q = 0$ in treatment 2, and $q = 1$ in treatment 3. This simplification greatly aids discussion without impacting the differences in the predictions of different theories.

For generality, we parameterize the payoffs in Game Γ_1 and Game Γ_2 to highlight the general features that allow us isolate the role of motives. Figure 3 below refers to the generalized version of Game Γ_1. The variable m is varied across treatments. In treatment 1, $m = x - 5k$, and in treatment 2 $m = x + 5k$. Note that all of the payoffs of player A are (at least weakly) larger than the payoffs of player B ($x > y + 2t$, $x > y + 4k$), and we further restrict t and k such that all payoffs are positive ($y > k$, $x > 5k$, $5k > t$).

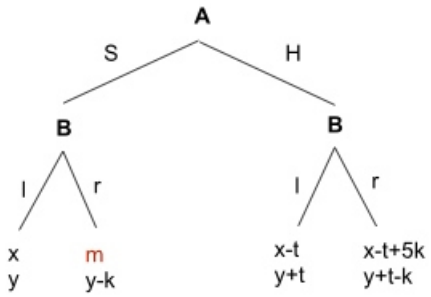


Figure 4: Generalized Game Γ_1

Figure 4 below refers to the generalized version of Game Γ_2. Note that all of the payoffs of player A are (at least weakly) larger than the payoffs of player B ($x > y$ and $x - y \geq 2t$), and we further restrict t and k such that all payoffs are positive. For simplicity of discussion, we also set $t = k$.

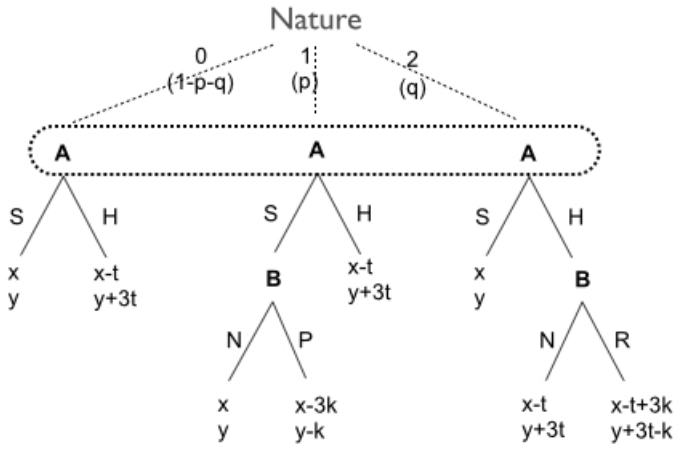


Figure 5: Generalized Game Γ_2

Dufwenberg and Kirchsteiger (2004)

The model respectively defines the perceived kindness of (H) and (S) from the perspective of B as $\kappa_B(H) = E_{BA}[\pi_B|H] - \frac{1}{2}\{E_{BA}[\pi_B|H] + E_{BA}[\pi_B|S]\}$ and $\kappa_B(S) = E_{BA}[\pi_B|S] - \frac{1}{2}\{E_{BA}[\pi_B|H] + E_{BA}[\pi_B|S]\}$, where $E_{BA}[\pi_B|S]$ denotes player B's beliefs regarding player A's expectations of player B's payoffs (π_B) if player A chooses (S).[32] The model posits that the degree of positive reciprocation to H increases in $\kappa_B(H)$ in the region where $\kappa_B(H) \geq 0$ and the degree of negative reciprocation to S increases in $|\kappa_B(S)|$ in the region where $\kappa_B(S) < 0$. The hypotheses in this paper are centered around the perceived kindness of (H), which depends on player B's second-order beliefs: what player B believes about what player A thinks player B will choose if player A chooses (H).

Experiment 1. Denote player B's second-order beliefs regarding the prevalence of (r) given action (H) as $b''_{t_1}(r|H)$ and $b''_{t_2}(r|H)$ in the treatments 1 and 2 respectively. Similarly, denote player B's second-order beliefs regarding the prevalence of (r) given action (S) as $b''_{t_1}(r|S)$ and $b''_{t_2}(r|S)$ in the two conditions.

Then, in treatment 2, $E_{BA,t_2}[\pi_B|H] = b''_{t_2}(r|H)(y+t-k)+(1-b''_{t_2}(r|H))(y+t)$ and in treatment 1, $E_{BA,t_1}[\pi_B|H] = b''_{t_1}(r|H)(y+t-k)+(1-b''_{t_1}(r|H))(y+t)$. And the perceived kindness of (H) across two treatments are $\kappa_{B,t_2}(H) = \frac{1}{2}\{b''_{t_2}(r|H)(y+t-k)+(1-b''_{t_2}(r|H))(y+t)\} - \{b''_{t_2}(r|S)(y-k)+(1-b''_{t_2}(r|S))(y)\}$ and $\kappa_{B,t_1}(H) = \frac{1}{2}\{b''_{t_1}(r|H)(y+t-k)+(1-b''_{t_1}(r|H))(y+t)\} - \{b''_{t_1}(r|S)(y-$

[32]Both (H) and (S) are in the efficient set of actions for player A and are the only actions player A can take.

$k) + (1 - b''_{t_1}(r|S))(y)\}$. Note that since the payoffs of player B are exactly the same across the two treatments, any differences in perceived kindness of (H) will stem from differences in second-order expectations. In order for the Dufwenberg and Kirchsteiger (2004) model to predict a higher degree of positive reciprocity to (H) in treatment 2, we either need to maintain $b''_{t_2}(r|H) < b''_{t_1}(r|H)$, which is inconsistent with the predicted behavior itself, or assume $b''_{t_2}(r|S) > b''_{t_1}(r|S)$ which contradicts the behavior and expectations in the data. Therefore, the Dufwenberg and Kirchsteiger (2004) model cannot explain the data from Experiment 1.

Experiment 2. In treatment 2, player B gets y if player A chooses (S) and she gets $y + 3t$ if he chooses (H). The perceived kindness of choosing (H) in treatment 2 can be calculated by comparing $E_{BA,t_2}[\pi_B|S] = y + 3t$ to the midpoint of possible outcomes, which is $\frac{1}{2}\{E_{BA,t_2}[\pi_B|H] + E_{BA,t_2}[\pi_B|S]\} = y + 1.5t$. Therefore, the preceived kindness of (H) in treatment 2 is $\kappa_{B,t_2}(H) = 1.5t$.

In treatment 3, player B's beliefs about player A's expectations regarding player B's payoffs if player A chooses H are given by $E_{BA,t_3}[\pi_B|H] = b''(R|H)(y + 3t - k) + (1 - b''(R|H))(y + 3t)$ and player B's beliefs about player A's expectations regarding player B's payoffs if player A chooses S ($E_{BA,t_3}[\pi_B|S]$) are simply y. If $b''(R|H) = 0$, then the preceived kindness of (H) is the same in treatment 2 and 3. If, $b''(R|H) > 0$, then the preceived kindness of (H) is strictly lower in treatment 3 than in treatment 2, since $\kappa_{B,t_3}(H) = 1.5t - 0.5b''(R|H)k$. Therefore the Dufwenberg and Kirchsteiger (2004) model would predict (at least weakly) higher level of positive reciprocity in treatment 2 than in treatment 3. This prediction is in line with the prediction in this paper and the results.

However, the model would produces a contradictory prediction of this paper in comparing the degree of positive reciprocity in treatment 1 compared to treatment 2. In treatment 1, player B's beliefs about player A's expectations regarding player B's payoffs if player A chooses S are given by $E_{BA,t_1}[\pi_B|S] = b''(P|S)(y-k)+(1-b''(P|S))(y)$ and player B's beliefs about player A's expectations regarding player B's payoffs if player A chooses H are simply $y + 3t$. Therefore, perceived kindness of (H) in this treatment is $\kappa_{B,t_1}(H) = 1.5t + 0.5b''(P|S)k$. If $b''(P|S) = 0$, then the perceived kindness of (H) is the same in treatments 1 and 2. If, $b''(P|S) > 0$, then perceived kindenss of (H) is higher in treatment 1 than in treatment 2, since choosing (S) may lead to player B sacrificing an amount k to punish player A in treatment 1. Therefore the modified Dufwenberg and Kirchsteiger (2004) model would predict (at least weakly) lower level of positive reciprocity in treatment 2 than

in treatment 1.

Falk and Fishbacher (2006)

In Game Γ_1 and Game Γ_2, we keep most of the features that would impact the degree of intentionality in the Falk and Fishbacher (2006) model constant: Player A has the same choice set (S, H) and full control over his actions across all treatments. Having fixed these dimensions, we can investigate how perceived kindness of player A's actions differ across treatments. The model respectively defines the perceived kindness of (H) and (S) from the perspective of B as $\kappa_B(H) = E_{BA}[\pi_B|H] - E_{BA}[\pi_A|H]$ and $\kappa_B(S) = E_{BA}[\pi_B|S] - E_{BA}[\pi_A|S]$, where $E_{BA}[\pi_B|S]$ denotes player B's beliefs regarding player A's expectations of player B's payoffs (π_B) if player A chooses (S) and $E_{BA}[\pi_B|S]$ denotes player B's beliefs regarding player A's expectations of player A's payoffs (π_A) if player A chooses (S). Therefore the Falk and Fischbacher (2006) model determines the perceived kindness of an action based the difference between player B's beliefs about player A's intended outcome for player B versus player A's intended outcome for himself.

Experiment 1. In treatment 1 where $m = x - 5k$, the relative outcome kindness of (S) is $\kappa_{B,t_1}(S) = b''_{t_1}(r|S)[(y-k) - (x-5k)] + (1 - b''_{t_1}(r|S))[y-x]$, and the relative outcome kindness of (H) is $\kappa_{B,t_1}(H) = b''_{t_1}(r|H)[(y+t-k) - (x-t+5k)] + (1 - b''_{t_1}(r|H))[(y+t) - (x-t)]$. In treatment 2, where $m = x + 5k$, the relative outcome kindness of (S) is $\kappa_{B,t_2}(S) = b''_{t_2}(r|S)[(y-k) - (x+5k)] + (1 - b''_{t_2}(r|S))[y-x]$, and the relative outcome kindness of (H) is $\kappa_{B,t_2}(H) = b''_{t_2}(r|H)[(y+t-k) - (x-t+5k)] + (1 - b''_{t_2}(r|H))[(y+t) - (x-t)]$.

If $b''_{t_2}(r|H) = b''_{t_1}(r|H)$, action (H) looks equally unkind in both treatments, as the payoffs are the same for this sub-game across the treatments and player B earns less than player A. If second order expectations are in line with predicted equlibrium play (and the SOE we elicit in the data), then $b''_{t_2}(r|H) > b''_{t_1}(r|H)$. Interestingly, according to the Falk and Fischbacher (2006) model this would imply that choosing (H) in treatment 1 is less unkind ($\kappa_{B,t_1}(H) > \kappa_{B,t_2}(H)$). This prediction would not be able to rationalize a higher degree of positive reciprocity in response to (H) in treatment 2.

Experiment 2. In treatment 2, since player B has no choice, this considerations is reduced to calculating the difference between player B's payoffs and player A's payoffs as a result of player A's choices. If player A chooses H, the distance between player B's payoff from player A's payoff is $\kappa_{B,t_2}(H) = y - x + 4t$.

In the other two treatments, the second order beliefs of player B matter. Let's denote player B's beliefs regarding player A's expectations of player B choosing R in response to H in treatment 3 as $b''(R|H)$. Similarly, let's denote player B's beliefs regarding player A's expectations of player B choosing P in response to S in treatment 1 as $b''(P|S)$. Then, in treatment 3, the relative outcome kindness of (H) is $\kappa_{B,t_3}(H) = b''(R|H)[(y+3t-k)-(x-t+3k)] + (1-b''(R|H))[(y+3t)-(x-t)]$. If $b''(R|H) = 0$, action (H) looks equally unkind in treatment 3 as it does in treatment 2. However, if $b''(R|H) > 0$, then action (H) looks more unkind in treatment 3 since it leads to a larger disadvantaged payoff for player B compared to the action (H) in treatment 2 (by an amount of $4k \cdot b''(R|H)$). However, action (H) looks equally unkind in treatment 1 compared to the action (H) in treatment 2, because $\kappa_{B,t_1}(H) = y - x + 4t$. Therefore the behavioral differences between treatments 1 and 2 cannot be captured by the Falk and Fishbacher (2006) model either.

Ancillary Results

Experiment 1

The first column of Table 5 displays the number and percentage of player As choosing the option that gives them the higher payoff (option 1) in the modified dictator games presented in part 1 of Experiment 1. The second and third columns respectively report player As' and player Bs' average beliefs regarding the proportion of player As choosing option 1 in the four modified dictator games presented in part 2. In line with early findings of Charness and Rabin (2002), dictators are more likely to sacrifice their own payoffs to help another person when their payoffs are higher than the other person, and when the sacrifice produces a larger gain on the part of the other person. Beliefs reflect an understanding of these preferences, as they follow the ordering of choice proportions. However, subjects seem to be averse to reporting beliefs close to the extremes (0% or 100%), thus displaying a slight conservatism bias - distortion towards the uniform prior - as documented in previous work (Huck and Weizsäcker, 2002).[33]

[33]We use a quadratic incentive scheme for eliciting beliefs. Huck and Weizsäcker (2002) show that the conservatism bias is smaller under this scheme compared to eliciting beliefs using the Becker-De Groot-Marshak bidding mechanism.

Table 5: Behavior and Beliefs regarding Behavior in Modified Dictator Games in Experiment 1

Choice Question	N		Player A's	Player B's
(Option 1) vs. (Option 2)		Option 1 Choice	Option 1 Beliefs	Option 1 Beliefs
		(1)	(2)	(3)
($4.50, $1.50) vs. ($4.00, $4.00)	129	37 (29%)		
($2.50, $0) vs. ($2.00, $1.50)	129	37 (29%)	43%	37%
($4.00, $1.00) vs. ($3.00, $2.00)	129	92 (71%)	60%	54%
($5.00, $2.00) vs. ($4.00, $4.00)	129	71 (55%)		
($1.00, $4.00) vs. ($0.50, $6.50)	129	89 (69%)	71%	77%
($2.00, $3.00) vs. ($1.50, $5.50)	129	95 (74%)	70%	77%

Experiment 2

Table 6 summarizes the choices of player As and player Bs in the dictator games presented in part 1 and the beliefs regarding behavior in these games as elicited in part 2 of Experiment 2. Again, we see that dictators are more likely to sacrifice their own payoff to help the other participant when their payoffs are larger than that of the other person's and when the sacrifice leads to a larger gain. Also, beliefs are in line with observed behavior, yet still too conservative.

Table 6: Behavior and Beliefs regarding Behavior in Modified Dictator Games

Choice Question	N	Option 1	Option 1	Option 1 Beliefs	Option 1 Beliefs
(Option 1) vs. (Option 2)		Player A	Player B	Player A	Player B
(800, 800) vs. (700, 1100)	88	70%	80%	75%	76%
(800, 200) vs. (600, 400)	88	60%	49%		65%
(900, 500) vs. (800, 800)	88	44%	41%		46%
(500, 900) vs. (400, 1200)	88	69%	73%	78%	
(500, 0) vs. (400, 300)	88	29%	20%	45%	44%
(900, 0) vs. (800, 200)	88	31%	27%		
(400, 600) vs. (300, 1100)	88	69%	72%	66%	
(500, 900) vs. (400, 600)	88	81%	79%		