

Motivations for indirect reciprocity: Good deeds or good people?

Kieran Gibson
Lionel Page
Vera L. te Velde*

October 24, 2025

Abstract

We investigate the leading motivations for indirect reciprocity by experimentally studying the role of type-based, intentions-based, and outcome-based preferences. We design a novel experimental setting in which participants can help other participants under varying probability of being seen by third parties. Those third-party observers can then reward participants based on both their history of helpful behavior and the observability of that behavior. Because good deeds done in public are often strategically motivated to build reputation, while only truly altruistic people help in private, we can identify whether indirect reciprocity is directed towards good deeds or good people. We find that indirect reciprocity towards an individual is influenced by their previous attempts to be helpful and their impact on others, but is surprisingly unaffected by the authenticity of past benevolent actions. That is, good deeds are rewarded regardless of whether these actions reflect true altruism or are likely a strategic play to encourage future reciprocal kindness. We discuss implications for theories of indirect reciprocity and sustained cooperation in large groups.

JEL classification: D90; C72; C91; D71.

Keywords: indirect reciprocity, altruism, intentions, signaling, guile

*All authors affiliated with the University of Queensland School of Economics. Gibson: k.gibson@uq.edu.au. Page: l.page@uq.edu.au. te Velde: v.tevelde@uq.edu.au. We thank Priscilla Man, Dorothea Kübler, Alessandra Cas-sar, Nick Netzer, Nicola Pavoni, Kenan Kalayci, David Smerdon, Zachary Breig, Antonio Rosato, Jamie Cross, Andy McLennan, Ben Grodeck, Elif Incekara-Hafalir, and several seminar audiences for helpful conversations. Experiments reported were conducted with University of Queensland Institutional Human Research Ethics Approval number 2020001732.

1 Introduction

The psychological motivations behind acts of kindness are multifaceted, ranging from genuine selflessness to covertly self-serving tactics. At one end of the spectrum, truly altruistic acts stem from empathy and a genuine concern for others. At the other end, kind actions are underpinned by a calculated self-interest, aimed at building a favorable reputation that could be beneficial in the future.¹ Distinguishing between these contrasting motivations seems particularly critical in the context of *indirect reciprocity*, the phenomenon in which individuals tend to reward those who have demonstrated kindness previously. Despite extensive research showing that indirect reciprocity can sustain cooperation in large groups (Mailath and Samuelson, 2006; Roberts et al., 2021), the criteria by which people judge acts of kindness, and choose how to reciprocate, is however not well understood. In particular, it is not clear how acts of kindness that are strategically motivated (driven by their associated reputational benefits) are assessed.

In this paper, we address this essential question: do we reward all good deeds, or do we only reward (genuinely) good people who perform them? To answer this, we developed a new experimental framework that allows us to distinguish between several drivers of indirect reciprocity. In our games, a participant, referred to as the Observer, has the opportunity to witness the decisions made by another participant, the Agent, to help other participants, the Recipients. The Observer then decides whether to reciprocally reward the Agent. By varying the information that the Observer has about the Agent and their actions, we can identify how indirect reciprocity depends on three potential factors: the outcomes of the Agent’s previous actions (either positive or negative), the actions themselves (kind or unkind in intention, regardless of the actual outcomes), or the motives behind those actions (truly selfless or strategically self-serving).

In our first game, we focus on disentangling the effects of the Agent’s actions from their underlying motivations. We introduce a variable probability p , known to both the Observer and Agent, that determines how likely it is that the Agent’s actions are observed. By manipulating this probability to be either high or low, we influence the Agent’s incentive to perform kind acts for reputation gains. Actions performed under low observability are more indicative of genuine altruism, while those under high observability are more likely to be driven by the desire to appear helpful and gain from the Observer’s indirect reciprocity later. Since the Observer knows the probability p , and are therefore aware of whether the Agent expected their actions to be seen, the Observer can infer how genuinely altruistic the Agent is likely to be. In this way, we can examine whether the Ob-

¹This is backed by studies indicating that participants tend to be almost twice as helpful when their actions are noticeable by a potential returner of favor. See, for example, Engelmann and Fischbacher (2009), and the results of this study.

server's reciprocity is guided more by the perceived motives of the Agent or simply by the actions themselves.

In our second game, we introduce random variation in the outcomes of the Agent's decisions, allowing us to separate the impact of the Agent's intended actions (kind or unkind) from the actual outcomes in the Observer's decision to reciprocate. In this setup, the Agent chooses whether to help two different Recipients; however, only one of these decisions is randomly implemented. The Observer sees one or both attempts and knows which choice was implemented, enabling them to base their reciprocation on these elements distinctly. Additionally, as in the first game, the observability of the Agent's second choice varies. This design also provides an opportunity to study how Observers respond to Agents who act guilefully—appearing kind in public towards one recipient while being unkind to another in less observable conditions.

Taken together, these two experimental games allow us to disentangle the motivations for indirect reciprocity. Surprisingly, we find that while Agents do strategically respond to the observability of their actions, Observers reward helpful behavior no differently when done under high or low observability. That is, Observers act as though they do not care about the inner motives of the Agents by rewarding good deeds regardless of whether they were done by genuinely altruistic people or not. Instead, both the outcome caused by the Agent's choices, and their attempts to be helpful regardless of outcomes, independently impact reciprocity. We conclude that, even if people internally care about others' types, their inferences about those types are not sophisticated, or people prefer to give the benefit of the doubt rather than inferring behavior probabilistically, so that type-based preferences do not accurately reflect choices. Intentions-based preferences, which predict similar first-order effects, are more consistent with the observed relationship between actions and indirect reciprocation.

This paper offers a substantive contribution to the literature on indirect reciprocity, a specific type of reciprocity that has been extensively studied in the context of cooperative behavior in groups.² Since the early origins of game theory, it has been known that positive reciprocity can be a stable outcome in repeated interactions between two given players due to the possibilities opened by the Folk Theorem. Such *direct reciprocity* between two players is however unable to explain the widespread nature of helpful behavior in large societies where most interactions do not take place within long-lasting dyadic relationships like the traditional associations between members of a small community (Nowak and Sigmund, 2005). To explain the widespread nature of

²The literature has distinguished two types of indirect reciprocity: downstream reciprocity, in which somebody is kind to another person after witnessing that person being kind to somebody else, and upstream reciprocity, in which somebody is kind to another person after having been the recipient of the kindness of somebody else. We focus here on downstream reciprocity, but see Simpson et al. (2018) for an integrative view of these two phenomena.

cooperation in groups (and not just in dyads) of people, Alexander (1987) proposed the notion of *indirect reciprocity*, whereby people's cooperative behavior with group members is associated with a positive reputation and where people cooperate more with those having a better reputation. The possibility for reputation-based indirect reciprocity to be an equilibrium of social interactions has been supported by standard game theory (Kandori, 1992) and evolutionary game theory models (Sugden, 1986; Nowak and Sigmund, 1998). An important question in this approach is whether the reputational effect of a failure to cooperate in the past depends on whether it was justified as punishment of others' non-cooperation (see Okada (2020) for a review).

Evidence of both negative (punishing those who are unkind to others) and positive (rewarding those who are kind to others) indirect reciprocity is widespread. Most broadly, third-party punishment, i.e. negative indirect reciprocity, is frequently documented in ethnographic studies across societies (Fessler, 2002; Greif, 1993, 1994; Mathew and Boyd, 2011). Both positive and negative indirect reciprocity have also been identified in field settings: Hairdressers who collect donations for charity receive higher tips (Khadjavi, 2017), online service requests are more likely to be honored when made by user profiles with a history of providing service to others (van Apeldoorn and Schram, 2016), and spectators are willing to punish those to violate social norms such as littering (Balafoutas, Nikiforakis and Rockenbach, 2014).

Laboratory experiments that implement repeated interactions in groups (Wedekind and Milinski, 2000; Milinski, Semmann and Krambeck, 2002; Wedekind and Braithwaite, 2002; Semmann, Krambeck and Milinski, 2004; Seinen and Schram, 2006) also confirm that subjects are more likely to help those with better public record of helpfulness, but Engelmann and Fischbacher (2009) find that approximately half of this helpfulness is due to strategic investment in reputation rather than indirect reciprocity. There is therefore heterogeneity in people's motivation to help, with some being genuinely altruistic and others motivated to gain future benefits of a good reputation.

The game theoretic models and repeated game experiments cited above have provided support to the idea that indirect reciprocity can be sustained in a population when players condition their kind behavior towards other people on those people's track records of past kindness. However, these studies are silent on the proximate psychological motivations behind the decision to reciprocate with people having a good reputation.

Several experimental studies have also found evidence of positive indirect reciprocity in one-shot interactions (e.g. Kahneman, Knetsch and Thaler, 1986; Turillo et al., 2002; Eckel and Grossman, 1996; Güth et al., 2001; Servátka, 2009; Stanca, 2009; Herne, Lappalainen and Kestilä-Kekkonen, 2013). And even in one-shot games, third-party punishment has been extensively documented in the experimental literature (e.g. Balafoutas, Grechenig and Nikiforakis, 2014; Fehr and

Fischbacher, 2004). Given the one-shot, anonymous setting of these studies, it is widely accepted that the game-theoretic explanation for the existence of indirect reciprocity in social interactions is only one part of the story. People likely do not engage in reciprocal behavior by consciously playing the equilibrium strategy of a game. Rather, they follow pro-social preferences that act as proximate psychological motivations inducing people to play equilibrium strategies (Binmore, 2005; Bowles and Gintis, 2011).

Our study helps to bridge the literature on cooperation that can be rational in repeated interactions and the literature on social preferences that aims at explaining the psychological motivations for cooperation (that can arise even in one-shot interactions). The literature on social preferences has provided several models for the psychological motivations driving direct reciprocity, and in the present study we investigate whether similar psychological motivations underpin indirect reciprocity. Three broad categories of models stand out. First, outcome-based preferences, whereby people may care about fair allocations as such and therefore have preferences for the outcomes resulting from interactions not to be too unequal (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002).³ Second, intentions-based preferences, whereby people want to reciprocate towards others who have been kind to them (e.g. Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010; Çelen, Schotter and Blanco, 2017). Third, type-based preferences, whereby altruists are willing to act kindly towards other altruists (e.g. Levine, 1998; Gul and Pesendorfer, 2016; Rotemberg, 2008). Type-based preferences, in particular, appear a natural candidate to explain indirect reciprocity as they predict that people will care about being kind towards altruistic people in general, without the requirement to have interacted with them in the past.⁴

Many game theoretic explanations of indirect reciprocity are based on the selection of partners who have demonstrated a more cooperative type in past interactions (Mailath and Samuelson, 2006). From that perspective, one could expect that type-based preferences are the natural model for the psychological motives underpinning indirect reciprocity. However, we find that such prefer-

³While outcome-based preferences were primarily proposed to explain spontaneous kindness and reciprocity (kindness as an answer to kindness), they can generate reciprocal behavior indirectly since a first act of kindness changes the allocation between players.

⁴Experimental evidence suggests that direct reciprocity is motivated by both the positive outcome from receiving someone's help and the intentions behind this help. This has been shown based on how people reciprocate actions that were chosen willfully versus randomly or by imposition (Rutte, Wilke and Messick, 1987; Cox, 2004; Blount, 1995; Offerman, 2002; Falk, Fehr and Fischbacher, 2008; Klempt, 2012; Charness, 2004; Charness and Levine, 2007), or when choices are made without knowledge of their consequences (Kagel, Kim and Moser, 1996) or by varying foregone options rather than the source of the decision, (Brandts and Solà, 2001; Nelson, 2002; Falk, Fehr and Fischbacher, 2003; Andreoni, Brown and Vesterlund, 2002; McCabe, Rigdon and Smith, 2003). Much of this evidence is also consistent with type-based reciprocity, which has been suggested as an alternative explanation that additionally explains some features of reciprocity that intentions-based models can't (Orhun, 2018).

ences are surprisingly unable to explain our results. Instead, we find that indirect reciprocity seems to be primarily driven by a mix of outcome-based and intentions-based preferences. These results provide novel insights into the role played by reputation in fostering cooperation in large populations. It may not be necessary for reputation to be a reflection of the inner motives of the players, but simply an indication of the predictable reliability of the pro-social behavior of other people so long as their reputations remain on the line. In other words, it may be enough to care about rewarding people doing good deeds for cooperative equilibrium to be sustained without caring about the possible instrumental reasons that may motivate such good deeds. We discuss whether and when preferences about the inner motives may also play a role.

Our results have clear implications for theories of social preferences: we suggest that models of intentions-based preferences should be adapted to allow for individuals to care about intentions displayed towards others. They also raise the question of what type of reputation agents care about building in the first place – if reciprocators do not engage in type inference, why should agents engage in type signaling, as suggested by many models?

Section 2 describes the experimental framework within the context of relevant theory, Section 3 describes the experimental procedures, Section 4 describes our main results, and Section 5 provides further analysis and discussion.

2 Experimental Design and Theoretical Predictions

2.1 Experimental Design

We use two games to disentangle the various motivations for indirect reciprocity, the “4-player game” and the “3-player game”. We introduce the 3-player game first, being simpler, followed by the 4-player game.

Figure 1 displays a summary of the design of the 3-player game. It involves three roles: the Agent, Observer, and Recipient. The initial endowments for the Agent and the Observer are 300 points each, while the Recipient begins with none. The game commences with the Agent deciding to either help (H) or not help (N) the Recipient. If the Agent opts for H, they lose 100 points and the Recipient receives 250 points. The choice of N leaves the endowments as they are.

In addition, we generate random variation in the observability of the Agent’s choice, without deception. Specifically, the Agent’s decision to help is “quasi-private”: the Observer gets to observe it with a probability (p) of either 0.1 or 0.9. After observing H, N, or U (unobserved), the Observer decides whether they will help the Agent or not (H or N). As in the Agent’s choice, if the

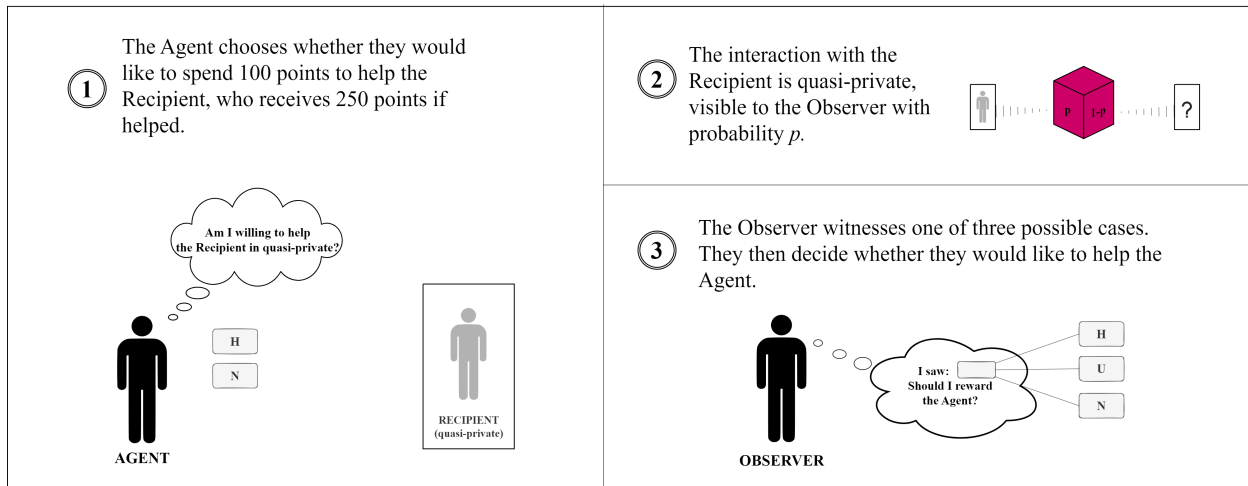


Figure 1: The Three-Player Game

Observer chooses H then they forego 100 points while the Agent receives 250; otherwise points are unchanged. The Recipient makes no decision and the mechanics of the game, including the value of p , is common knowledge to all three players.

This setting is inspired by the repeated helping game (Engelmann and Fischbacher, 2009), but, in our experiment, indirect reciprocity occurs in one-shot interactions: the Observer's choice to reciprocate is not simultaneously the basis for others' later reciprocation.

Figure 2 displays a summary of the design of the 4-player game. It extends the 3-player game by giving the Agent the opportunity to try to help two different Recipients. The game consists of the Agent, Observer, Recipient 1, and Recipient 2, who begin with endowments of 300, 300, 0, and 0 points respectively. First, the Agent chooses H or N for Recipient 1, and then chooses H or N for Recipient 2, with payoff consequences exactly as in the 3-player game if implemented. The Agent's decision towards Recipient 1 is always visible to the Observer, but their decision towards Recipient 2 is quasi-private, observed with probability $p \in \{0.1, 0.9\}$. The Agent also knows that exactly one of these decisions will be randomly chosen to be implemented so that final payoffs correspond exactly to the 3-player game. The Observer thus witnesses one of six possible combinations of the Agent's intentions towards Recipient 1 (H or N) and Recipient 2 (H, N, or U), at each level of p . After witnessing the Agent's choices and which one is (randomly) implemented, the Observer decides whether to help the Agent, just like in the 3-player game. Both Recipients make no decisions and the mechanics of the game, including the value of p , are common knowledge to the players.

Altering the value of p in either the 3-player game or 4-player game impacts the strategic incentives for the Agent. A higher p gives the Observer more opportunity to indirectly reciprocate

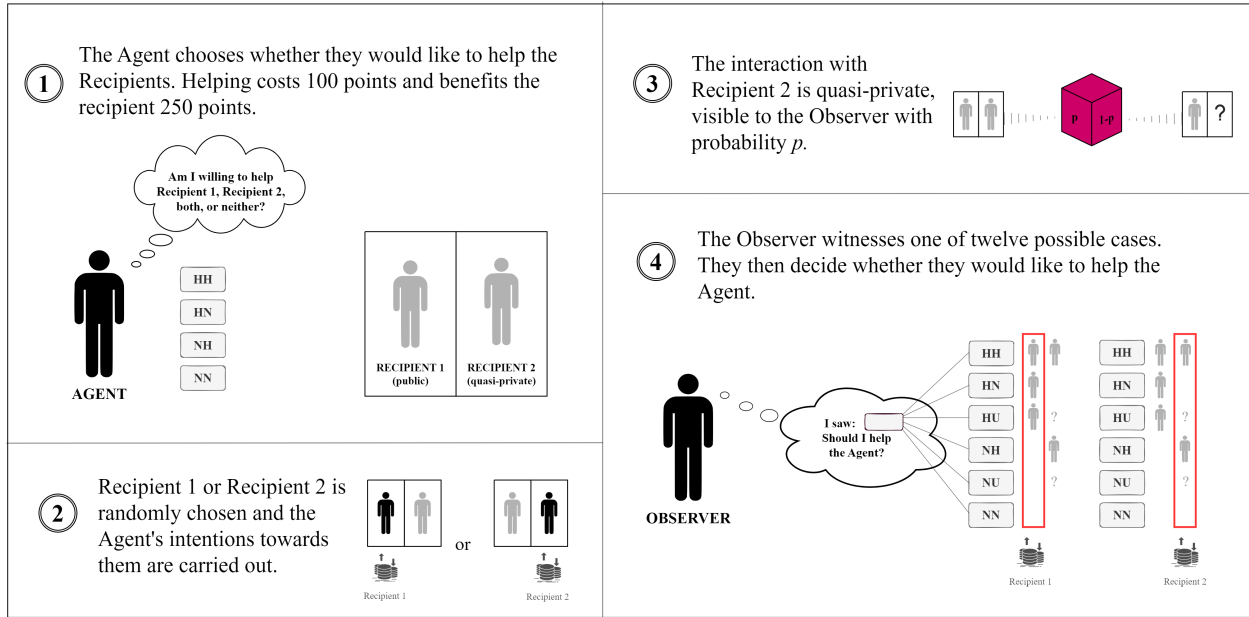


Figure 2: The Four-Player Game

based on the Agent's choice(s). This then correspondingly influences the Observer's perception of the Agent's overall altruism. These variations allow us to test for type-based reciprocity.

In the 4-player game, the random implementation of one of the two choices made by the Agent allows us to compare the level of indirect reciprocity when outcomes change while intentions are constant or vice versa. For example, an Agent choosing HN might experience different reciprocation based on which choice is selected for payment (H with Recipient 1 or N with Recipient 2). Furthermore, an Observer may reciprocate differently after observing HH versus HN, even if the choice implemented is H, the same in both cases.

2.2 Theoretical predictions

Theoretical approaches to reciprocity can be broadly categorized into three classes: 1) outcome-based preferences, 2) intentions-based preferences, and 3) type-based preferences. Outcome-based social preferences, which encompass pure altruism and distributional preferences, are naturally applicable to indirect reciprocal interactions (e.g. Fehr and Gächter, 2000; Bolton and Ockenfels, 2000; Charness and Rabin, 2002). Intentions-based preferences (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010; Çelen, Schotter and Blanco, 2017) encapsulate the extent to which people react to kind intentions. Good deeds, measured by the helpfulness of one's intentions, are reciprocated. While this could also drive indirect reciprocity,

it is unclear whether we would feel compelled to reciprocate intentions on someone else's behalf, and existing models do not accommodate this possibility. Type-based models of reciprocity, also known as interdependent preferences (Levine, 1998; Gul and Pesendorfer, 2016; Rotemberg, 2008), in which we treat others according to *their* levels of altruism, are naturally applicable to indirectly reciprocal interactions. These models suggest we are more altruistic towards people who are (inferred to be) the most purely altruistic overall, rather than reciprocating specific outcomes or good deeds towards someone else.

We now formulate hypotheses based on these alternative models of reciprocity: outcome-based preferences, intentions-based preferences, and type-based preferences.

2.2.1 Outcome-based Analysis

Outcome-based models are intuitively applicable to both the 3-player game and 4-player game. Concepts like inequality aversion, pure altruism, and similar models tend to predict that an Observer will be more inclined to opt for 'H' (help) when they witness an Agent's choice of 'H' being implemented.⁵ More broadly, we hypothesize that reciprocity may be directed at others who are judged favorably on a consequentialist basis. Hence, we make the following hypothesis:

Hypothesis 1. *Due to outcome-based reciprocity:*

1. *Observers will reciprocate more frequently after witnessing H than N in the 3-player game, regardless of p.*
2. *Observers will reciprocate more frequently after witnessing H being implemented in the 4-player game than after witnessing N being implemented.*

2.2.2 Intentions-based Analysis

Intentions-based models of direct reciprocity have received a great deal of attention (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010; Çelen, Schotter and Blanco, 2017). Despite variations in the definition of kind intentions, a common thread unifying these models involves player A's kindness towards player B being gauged through the expected payoff B will receive as a result of A's choices. This judgement is relative to the options available

⁵After an Agent chooses to help the payoffs for (Agent,Observer,Recipient) stand at (200,300,250) while after an Agent chooses not to help they stand at (300,300,0). A purely altruistic Observer would therefore put larger weight on the first Agent if utility is concave. A strictly inequality-averse Observer would not want to cause disadvantageous inequality in either case, but would cause worse inequality if helping the latter.

to player A, given A's expectations of other players' choices. If player A exhibits kind intentions towards player B, this then motivates player B to reciprocate this kindness with kindness. The corollary is also true: unkind actions trigger negative reciprocity. In equilibrium, all players reciprocate optimally, in accordance with their rational expectations of the other players' kindness.

Analyzing intentions-based models of reciprocity presents a certain level of complexity due to the dependence on psychological equilibrium concepts (Geanakoplos, Pearce and Stacchetti, 1989; Battigalli and Dufwenberg, 2009) and rational higher-order expectations. Moreover, the models are built to analyze *direct* reciprocity: The issue with applying these models to indirect reciprocity is that the observed kindness is directed toward others, not me. However, what is clear is that these models propose that kindness is evaluated based on intentions, and we conceptually extrapolate that even an uninvolved third party may reward kind actions. We do not endeavor to formally adapt these models to our setting, but rather, in our games, we interpret "intentions-based" indirect reciprocity as involving rewarding kind intentions demonstrated *towards someone else*.

In the context of our 3-player game, it is clear that H is kinder than N. We further assume that the Observer's judgement of kindness towards the Recipient(s) does not rely on p , given that the Recipient's payoff does not directly hinge on p . In the case of the 4-player game, it is similarly clear that HH is kinder than HN or NH, both of which are kinder than NN. We further assume that HN and NH convey equal kindness in the 4-player game, because the greater observability of the choice towards Recipient 1 does not directly affect the outcomes of either Recipient 1 or Recipient 2.

Hypothesis 2. *Due to intentions-based reciprocity:*

1. *Observers will reciprocate more frequently after witnessing H than N in the 3-player game, regardless of p .*
2. *Observers will reciprocate most frequently in the 4-player game after witnessing HH. They will reciprocate equally after witnessing HN or NH, regardless of which outcome is implemented. They will reciprocate least after witnessing NN. Reciprocation rates will not depend on p .*

Note that Hypothesis 1 and Hypothesis 2 make the same prediction in the 3-player game, as helpful intentions and outcomes are inseparable in this context. The 4-player game, however, allows us to test intentions and outcomes independently.

2.2.3 Type-based Analysis

Turning now to an analysis of motives, we consider models of type-based reciprocity, such as that proposed by Levine (1998). In these models, Observers reward individuals based on their perceived character or “type”. Here, the critical feature is the ability of even uninvolved Observers to glean insights into the altruistic motivations of other players, and later make decisions grounded in those inferences. These models are well-suited to explain indirect reciprocity, and we adapt Levine’s (1998) model to our experimental games.

Every player, denoted by i , is characterized by their individual level of altruism, α_i . The distribution of α_i is assumed to be uniform, with $\alpha_i \sim \phi = U[0, A]$. While this assumption does not alter the qualitative conclusions derived from the model, it greatly simplifies the analysis and allows for closed-form equilibrium calculations. We also restrict our attention to non-pooling equilibria in which both H and N are chosen by agents in the 3-player game and both HH and NN are chosen in the 4-player game. This would trivially be true if α_i had full support on \mathbb{R} , but with a uniform distribution, this requires that A is large enough that some agents help but not so large that all help in order to be rewarded: $A_{\min} < A < A_{\max}$. Precise values for these bounds are derived in the proofs of Propositions 1 and 3 and are shown to be consistent with reasonable parameter values.

We refer to the Observer’s altruism parameter as α_O and the Agent’s as α_A . Player i ’s overall altruism towards another player j is contingent on α_i and i ’s expectation of α_j according to the following equation:

$$v_i = u_i + (\alpha_i + \lambda E_i[\alpha_j]) u_j, \quad (1)$$

In this equation, the total utility of player i , v_i , is the sum of i ’s personal consumption utility u_i and the consumption utility of the other player j , weighted by the sum of player i ’s altruism level α_i and their expectation of player j ’s altruism level α_j , multiplied by a weighting factor $\lambda \in [0, 1]$.⁶

We analyze the 3-player game using the solution concept of Perfect Bayesian Equilibrium (PBE). A PBE consists of a strategy profile and a system of beliefs such that (i) the strategies are sequentially rational given the beliefs, and (ii) the beliefs are consistent with the strategies and are updated using Bayes’ rule whenever possible. Note that because we restrict attention to situations in which both H and N are chosen by agents in equilibrium, there are no off-equilibrium path

⁶We follow Levine (1998) in assuming that consumption utility is linear; given our binary choice setting this is an immaterial simplification. We also omit a normalizing constant of $1 + \lambda$ that would turn the weight on u_j into a weighted average of altruism parameters because this is also immaterial to the qualitative predictions and simplifies notation.

beliefs to specify.

In a PBE of this game, the Agent's decision to help depends on their altruism level α_A and the Observer's equilibrium strategy. The Observer's decision to help, in turn, depends on their altruism level α_O and their belief about the Agent's type (altruism level) conditional on the observed action. When considering whether to help the Agent with a benefit b at a cost c to the Observer, the Observer simply weighs the altruistic utility obtained from helping against its monetary cost. Therefore, the difference in utility between helping and not helping is $(\alpha_O + \lambda E[\alpha_A])b - c$. Since this value is monotonically related to α_O , Observers with sufficiently high types opt to help in equilibrium. We can therefore define cutoff values O_H , O_N , and O_U as the values of α_O above which the Observer chooses to help upon witnessing H, N, and U respectively. Hence, the expected utility of helping is:

$$\underbrace{(pP(\alpha_O > O_H) + (1 - p)P(\alpha_O > O_U))b}_{\text{Expected reciprocation utility}} + \underbrace{(\alpha_A + \lambda E[\alpha_R])b}_{\text{Altruism utility}} - \underbrace{c}_{\text{Helping cost}}.$$

Conversely, the expected utility from choosing N only features the possible reciprocation by the Observer:

$$(pP(\alpha_O > O_N) + (1 - p)P(\alpha_O > O_U))b.$$

Similarly to Observers, we can define a cutoff value A_H for the Agent's altruism parameter α_A above which the Agent chooses to help. Given the cutoff strategy employed by the Agent, the Observer, upon witnessing H, can infer that the Agent's altruism level is at least A_H . Conversely, if the Observer observes N, they can deduce that the Agent's altruism level is at most A_H . Armed with this knowledge, the Observer can choose whether to extend help to the Agent as described above.

Proposition 1. *In the 3-player game with utility determined by (1), there exists a Perfect Bayesian Equilibrium in which the Agent's and the Observer's decisions to help are determined by whether their respective altruism levels, α_A and α_O , are higher than thresholds A_H , O_U , O_N , O_H , such that:*

- *The Agent helps if $\alpha_A \geq A_H$*
- *The Observer helps when the Agent's action is unobserved if $\alpha_O \geq O_U$*
- *The Observer helps when the Agent is observed not to have helped if $\alpha_O \geq O_N$*

- The Observer helps when the Agent is observed to have helped if $\alpha_O \geq O_H$

In this equilibrium, we have $O_H < O_U < O_N$ and $0 < A_H, O_N, O_U, O_H < A$.

All proofs are in Appendix A. Figure 3 summarizes the cutoff values used to govern the Observer's decision, based on their Bayesian beliefs about the Agent. The region below O_H represents Observers who withhold helping the Agent even if they see them help the Recipient. Conversely, Observers who fall in the region above O_N are willing to help regardless of what they saw. In between are Observers that reward the Agent reciprocally.

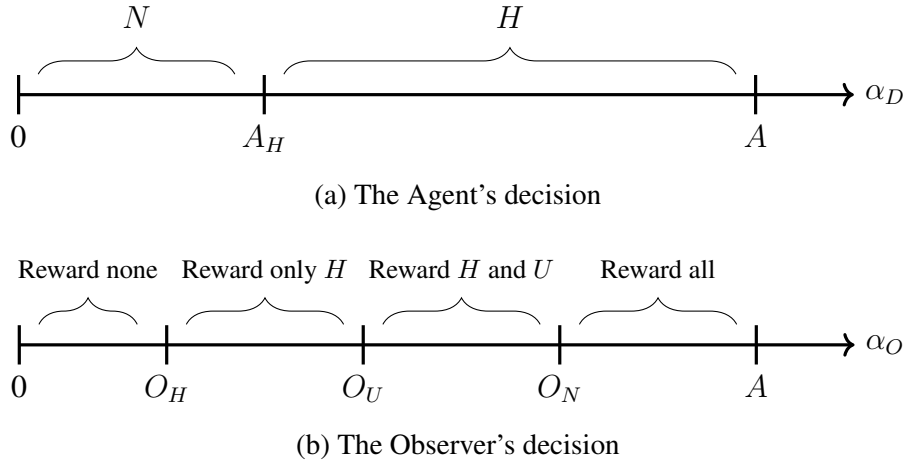


Figure 3: Agent's and Observer's decisions in the 3-player game, as a function of their altruism type parameters. An Agent helps if $\alpha_A > A_H$. An Observer reciprocates after witnessing $X \in \{H, N, U\}$ if their altruism parameter exceeds O_X .

A feature of type-based preferences is that they generate strategic motives to signal desired types. In our setting, Agents with high α are more likely to help regardless, so H signals higher altruism than N. But the value of α_A that Observers can infer also depends on the Agent's strategic incentives to help, which are stronger when p is high. A high p can prompt Agents with lower α to extend help since they now have a better chance of being observed and subsequently rewarded. The values of A_H , O_H and O_N therefore depend on p according to Proposition 2:

Proposition 2. *In the 3-player game,*

1. *More Agents choose H when p is high. That is, A_H is decreasing in p .*
2. *Fewer Observers reciprocate after witnessing H when p is high. That is, O_H is increasing in p .*

3. Fewer Observers reciprocate after witnessing N when p is high. That is, O_N is increasing in p .

These comparative static predictions form our key experimental test of type-based preferences in the 3-player game.

Turning to the 4-player game, the Agent now has the opportunity to try to help two Recipients. The Agent's first decision of whether to help Recipient 1 is always clearly observed by the Observer. In contrast, their subsequent choice of whether to help Recipient 2 is quasi-private (observable with probability p), as in the 3-player game. It is common knowledge that only one of these choices will be implemented, and the Observer will be able to see which one was randomly implemented. At that point, the Observer will have the same opportunity for indirect reciprocity as in the 3-player game. In this game, we can define the Agent's strategy as $x_A \in \{HH, HN, NH, NN\}$, where the first letter denotes the public choice towards Recipient 1 and the second letter refers to the quasi-private choice towards Recipient 2.

In the 3-player game, all possible scenarios are on the equilibrium path in the non-pooling equilibria we consider. In contrast, in the 4-player game some actions may not be supported in any PBE. Specifically, while $A_{\min} < A < A_{\max}$ guarantees that HH and NN are both chosen in equilibria, HN and/or NH may not be. We prove the existence of an equilibrium that takes one of two forms depending on observability p . For sufficiently low p , Agents will choose one of the three action profiles: HH , HN , and NN . We call this a Type 1 equilibrium. When p exceeds a threshold \bar{p} , only HH and NN are supported in equilibrium. We call this a Type 2 equilibrium. In both forms, NH is off the equilibrium path if we restrict attention to equilibria that satisfy the D_1 Criterion (Cho and Kreps, 1987), with HN similarly off the equilibrium path in Type 2 equilibrium. Additionally, a range of alternative off-path beliefs result in sufficiently unfavorable evaluations of these actions and can be used to sustain the equilibrium. Further details are provided in the appendix.

Proposition 3. *In the 4-player game with utility determined by (1), there exists a Perfect Bayesian Equilibrium characterized by thresholds for the Agent (A_{HH}, A_{HN}, A_{2HH}) and the Observer ($O_{HH}, O_{HN}, O_{HU}, O_{NN}$) that depend on the observability level p relative to a threshold \bar{p} . The equilibrium takes one of two forms:*

1. *Type 1 equilibrium: If $p < \bar{p}$, the Agent chooses HH if $\alpha_A > A_{HH}$, HN if $A_{HN} < \alpha_A < A_{HH}$, and NN if $\alpha_A < A_{HN}$, satisfying $0 < A_{HN}, A_{HH} < A$. The Observer reciprocates after witnessing $x \in \{HH, HU, HN, NN, NU\}$ if $\alpha_O > O_x$, satisfying $0 < O_{HH} < O_{HU} < O_{HN} < O_{NN} = O_{NU} < A$.*

2. *Type 2 equilibrium: If $p > \bar{p}$, the Agent chooses HH if $\alpha_A > A_{2HH}$ and NN otherwise, satisfying $0 < A_{2HH} < A$. The Observer's thresholds for reciprocation satisfy $0 < O_{HH} = O_{HU} < O_{NU} = O_{NN} < A$.*

Intuitively, for high values of p , the HN strategy loses appeal to Agents. The reason is the smaller chance of being able to “get away with” helping only when observed. The value of HN to the Agent arises when Observers witness HU : Observers cannot distinguish those who chose HH or HN in this case and thus reciprocate towards the average altruism level represented by either action. But when p is large, HU is infrequent and the likelihood of being pooled with altruistic players diminishes. Instead, being seen to be willing to help in public but not in quasi-private simply reveals that the Agent is not altruistic enough to help unconditionally.

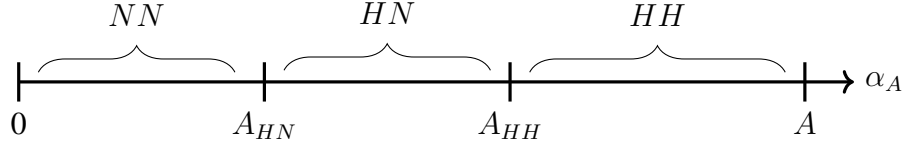
In a Type 1 equilibrium, the Observer witnesses one of five scenarios: HH , HN , HU , NU , or NN . Given that the NH strategy is never adopted in equilibrium (if the Observers' beliefs after witnessing NH off the equilibrium path are sufficiently unfavorable), it is understood that NU is a disguise for NN . Consequently, the Observer adopts one of four cutoff strategies defined by the altruism thresholds O_{HH} , O_{HN} , O_{NN} , and O_{HU} above which the Observer will help after observing each scenario, similarly to the 3-player game detailed above.

Working backwards, we can show that very unaltruistic Agents choose NN , mid-altruism Agents choose HN (i.e. helping only in public), and high-altruism Agents choose HH . Formally, the strategy of the Agent is defined by two cutoff values in α_A — A_{HN} , below which the Agent chooses NN and above which they switch to HN , and A_{HH} (greater than A_{HN}), above which they transition to HH .

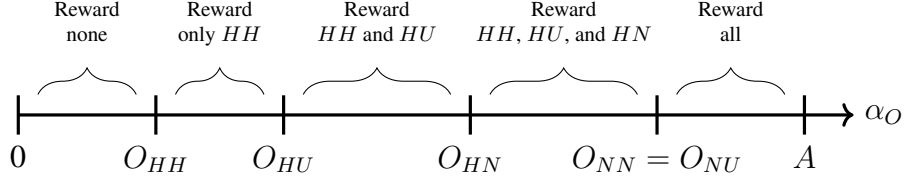
In a Type 2 equilibrium, the Agent's behavior is defined by a single cutoff value A_{2HH} : if $\alpha_A > A_{2HH}$ the Agent will choose HH , and otherwise will choose NN . The Observer can therefore infer exactly what the Agent chose based on the first, public choice towards Recipient 1, and reciprocates accordingly.

Type 1 equilibrium is visually represented in Figure 4 which delineates the cutoff values for the Agent and the Observer, dictating their actions and demonstrating their beliefs.

As in the 3-player game, the thresholds A_{HN} , A_{HH} and O_x in a Type 1 equilibrium are implicitly functions of p , as derived in detail in Appendix A. In a Type 2 equilibrium, there is no longer any ambiguity about the Agent's choice towards Recipient 2 because it is always identical to their choice towards Recipient 1, and so changes in p do not entail a real change in observability. Because Type 1 equilibria involve subtler tradeoffs between more than two strategies, the relationship between strategies and p is not as straightforward as in the 3-player game, but the key relationship between rates of choosing HH and rates of reciprocation towards HH mirrors the 3-player game



(a) The Agent's decision



(b) The Observer's decision

Figure 4: The Agent's and Observer's decisions in the 4-player game, as a function of their altruism type parameters. The Agent helps both recipients if $\alpha_A > A_{HH}$, only Recipient 1 (in public) if $\alpha_A > A_{HN}$, and neither recipient otherwise. The Observer reciprocates after witnessing scenario $x \in \{HH, HU, HN, NN, NU\}$ if their type parameter exceeds O_x .

analysis. Specifically, as long as more Agents choose HH when p rises (as is empirically the case), fewer Observers will indirectly reciprocate after witnessing HH . We have the following result:

Proposition 4. *In the 4-player game:*

1. *In a Type 1 equilibrium, reciprocity towards HH is decreasing if HH becomes more prevalent when p rises. That is, O_{HH} rises (making it a threshold that is met for fewer Observers) if A_{HH} falls.*
2. *In a Type 2 equilibrium, the equilibrium characterization is not dependent on p , so changes in p do not affect the equilibrium as long as p remains above \bar{p} .*

In a Type 1 equilibrium, if Agents generally respond to increasing observability by helping more often in quasi-private (rather than, on the other hand, by switching from HN to NN as HN becomes less attractive), then this leads to a dilution of the average altruism of HH -choosing Agents. In response, Observers are less likely to reciprocate towards HH . When p reaches a threshold \bar{p} , no Type 1 equilibrium exists anymore, and the Type 2 equilibrium that prevails is not affected by further increases in p because HU is perfectly understood to represent HH .

We can now define our last experimental hypothesis based on propositions 2 and 4.

Hypothesis 3. *Due to type-based reciprocity:*

1. *When observability (p) is high, more Agents choose to Help (H) in the 3-player game.*

2. *When observability (p) is sufficiently high, no Agents will choose to help only in public (HN) in the 4-player game.*
3. *When observability (p) increases, fewer Observers reciprocate towards Agents that choose H in the 3-player game, and similarly reciprocate less towards Agents that choose HH in the 4-player game so long as Agents choose HH more often at higher p .*

3 Experimental Procedures

We conducted nine experimental sessions between October 2021 and March 2022, with a total of 168 participants. Two participants were dropped from analysis due to prior familiarity with the research project, leaving a sample of 166. Sessions were advertised through the University of Queensland’s SONA participant database, which includes several thousand students and staff members. Each session was conducted in person at the School of Economics Laboratory, lasted one hour, and paid an average of AUD \$28.31. All tasks were computerized using oTree (Chen, Schonger and Wickens, 2016).

Figure 5 shows a timeline of the experiment. For every distinct game, participants were anonymously and randomly re-matched with other subjects in the room. The complete protocol is shown in Online Appendix D.

The experiment began with brief instructions outlining the session without revealing details of the upcoming games. Participants were then randomly paired and played a mini-dictator game. In each pair, both participants chose between two allocations, (\$1,\$1) and (\$2,\$0), with one of the decisions randomly selected to be implemented. Outcomes were not revealed until the end of the session. This task provided a simple individual measure of altruism, which we later relate to behavior in our games to test the predictions of type-based reciprocity.

Parts 3 and 4 formed the core of the experiment, and their order was counterbalanced across sessions to control for order effects. In part 3, participants first learned the rules of the 3-player game and were required to answer several comprehension checks correctly before proceeding. They then played six rounds of the game using the direct method, playing once at each level of observability ($p = 0.1$ and 0.9) in each role. This phase allowed participants to become familiar with the game through experience, gaining an understanding of its equilibrium, the perspective associated with each role, and the implications of low and high observability. Outcomes of these rounds were revealed immediately, although only one round was randomly selected for payment, and the paid round was not disclosed until the end of the session.

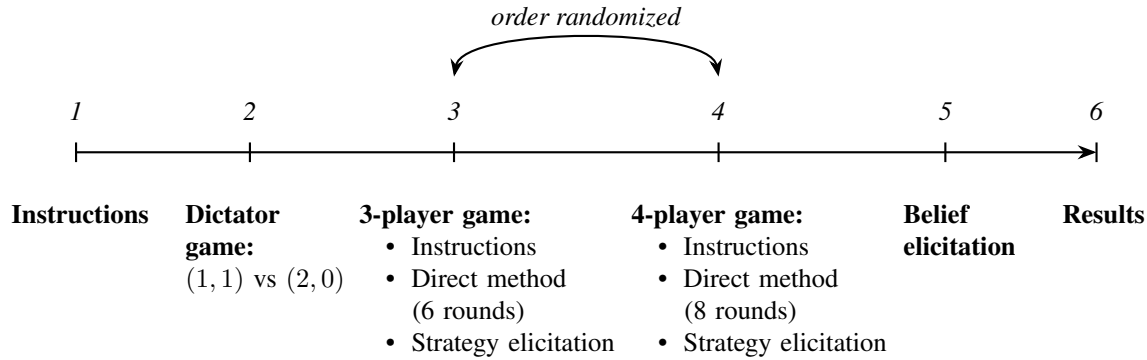


Figure 5: Experimental session timeline.

After the direct-method rounds, we elicited each participant’s complete strategy profile by presenting six additional rounds. At each level of observability, observers were asked to specify their full contingency plans as a function of what they might observe of Agents. The use of the strategy method was essential for capturing observers’ responses to Agent behavior under low observability. To prevent additional learning during strategy elicitation, no feedback was provided between decisions, and the outcome of the randomly selected paid round was revealed only at the end of the session.

Without contemporaneous feedback, decisions could be made in any order. Each participant first played as an Agent under one value of p , then as an Observer with the same p value, followed by the same sequence with the other p value, and finally (purely passively) as a Recipient under each p value. The ordering of high and low observability was randomized. While within-subject designs can introduce order effects, this design was advantageous for our purposes because it allowed Observers to form well-informed beliefs about Agents. After having experienced both roles and both observability levels, Observers could respond with a full understanding of how observability shapes Agent behavior. Since both the order of games (Parts 3 and 4) and the order of p values were randomized, these potential order effects can be identified and controlled for; Online Appendix C shows that neither substantially affected behavior. Moreover, our within-subject design also allows us to examine whether participants’ behavior as Agents is related to how they reciprocate as Observers.

Part 4 followed the same structure as Part 3. Participants were introduced to the rules of the 4-player game and were required to answer several comprehension checks correctly before proceeding. They then played eight rounds of the game using the direct method, playing once at each level of observability for each role ($p = 0.1$ and 0.9). Outcomes of these rounds were revealed immediately, although only one round was selected for payment, and the paid round was not dis-

closed until the end of the session. We then elicited each participant’s strategy profile by presenting sixteen additional rounds, in which Observers specified their contingency plans as a function of the six possible action profiles they might observe. The ordering of p values was randomized as in Part 3. Sixteen rounds, rather than eight, were required to elicit Observer strategies contingent on both possible realizations of the randomly selected Recipient who the Agent was able to help. Two of these rounds were selected for payment at the end of the session.

In Part 5, participants reported their beliefs about how people played as Agents in earlier parts. They reported rates of choosing H at both levels of p in the 3-player game, along with rates of choosing each of HH, HN, NH, and NN at both levels of p in the 4-player game. They also provided beliefs about choosing (\$1,\$1) in the mini-dictator game. All rates were reported as integer values between 0 and 100. One of these eleven belief reports, each representing the perceived proportion of participants choosing a given action, was randomly selected for payment according to the quadratic scoring rule. Finally, in Part 6, the results of the rounds selected for payment were revealed in full, along with participants’ corresponding earnings.

We preregistered our hypotheses, analysis plan, sample size, and experimental approach, which can be accessed at <https://aspredicted.org/96k6f.pdf>.

4 Results

All analysis focuses exclusively on data from the strategy method rounds of the 3-player game and 4-player game, except where noted otherwise. Online Appendix B shows that the differences between strategy method choices and direct method data from practice rounds, where comparison is feasible, are minimal. For exposition purposes, we start with the tests of the Type-based Reciprocity, i.e. Hypothesis 3.

4.1 Typed-based Reciprocity

We begin with the analysis of the 3-player game results, since it is designed to be the clearest and strongest test for our type-based model of indirect reciprocity. Notably, the only contextual information available to the Observer in this setup is observability p .

The choice data from the 3-player game for both Agents and Observers are presented in Figure 6. The data unveil two distinct patterns. First, Agents significantly adjust their behavior based on the probability of being observed. 69% of Agents choose to help when observability p is high, while a mere 33% do so when p is low (a significant difference with a two-sample Z-test p -value

< 0.001). In contrast, Observers seem indifferent to the strategic shifts in the Agents' actions— they reciprocate to the Agents 62% of the time when observability is low and 61% when it is high.⁷ This finding strikingly contradicts part 3 of Hypothesis 3.

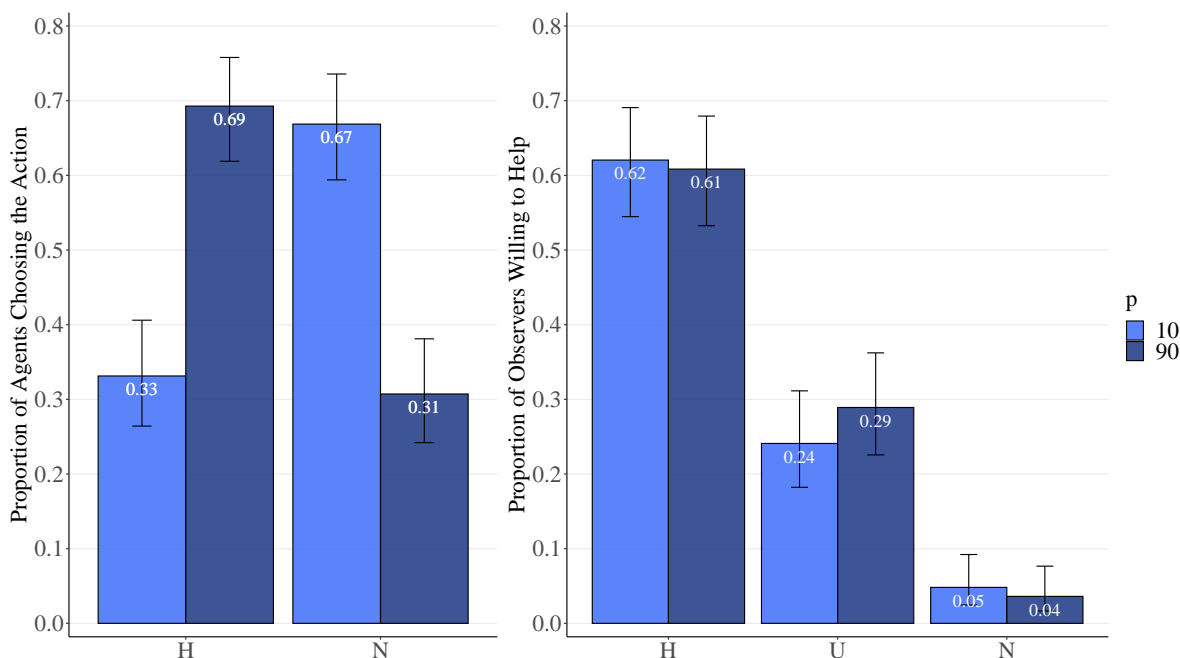


Figure 6: Observer and Agent choices in the 3-player game by observability, i.e. the probability that the Agent's quasi-private decision was witnessed by the Observer. The left panel shows the rates of Agents choosing to be helpful (H) or not (N), while the right panel shows rates of reciprocation in the three possible scenarios the Observer may witness (H, N, or Unobserved (U)). Wilson score 95% confidence intervals are indicated.

This result extends to the 4-player game as well. Figure 7 shows the choice rates for Agents and Observers in the 4-player game, contingent on observability.⁸ The HH case again clearly contradicts Hypothesis 3 part 3, mirroring the 3-player game results. While a considerably higher number of Agents choose HH when $p = 0.9$ than when $p = 0.1$ (59% versus 33% respectively, two-sample Z-test p -value < 0.001), Observers' reciprocation rates are again insensitive to observability (64% when $p = 0.9$ versus 63% when $p = 0.1$). Agents' behavior is in line with the rationale that those with lower altruism are drawn to HH when their secondary choices are more

⁷To provide more statistical perspective on the null result, the observed mean difference is 0.012 with a 90% CI of $[-0.046, 0.070]$. Using a mean equivalence test (TOST) with $\alpha = 0.05$, equivalence requires the CI to lie within $\pm\Delta$; here, the minimum bound is $\Delta_{\min} = 0.071$. Although the estimate is close to zero, the relatively large Δ_{\min} implies that statistical evidence for equivalence rests on wide margins. Nonetheless, we replicate this result in the 4-player game, lending support to its robustness.

⁸ See Appendix B Figure 11 for a complete breakdown of reciprocation rates in the 4-player game by both observability and the recipient randomly chosen to have their outcome implemented.

likely to enhance their reputations. However, the defining factor in Observers’ decisions appears to be the concrete actions of the Agent, rather than the implications of those actions for the Agent’s type.

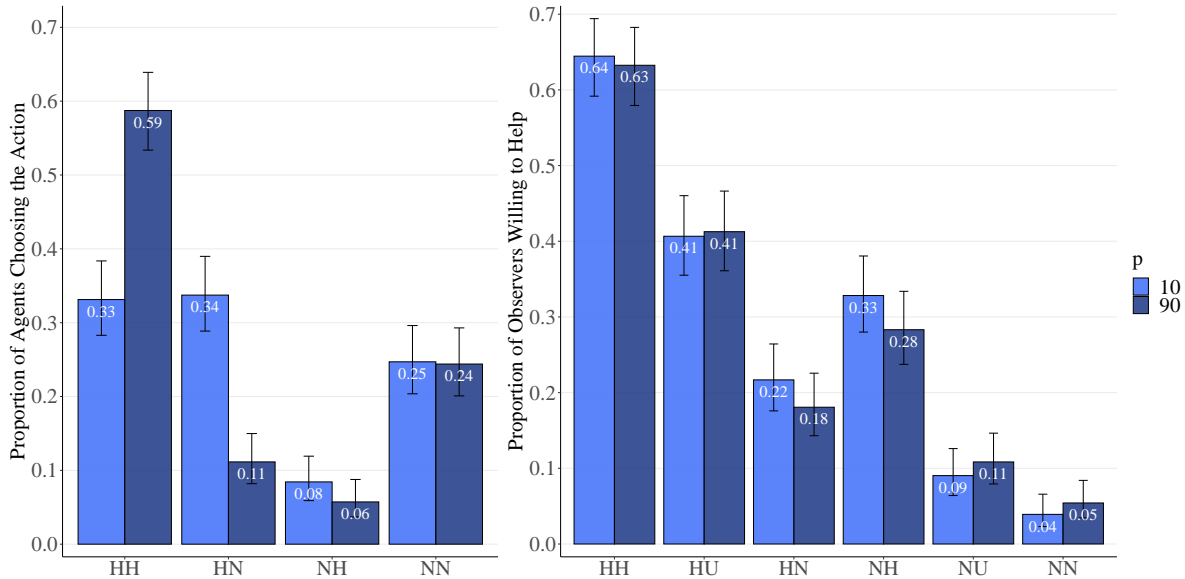


Figure 7: Observer and Agent choices in the 4-player game, by observability p of the Agent’s second quasi-private choice. The left panel shows the rates of Agents choice combinations. The right panel shows Observers’ rates of reciprocation contingent on the six possible scenarios they may have witnessed — Help (H) or Not (N) towards Recipient 1 in public and H, N, or U (Unobserved) towards Recipient 2 in quasi-private. Wilson score 95% confidence intervals are indicated.

Notice that in Figure 7, the increase in Agents’ choice of HH coincides with an (all but) disappearance of HN when observability is high, consistent with Hypothesis 3 part 2. Approximately half of the Agents who opt for HH when $p = 0.9$ deviate from this choice when $p = 0.1$, pivoting towards HN to exploit the ambiguous case of HU. This demonstrates the reputation-building motive of a substantial number of Agents and brings into question the altruistic motivations of any HH-choosing Agent when observability is high. This reinforces our evidence against type-based preferences — while Agents are indeed strategically tailoring their decisions to appear altruistic when a reputational reward is at stake, Observers seem to disregard this information entirely. Importantly, we also find evidence that Observers understand the implications of changes in p and the possible strategic motives of Agents (see Section 5); indeed, Observers always make their choices immediately after encountering the same situation as Agents in order to maximize the salience of the impact of observability on Agents. The Observers’ decisions do not stem from a misunderstanding or ignorance of the Agents’ strategic motives, but seem to be motivated purely by the

Agents' actions themselves rather than inferences about Agents' inherent altruism.

We summarize these findings with the following two Results:

Result 1. *Agents respond strategically to the changing probability of being observed. When observability is higher, they are significantly more helpful in the 3-player game, more likely to choose HH in the 4-player game, and unlikely to choose HN in the 4-player game, consistent with Hypothesis 3 parts 1 and 2.*

Result 2. *Observers do not condition their indirect reciprocity on the observability of the Agents' action as predicted by type-based reciprocity. Specifically, Agents that are seen to choose H in the 3-player game or HH in the 4-player game receive the same reciprocation from the Observer regardless of observability, contrary to Hypothesis 3 part 3.*

4.2 Outcome-based Reciprocity

We next consider the effect of outcomes on Observers' indirect reciprocity. To investigate this question, we use the 4-player game in which random variations in outcome are generated given certain decisions by the Agent. Such randomness is absent when the Agent opts for either HH or NN, as one Recipient invariably receives something in the former scenario and neither Recipient ever benefits in the latter. However, when the Agent opts for choice combinations HN or NH, only one choice is randomly selected to be implemented. This allows us to study how the Observers' reciprocity fluctuates in response to the outcome stemming from the Agent's choice, while keeping that choice—either HN or NH—constant.

Figure 8 shows how Observers' reciprocation varies contingent on the Agent's decision x_A and the randomly chosen Recipient whose outcome is implemented.⁹ As anticipated, reciprocation rates following the observation of HH or NN do not hinge on the Recipient chosen for payment, given that the Agent's outcome remains constant in either situation. However, when the Observer witnesses HN or NH, they demonstrate a stronger inclination to reciprocate if the first or second Recipient, respectively, is chosen for payment. In such circumstances, the Agent's benevolent intent comes to fruition—they incur a cost to assist, and the Recipient duly receives the benefit—resulting in an uptick in the Observer's reciprocation towards them. Thus, Hypothesis 1 is supported in our results. Past outcomes, independent of the actual decisions made, play a significant role in indirect reciprocity.

To summarize,

⁹For simplicity, we pool data across p when describing these results but they continue to hold in the full breakdown of reciprocation rates, as can be seen in the raw data provided in Appendix B.

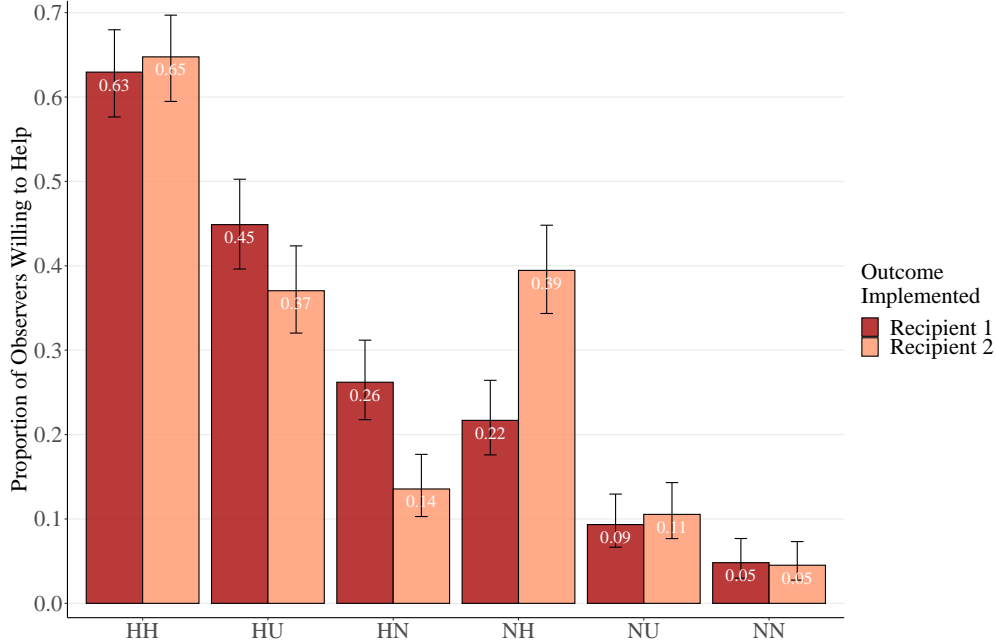


Figure 8: Observer rates of reciprocation in the 4-player game according to whether Recipient 1 or Recipient 2 was randomly selected to have the Agent’s choice towards them implemented. Wilson score 95% confidence intervals are indicated.

Result 3. *Observers indirectly reciprocate more towards Agents who achieve helpful outcomes, consistent with Hypothesis 1.*

4.3 Intentions-based Reciprocity

To study the effect of Agents’ helpful intentions on Observers’ indirect reciprocity we can ask the converse question in the 4-player game: Is indirect reciprocity influenced by Agents’ intentions, holding final outcomes constant? Figure 9 compares these rates of reciprocation. The darker bars to the left represent rates of reciprocation when one Recipient received help. Despite the identical outcomes in terms of help received, Observers exhibit a higher propensity to reciprocate if the Agent intended to help both Recipients (64% for HH) rather than just one (26% for HN and 39% for NH). The differences between HH and HN, as well as between HH and NH, are highly statistically significant (both two-sample Z-test p -values < 0.001).¹⁰

The lighter bars on the right of Figure 9 represent scenarios where the Agent did not provide help to any Recipient, either by choice (NN) or because the selected Recipient was the one not offered assistance (HN or NH). Here too, we observe that Observers are more inclined to reciprocate

¹⁰Comparisons remain highly statistically significant when also controlling for observability.

when Agents intend to help at least one Recipient. The reciprocation rate is only 5% in the NN scenario, but jumps to 14% and 22% in the HN and NH scenarios respectively. The differences between NN and NH, as well as between NN and HN, are also highly statistically significant (both two sample Z-test p -values < 0.01).

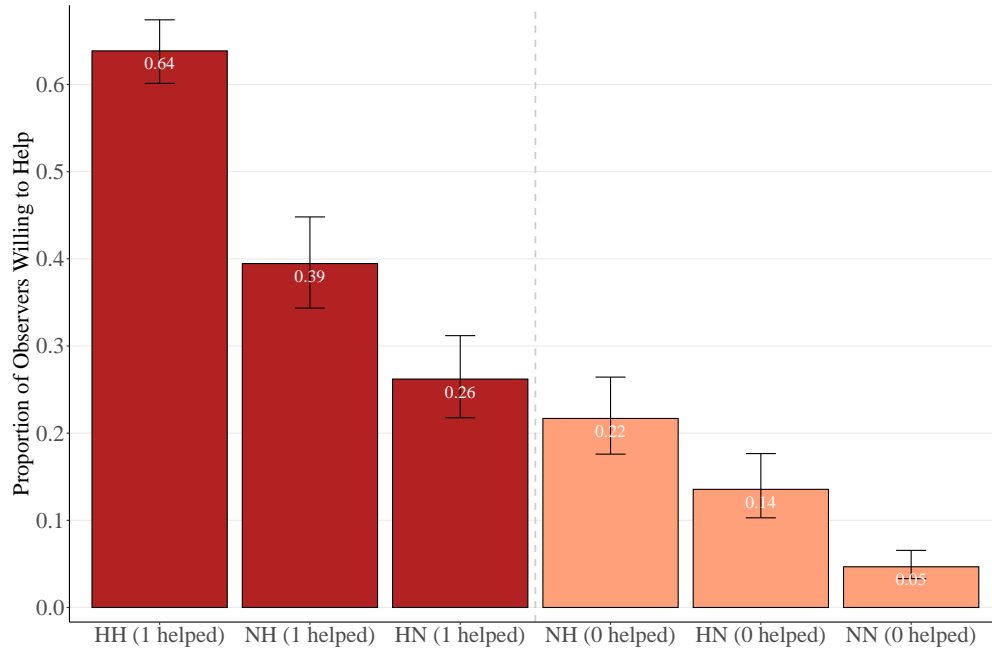


Figure 9: Observer rates of reciprocation in the 4-player game according to whether a helpful outcome was achieved as a result of the Agent’s intentions. Wilson score 95% confidence intervals are indicated.

Hypothesis 2 is supported. When combined with our results that reject the role of type-based preferences, this stands as a remarkable finding. It appears that intentions are inherently significant, rather than primarily serving as indicators of an Agent’s underlying altruistic character. This is true even when these intentions are directed towards someone else, supporting the general approach of intentions-based models of reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010; Çelen, Schotter and Blanco, 2017). However, intentions-based preferences were initially postulated to explain direct reciprocity specifically. It’s not readily apparent that third-party observers should concern themselves with the intentions someone demonstrated towards another, particularly when the intended help might not materialize, and these models do not allow for this possibility. Yet, our results suggest that these intentions provide a strong motive for indirect reciprocity.

To summarize,

Result 4. *Consistent with Hypothesis 2, Observers indirectly reciprocate more towards Agents who display an intent to help a greater number of Recipients.*

Table 1 decomposes the effects of outcome-based and intentions-based reciprocity in the 4-player game using probit regression. “Helpful Outcome?” is a dummy variable taking the value 1 when the Recipient, randomly chosen for payment, benefits from an Agent’s decision to help. “Intentions” is a categorical variable taking values 0, 1, or 2, based on the number of Recipients the Agent opts to help. Data from HU and NU scenarios are omitted as these independent variables aren’t clearly defined in these cases.¹¹ Outcomes and intentions both separately strongly influence indirect reciprocity, in line with our results above.

Dependent Variable: Observer helped Agent		
Case Observed		
	(1)	(2)
Helpful Outcome	0.4830*** (0.0814) [0.1404]	0.5093*** (0.0877) [0.1471]
Intentions	0.7825*** (0.0712) [0.2140]	0.7756*** (0.0730) [0.2101]
Observations	2,656	2,368
Demographic controls	No	Yes

Table 1: Probit regression analysis of Observer reciprocation rates in the 4-player game after witnessing HH, HN, NH, or NN, as a function of the number of recipients the Agent intended to help and whether a helpful outcome was implemented. Standard errors are clustered by participant. Demographic controls include gender, international student status, English as a first language, and previous training in economics. Average marginal effects are shown in square brackets. Statistical significance indicated at * 10%, ** 5%, and *** 1% levels.

All coefficients in the regression analysis are highly statistically significant with or without demographic controls. Using regression (2), the average marginal effect of a helpful outcome suggests that an Observer is 14.71 percentage points more likely to help the Agent if they achieve a helpful outcome. Intentions have a larger average marginal effect of 21 percentage points and,

¹¹Including HU and NU data in the regressions and using empirical expectations of intentions and outcomes as the independent variables yields the same qualitative pattern of findings.

as an Agent can have up to 2 helpful intentions, they emerge as the dominant factor in indirect reciprocity even though these intentions are directed towards third parties.

5 Additional analysis and discussion

5.1 The role of type-based preferences in indirect reciprocity

Our experimental results suggest that type-based preferences play a relatively minor role, if any, in driving indirect reciprocity in our setting. This is a surprising finding given that type-based models seem well-suited to capture reputational concerns that are central to indirect reciprocity (Nowak and Sigmund, 1998; Panchanathan and Boyd, 2003). However, it is important to note that the key predictions distinguishing type-based models from intention-based and outcome-based models concern second-order effects: while all three models predict that helpful actions will be rewarded, type-based models go further in predicting that the degree of reward should depend on how informative the action is about the underlying altruism of the Agent.

Our experimental design allows us to cleanly test for these second-order effects, and we find little evidence of Observers making such fine distinctions. The 3-player game illustrates this most vividly as it maximizes the prominence of p , minimizes factors apart from the signaling value of the decision to help, and thereby makes the motives behind the Agents' choices as simple as possible to understand. All Observers also act as Agents, further facilitating inferences about the Agent's motivations. It's noteworthy that while Agents *do* strongly respond to strategic signaling incentives, the same participants acting as Observers do not condition their reciprocity on this when rewarding the Agent. This suggests that type-based considerations are not a dominant concern, at least in our one-shot strategic settings (we discuss roles for type-based preferences that are not ruled out by our results below).

Given the surprising nature of our results on type-based preferences, we consider three alternative explanations for our findings. First, it could be that Observers do not fully appreciate the strategic incentives faced by Agents in our experiment. However, our belief elicitation data suggests that Observers generally have accurate expectations about how Agents behave at different levels of p , casting doubt on this interpretation. As shown in Table 2, participants accurately predict the substantial difference in Agent helpfulness between low and high observability treatments. For instance, in the three player game, participants believe that 32% of Agents are helpful when $p = 0.1$ and 66% are helpful when $p = 0.9$. This demonstrates the participants' awareness of the strategic incentives at play.

Game	Action	p	Belief	Truth	Difference
Mini-dictator game	H	–	59%	61%	2%
3-player game	H	10	32%	33%	1%
3-player game	H	90	66%	69%	3%
4-player game	HH	10	21%	33%	12%
4-player game	HH	90	43%	59%	16%
4-player game	HN	10	42%	34%	-8%
4-player game	HN	90	23%	11%	-12%
4-player game	NH	10	10%	8%	-2%
4-player game	NH	90	11%	6%	-5%
4-player game	NN	10	27%	25%	-2%
4-player game	NN	90	23%	24%	1%

Table 2: Actual rates and elicited beliefs about the rates of choosing each action in each of the three games played (conditional on observability p).

Second, it’s possible that some types of Observers do engage in type-based reciprocity, but this is obscured by our aggregate analysis. For example, it’s possible that only strategic agents realize that others will be strategic, or perhaps that strategic agents are kinder to agents likely also to be strategic (when $p = 0.9$) due to homophily or a similar motivation. To investigate this, we categorize Observers into different types based on their own behavior as Agents. As all participants play both as Agent and Observer, we can use their choices as Agent in the 3-player game to categorize Observers as Altruists (who help regardless of observability), Strategists (who help only when observability is high), and Selfish (who never help).¹²

Table 3 presents Observer behavior and beliefs for these types. All types clearly recognize that Agents are more helpful when $p = 0.9$. Furthermore, all types, separately, reciprocate equally towards high and low observability helpfulness. Strategists very slightly increase their reciprocation towards helpful Agents when observability is lower, from 71% to 75%, but this difference is not significant and is dwarfed by the effects of intentions or outcomes. In conclusion, we find that agents do not condition their reciprocity on the reputational incentives at play. These results provide further evidence that type-based considerations are not a key driver of reciprocity even when accounting for heterogeneity in Observer types.

Finally, one may wonder whether our manipulation of p successfully varies the reputational incentives and therefore the informativeness of the Agent’s action. Our type-based reciprocity

¹² 5% of participants chose to help when $p = 0.1$ but not when $p = 0.9$. We ignore this small subsample in this analysis, although they are still represented in the aggregate row.

	Count	%	H		U		N		Beliefs	
			90	10	90	10	90	10	90	10
Altruist	47	28%	83%	83%	53%	51%	2%	4%	77%	44%
Strategist	68	41%	71%	75%	28%	16%	1%	1%	75%	26%
Selfish	43	26%	21%	19%	7%	5%	7%	7%	48%	24%
Aggregate	166	100%	61%	62%	29%	24%	4%	5%	66%	32%

Table 3: Observer behavior and beliefs in the 3-player game as a function of their behavior as Agents. “Altruists” types always help as the Agent, “Strategists” help only when observability p is high (90%), and “Selfish” never help. The first two columns show numbers and percentages of each type. The middle six columns show reciprocation rates conditional on observing each of the three scenarios H, U, and N, by observability. The last two columns show beliefs about Agents’ helping rates at each level of observability.

model predicts that if such reputational incentives are larger for $p = 0.9$, then Agents choosing to Help when $p = 0.1$ are more altruistic than those who help when $p = 0.9$. If this were not true, an Observer with type-based reciprocal preferences shouldn’t condition their behavior on observability. The mini-dictator game played by all participants at the start of the experiment provides a rough measure of altruism and the results are consistent with Agent helpfulness being correlated with altruism. The conditional expectation of a participant’s sharing in the mini-dictator game, given their 3-player game type as defined in Table 3, is 77% for Altruists and 63% for Strategists. This difference aligns with the predictions of type-based reciprocity, and is marginally statistically significant with a p -value of 0.06 (one-tailed two-sample Z-test). Thus, our manipulation of observability does seem to reveal differences in underlying altruism, yet Observers do not condition their reciprocity on this information, further reinforcing our main conclusions.

There are several possible interpretations of this collection of results. The evolutionary literature on indirect reciprocity and partner choice suggests that it is a good policy to make inferences about others’ types and to condition future reciprocal behavior based on these inferences (Roberts et al., 2021). However, making these inferences correctly might be too complex in most scenarios, leading us to rely on simpler heuristics that provide reasonable proxies for types. A simple heuristic based on observed intentions could be more effective and less susceptible to errors than complex inferences about motivations (Haselton et al., 2009). That is, it’s possible that people internally do care about types, rather than intentions, but because our inferences about types are limited in their sophistication and do not seem to be affected by probabilistic observability, models of intentions are more consistent with actual behavior. Another related possibility is that people prefer to grant the benefit of the doubt about unobserved behavior: people may infer that someone

who is observed being nice in public is likely to shirk when they get the chance, but without proof, they may wish to give the benefit of the doubt by reciprocating equivalently as to someone who has sent a stronger signal by being nice in quasi-private.

Another possible interpretation is based on social norms. If indirect reciprocity is governed by norms, it is more likely to be built on discrete categorizations rather than on whether a continuous variable (inferences about altruism levels) meets a threshold (Yoeli et al., 2022). For instance, a social norm for Agents to choose HH in the 4-player game may be supported by a norm for Observers to reward HH and HU, punish NN and NU, and to penalize HN because HN is seen as an attempt to exploit Observer reciprocity towards HU. Exploring these norm-based explanations and how they compare to type-based models is an important direction for future research.

Our results also do not rule out the importance of type-based preferences in other settings. While our games reflect typical models of indirect reciprocity where the Observer engages in a one-shot interaction with the Agent, many real-world settings involve repeated interactions and the opportunity to select partners. These *partner choice* situations may justify caring for the inner pro-social preferences potential partners have displayed in the past as a guarantee that selected partners will reliably act cooperatively in future interactions instead of possibly opting for uncooperative strategies if there happen to be material incentives to do so (Baumard, André and Sperber, 2013; Bliege Bird, Ready and Power, 2018; Ågren, Davies and Foster, 2019). Previous studies have argued for the importance of type signaling in settings involving partner choice (Barclay and Willer, 2007; Sylwester and Roberts, 2010; Fehrler and Przepiorka, 2016); however these results are also consistent with intentions-based preferences as we conceptualize them in the current study. Extending our experimental paradigm to a partner choice setting is a promising direction to test whether type-based preferences play a larger role in those environments.

In conclusion, our results indicate that type-based preferences are not a dominant driver of indirect reciprocity in our experimental setting. Instead, we find that intention-based and outcome-based preferences seem to play a more central role. Nevertheless, type-based preferences may still matter in settings beyond the scope of our experiment, such as real-world environments with partner choice. An important direction for future work is to further clarify the conditions influencing the relevance of type-based preferences.

5.2 Guile aversion

The results from the 4-player game experiment point to an additional factor that current models don't account for. If Observers are driven by intentions-based indirect reciprocity, as our other results indicate, then Observers should treat NH and HN identically since both entail an attempt to

help exactly one Recipient. Figure 7 shows that, instead, Observers consistently reciprocate more towards NH, regardless of visibility. Observers reward HN at rates of 22% and 18% at low and high observability respectively, which are each lower than the respective rates of 33% and 28% for NH. These differences are both statistically significant (two sample Z-test p -values < 0.05).

Table 4 introduces a dummy variable named “guile”, taking a value of 1 when an Observer encounters HN, and 0 otherwise, to the previous regression analysis in Table 1 which is duplicated with controls in column 1. Guile has an average marginal effect of 7.4% in column 2, which is roughly half the magnitude of the effect of a helpful outcome. This result points towards Observers penalizing Agents who are seen to help publicly but who avoid doing so in more private settings (i.e., HN), which we interpret as Observers punishing what they perceive as guileful.

This finding contributes to the discourse on strategic reputation building. Previous work by Engelmann and Fischbacher (2009) found that 25% of their participants in a repeated helping game were purely strategic, choosing to help publicly but never in private. These strategic types dominated their sessions, receiving 1.23 times the average session payoff (while weakly or non-strategic reciprocal players received 0.69 times the session average). Higher-order information was posited as a potential remedy for the strategic behavior of these types and as an explanation for how these types could coexist with non-strategic types in the long run. In our experiment, information about multiple one-shot decisions plays a similar role to higher-order information in a repeated game, and this is available to the Observer when the quasi-private interaction is revealed. Observers can then distinguish consistently helpful players (HH) from their strategic counterparts (HN). The penalty assigned specifically to HN suggests that Observers are using the additional information to punish Agents who are being purely strategic, which is broadly consistent with Engelmann and Fischbacher’s (2009) predictions.

However, the precise nature of the guile aversion we observed raises some questions. Notice that the intuition that people may be disinclined to reward strategic altruism is already built into type-based reciprocity: according to type-based reciprocity, if someone’s helpfulness is not motivated by pure altruism, they should be rewarded less than a pure altruist.¹³ This is exactly what we reject with our 3-player game results, and yet the 4-player game shows that Observers who can *confirm* that someone is only altruistic in public, not just infer it with high likelihood, do in fact punish that inconsistency. This insensitivity to probabilistic inferences, in lieu of exclusive focus on observed choices, extends so far that Observers favor NH over HN even when p is high, that is when HN is so likely to be “caught” that it can hardly be considered guileful. This is shown in

¹³ There are a few studies that claim to demonstrate this effect in directly reciprocal interactions (Lin and Ong, 2011; Stanca, Bruni and Corazzini, 2009; Johnsen and Kvaløy, 2016) but these results are also consistent with intentions-based direct reciprocity.

Dependent Variable: Observer helped Agent			
Case Observed			
	(1)	(2)	(3)
Helpful Outcome	0.5093*** (0.0877) [0.1471]	0.5289*** (0.0921) [0.1523]	0.5284*** (0.0924) [0.1520]
Intentions	0.7756*** (0.0730) [0.2101]	0.7285*** (0.0692) [0.1950]	0.7288*** (0.0694) [0.1949]
Guile		-0.2779*** (0.0727) [-0.0740]	-0.2235** (0.0873) [-0.0596]
High observability			-0.0542 (0.0389)
Guile × High observability			-0.1135 (0.0888)
Observations	2,368	2,368	2,368
Demographic Controls	Yes	Yes	Yes

Table 4: Probit regression analysis of Observer reciprocation rates in the 4-player game after witnessing HH, HN, NH, or NN. Standard errors are clustered by participant. Demographic controls include gender, international student status, English as a first language, and previous training in economics. Average marginal effects are shown in square brackets. Statistical significance indicated at * 10%, ** 5%, and *** 1% levels.

Table 4 Column 3 which shows that guile does not substantially vary with observability.

A possible explanation for this guile aversion builds on one interpretation of our main results discussed above: sophisticated Type inferences involving probabilistic observability are not consistent with behavior, but people nonetheless internally care about others' types and use observed intentions to judge them. Intentions-based preferences are therefore broadly consistent with the data. Provably strategically inconsistent behavior (HN) is additionally easily enough understood to be a product of insincere altruism, and is therefore punished relative to other forms of inconsistency that nonetheless embody the same kind intentions (NH).

Our finding of guile aversion has connections with a nascent literature on an aversion to inauthentic behavior like hypocrisy. Jordan et al. (2017) proposed that hypocrisy is disliked because of "false signaling". If someone condemns an immoral behavior but then behaves immorally themselves, it's considered a more significant betrayal than merely lying about one's behavior. Several patterns of judgements predicted by this theory were supported in a series of survey studies. Relatedly, Koehler and Gershoff (2003) reported that people react more negatively to betrayal than to bad outcomes alone. Future work will be required to fully understand the complex strategies deployed by people when they are engaged in repeated interactions featuring opportunities to cooperate or help others.

6 Conclusion

This paper explores the underlying motivations for indirect reciprocity, specifically investigating whether individuals reward good deeds or good people. To do this, we developed a new experimental framework that distinguishes between the effects of the actions that an agent takes, the motives behind them, and the resulting outcomes, on future indirect reciprocity. We observe extensive indirect reciprocity, but, surprisingly, find that type-based preferences are unable to explain it. Although our Observers could detect the strategic motives behind the Agents' helpful actions, these perceptions did not influence the Observers' willingness to reciprocate such behaviors.

The main drivers of indirect reciprocity instead are the outcomes of past helping decisions by the Agents, even when these outcomes were explicitly random, and the intentions of the Agents, even when these did not lead to positive outcomes. We therefore conclude that people tend to reward good deeds rather than inherently good people. This result is particularly striking as it challenges the prevailing assumption underpinning type-based preferences: that people are motivated to reward individuals perceived as good. Given the widespread belief that type-based preferences are well-suited to explain indirect reciprocity, this finding represents a significant departure from

established thinking.

We also identify a specific aversion to rewarding guileful behavior that was successfully “caught” by the Observer. In our experiments, participants who helped in public but chose not to when their actions could be unobservable received significantly less help from the Observers than those who were equally helpful but whose helpfulness was less easily observable. This effect, which we label as “guile aversion,” raises a critical question: how should people react towards those who not only engage in cooperative behavior without genuine altruistic preferences but also without the intention to continue doing so when their actions are not fully observed?

Our findings shed a new light on the psychological mechanisms that sustain indirect reciprocity within large groups. While type-based preferences seem naturally suited to maintaining such an equilibrium, they may not be required. Instead, this equilibrium might rely on people valuing others reputation for doing the right thing, without necessarily caring about whether this reputation reflects true altruism or strategic motives. Our results suggest that the good and bad intentions displayed towards previous other partners are an important factor that future models of the motivations for indirect reciprocity should account for.

While we find no evidence for type-based preferences in our setting, this does not preclude their possible role in other types of social interactions. In scenarios where individuals select partners for medium to long-term relationships—such as coworkers, friends, or mates—choosing partners with prosocial types might be the optimal strategy. This selection may ensure that partners will do the right thing even in future situations where they face minimal scrutiny and might benefit from defecting from cooperation. Such preferences for the reputation of one’s social partners likely vary depending on the nature of the interaction. For transient interactions, valuing a reputation for doing the right thing might suffice to sustain cooperation, but in long-term partnerships, the intrinsic nature of individuals may become more significant. These distinctions open avenues for future research to explore more deeply why we value others’ reputations differently across various social contexts.

References

- Ågren, J Arvid, Nicholas G Davies, and Kevin R Foster. 2019. “Enforcement is central to the evolution of cooperation.” *Nature Ecology & Evolution*, 3(7): 1018–1029. DOI: <https://doi.org/10.1038/s41559-019-0907-1>.
- Alexander, Richard D. 1987. *The Biology of Moral Systems. Foundations of Human Behavior*, Hawthorne, N.Y:A. de Gruyter.

- Andreoni, James, Paul M. Brown, and Lise Vesterlund.** 2002. “What Makes an Allocation Fair? Some Experimental Evidence.” *Games and Economic Behavior*, 40(1): 1–24. DOI: <https://doi.org/10.1006/game.2001.0904>.
- Balafoutas, Loukas, Kristoffel Grechenig, and Nikos Nikiforakis.** 2014. “Third-Party Punishment and Counter-Punishment in One-Shot Interactions.” *Economics Letters*, 122(2): 308–310. DOI: <https://doi.org/10.1016/j.econlet.2013.11.028>.
- Balafoutas, Loukas, Nikos Nikiforakis, and Bettina Rockenbach.** 2014. “Direct and Indirect Punishment among Strangers in the Field.” *Proceedings of the National Academy of Sciences*, 111(45): 15924–15927. DOI: <https://doi.org/10.1073/pnas.1413170111>.
- Barclay, Pat, and Robb Willer.** 2007. “Partner Choice Creates Competitive Altruism in Humans.” *Proceedings of the Royal Society B: Biological Sciences*, 274(1610): 749–753. DOI: <https://doi.org/10.1098/rspb.2006.0209>.
- Battigalli, Pierpaolo, and Martin Dufwenberg.** 2009. “Dynamic Psychological Games.” *Journal of Economic Theory*, 144(1): 1–35. DOI: <https://doi.org/10.1016/j.jet.2008.01.004>.
- Baumard, Nicolas, Jean-Baptiste André, and Dan Sperber.** 2013. “A Mutualistic Approach to Morality: The Evolution of Fairness by Partner Choice.” *Behavioral and Brain Sciences*, 36(1): 59–78. DOI: <https://doi.org/10.1017/S0140525X11002202>.
- Binmore, Kenneth G.** 2005. *Natural Justice*. , Oxford:Oxford University Press.
- Bliege Bird, Rebecca, Elspeth Ready, and Eleanor A. Power.** 2018. “The Social Significance of Subtle Signals.” *Nature Human Behaviour*, 2: 452–457. DOI: <https://doi.org/10.1038/s41562-018-0298-3>.
- Blount, Sally.** 1995. “When Social Outcomes Aren’t Fair: The Effect of Causal Attributions on Preferences.” *Organizational Behavior and Human Decision Processes*, 63(2): 131–144. DOI: <https://doi.org/10.1006/obhd.1995.1068>.
- Bolton, Gary E., and Axel Ockenfels.** 2000. “ERC: A Theory of Equity, Reciprocity, and Competition.” *American Economic Review*, 90(1): 166–193. DOI: <https://doi.org/10.1257/aer.90.1.166>.
- Bowles, Samuel, and Herbert Gintis.** 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton University Press.

- Brandts, Jordi, and Carles Solà.** 2001. “Reference Points and Negative Reciprocity in Simple Sequential Games.” *Games and Economic Behavior*, 36(2): 138–157. DOI: <https://doi.org/10.1006/game.2000.0818>.
- Çelen, Boğaçhan, Andrew Schotter, and Mariana Blanco.** 2017. “On Blame and Reciprocity: Theory and Experiments.” *Journal of Economic Theory*, 169: 62–92. DOI: <https://doi.org/10.1016/j.jet.2017.01.006>.
- Charness, Gary, and David I. Levine.** 2007. “Intention and Stochastic Outcomes: An Experimental Study.” *The Economic Journal*, 117(522): 1051–1072. DOI: <https://doi.org/10.1111/j.1468-0297.2007.02066.x>.
- Charness, Gary B.** 2004. “Attribution and Reciprocity in an Experimental Labor Market.” *Journal of Labor Economics*, 22(3): 665–688. DOI: <https://doi.org/10.1086/383111>.
- Charness, Gary B., and Matthew Rabin.** 2002. “Understanding Social Preferences with Simple Tests.” *Quarterly Journal of Economics*, 117(3): 817–869. DOI: <https://doi.org/10.1162/003355302760193904>.
- Chen, Daniel L, Martin Schonger, and Chris Wickens.** 2016. “oTree—An open-source platform for laboratory, online, and field experiments.” *Journal of Behavioral and Experimental Finance*, 9: 88–97. DOI: <https://doi.org/10.1016/j.jbef.2015.12.001>.
- Cho, In-Koo, and David M. Kreps.** 1987. “Signaling Games and Stable Equilibria.” *Quarterly Journal of Economics*, 102(2): 179–221.
- Cox, James C.** 2004. “How to Identify Trust and Reciprocity.” *Games and Economic Behavior*, 46: 260–281. DOI: [https://doi.org/10.1016/S0899-8256\(03\)00119-2](https://doi.org/10.1016/S0899-8256(03)00119-2).
- Dufwenberg, Martin, and Georg Kirchsteiger.** 2004. “A Theory of Sequential Reciprocity.” *Games and Economic Behavior*, 47(2): 268–298. DOI: <https://doi.org/10.1016/j.geb.2003.06.003>.
- Eckel, Catherine C., and Philip J. Grossman.** 1996. “The Relative Price of Fairness: Gender Differences in a Punishment Game.” *Journal of Economic Behavior & Organization*, 30(2): 143–158. DOI: [https://doi.org/10.1016/S0167-2681\(96\)00854-2](https://doi.org/10.1016/S0167-2681(96)00854-2).
- Engelmann, Dirk, and Urs Fischbacher.** 2009. “Indirect Reciprocity and Strategic Reputation Building in an Experimental Helping Game.” *Games and Economic Behavior*, 67(2): 399–407. DOI: <https://doi.org/10.1016/j.geb.2008.12.006>.

- Falk, Armin, and Urs Fischbacher.** 2006. "A Theory of Reciprocity." *Games and Economic Behavior*, 54(2): 293–315. DOI: <https://doi.org/10.1016/j.geb.2005.03.001>.
- Falk, Armin, Ernst Fehr, and Urs Fischbacher.** 2003. "On the Nature of Fair Behavior." *Economic Inquiry*, 41(1): 20–26. DOI: <https://doi.org/10.2139/ssrn.203289>.
- Falk, Armin, Ernst Fehr, and Urs Fischbacher.** 2008. "Testing Theories of Fairness - Intentions Matter." *Games and Economic Behavior*, 62(1): 287–303. DOI: <https://doi.org/10.1016/j.geb.2007.06.001>.
- Fehr, Ernst, and Klaus M. Schmidt.** 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114(3): 817–868.
- Fehr, Ernst, and Simon Gächter.** 2000. "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, 14(3): 159–182. DOI: <https://doi.org/10.1257/jep.14.3.159>.
- Fehr, Ernst, and Urs Fischbacher.** 2004. "Third-Party Punishment and Social Norms." *Evolution and Human Behavior*, 25(2): 63–87. DOI: [https://doi.org/10.1016/S1090-5138\(04\)00005-4](https://doi.org/10.1016/S1090-5138(04)00005-4).
- Fehrler, Sebastian, and Wojtek Przepiorka.** 2016. "Choosing a Partner for Social Exchange: Charitable Giving as a Signal of Trustworthiness." *Journal of Economic Behavior & Organization*, 129: 157–171. DOI: <https://doi.org/10.1016/j.jebo.2016.06.006>.
- Fessler, Daniel M. T.** 2002. "Windfall and Socially Distributed Willpower: The Psychocultural Dynamics of Rotating Savings and Credit Associations in a Bengkulu Village." *Ethos*, 30(1-2): 25–48. DOI: <https://doi.org/10.1525/eth.2002.30.1-2.25>.
- Geanakoplos, John, David Pearce, and Ennio Stacchetti.** 1989. "Psychological Games and Sequential Rationality." *Games and Economic Behavior*, 1(1): 60–79. DOI: [https://doi.org/10.1016/0899-8256\(89\)90005-5](https://doi.org/10.1016/0899-8256(89)90005-5).
- Greif, Avner.** 1993. "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition." *The American Economic Review*, 83(3): 525–548.
- Greif, Avner.** 1994. "Cultural Beliefs and the Organization of Society: A Historical and Theoretical Reflection on Collectivist and Individualist Societies." *Journal of Political Economy*, 102(5): 912. DOI: <https://doi.org/10.1086/261959>.

- Gul, Faruk, and Wolfgang Pesendorfer.** 2016. “Interdependent Preference Models as a Theory of Intentions.” *Journal of Economic Theory*, 165: 179–208. DOI: <https://doi.org/10.1016/j.jet.2016.04.009>.
- Güth, Werner, Manfred Königstein, Marchand Nadège, and Klaus Nehring.** 2001. “Trust and Reciprocity in the Investment Game with Indirect Reward.” *Homo Oeconomicus*, 18: 241–262.
- Haselton, Martie G., Gregory A. Bryant, Andreas Wilke, David A. Frederick, Andrew Galperin, Willem E. Frankenhuys, and Tyler Moore.** 2009. “Adaptive Rationality: An Evolutionary Perspective on Cognitive Bias.” *Social Cognition*, 27(5): 733–763. DOI: <https://doi.org/10.1521/soco.2009.27.5.733>.
- Herne, Kaisa, Olli Lappalainen, and Elina Kestilä-Kekkonen.** 2013. “Experimental Comparison of Direct, General, and Indirect Reciprocity.” *The Journal of Socio-Economics*, 45: 38–46. DOI: <https://doi.org/10.1016/j.socec.2013.04.003>.
- Johnsen, Åshild A., and Ola Kvaløy.** 2016. “Does Strategic Kindness Crowd out Prosocial Behavior?” *Journal of Economic Behavior & Organization*, 132: 1–11. DOI: <https://doi.org/10.1016/j.jebo.2016.09.016>.
- Jordan, Jillian J., Roseanna Sommers, Paul Bloom, and David G. Rand.** 2017. “Why Do We Hate Hypocrites? Evidence for a Theory of False Signaling.” *Psychological Science*, 28(3): 356–368. DOI: <https://doi.org/10.1177/0956797616685771>.
- Kagel, John H., Chung Kim, and Donald Moser.** 1996. “Fairness in Ultimatum Games with Asymmetric Information and Asymmetric Payoffs.” *Games and Economic Behavior*, 13(1): 100–110. DOI: <https://doi.org/10.1006/game.1996.0026>.
- Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler.** 1986. “Fairness as a Constraint on Profit Seeking: Entitlements in the Market.” *American Economic Review*, 76(4): 728–741.
- Kandori, Michihiro.** 1992. “Social Norms and Community Enforcement.” *The Review of Economic Studies*, 59(1): 63–80. DOI: <https://doi.org/10.2307/2297925>.
- Khadjavi, Menusch.** 2017. “Indirect Reciprocity and Charitable Giving: Evidence from a Field Experiment.” *Management Science*, 63(11): 3708–3717. DOI: <https://doi.org/10.1287/mnsc.2016.2519>.

- Klempt, Charlotte.** 2012. "Fairness, Spite, and Intentions: Testing Different Motives behind Punishment in a Prisoners' Dilemma Game." *Economics Letters*, 116(3): 429–431. DOI: <https://doi.org/10.1016/j.econlet.2012.04.048>.
- Koehler, Jonathan J, and Andrew D Gershoff.** 2003. "Betrayal Aversion: When Agents of Protection Become Agents of Harm." *Organizational Behavior and Human Decision Processes*, 90(2): 244–261. DOI: [https://doi.org/10.1016/S0749-5978\(02\)00518-6](https://doi.org/10.1016/S0749-5978(02)00518-6).
- Levine, David K.** 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics*, 1(3): 593–622. DOI: <https://doi.org/10.1006/redy.1998.0023>.
- Lin, Hong (Hannah), and David Ong.** 2011. "Separating Gratitude from Guilt in the Laboratory." SSRN Working Paper 1943949.
- Mailath, George J, and Larry Samuelson.** 2006. *Repeated Games and Reputations: Long-run Relationships*. Oxford University Press.
- Mathew, Sarah, and Robert Boyd.** 2011. "Punishment Sustains Large-Scale Cooperation in Prestate Warfare." *Proceedings of the National Academy of Sciences*, 108(28): 11375–11380. DOI: <https://doi.org/10.1073/pnas.1105604108>.
- McCabe, Kevin A., Mary L. Rigdon, and Vernon L. Smith.** 2003. "Positive Reciprocity and Intentions in Trust Games." *Journal of Economic Behavior & Organization*, 52(2): 267–275. DOI: [https://doi.org/10.1016/S0167-2681\(03\)00003-9](https://doi.org/10.1016/S0167-2681(03)00003-9).
- Milinski, Manfred, Dirk Semmann, and Hans-Jürgen Krambeck.** 2002. "Reputation Helps Solve the 'Tragedy of the Commons'." *Nature*, 415(6870): 424–426. DOI: <https://doi.org/10.1038/415424a>.
- Nelson, William Robert.** 2002. "Equity or Intention: It Is the Thought That Counts." *Journal of Economic Behavior & Organization*, 48(4): 423–430. DOI: [https://doi.org/10.1016/S0167-2681\(01\)00245-1](https://doi.org/10.1016/S0167-2681(01)00245-1).
- Nowak, Martin A., and Karl Sigmund.** 1998. "Evolution of Indirect Reciprocity by Image Scoring." *Nature*, 393(6685): 573–577.
- Nowak, Martin A., and Karl Sigmund.** 2005. "Evolution of Indirect Reciprocity." *Nature*, 437(October): 1291–11298. DOI: <https://doi.org/10.1038/31225>.

- Offerman, Theo.** 2002. “Hurting Hurts More than Helping Helps.” *European Economic Review*, 46(8): 1423–1437. DOI: [https://doi.org/10.1016/S0014-2921\(01\)00176-3](https://doi.org/10.1016/S0014-2921(01)00176-3).
- Okada, Isamu.** 2020. “A Review of Theoretical Studies on Indirect Reciprocity.” *Games*, 11(3). DOI: <https://doi.org/10.3390/g11030027>.
- Orhun, A. Yeşim.** 2018. “Perceived Motives and Reciprocity.” *Games and Economic Behavior*, 109: 436–451. DOI: <https://doi.org/10.1016/j.geb.2018.01.002>.
- Rabin, Matthew.** 1993. “Incorporating Fairness into Game Theory and Economics.” *American Economic Review*, 83(5): 1281–1302.
- Roberts, Gilbert, Nichola Raihani, Redouan Bshary, Héctor M. Manrique, Andrea Farina, Flóra Samu, and Pat Barclay.** 2021. “The Benefits of Being Seen to Help Others: Indirect Reciprocity and Reputation-Based Partner Choice.” *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376: 20200290. DOI: <https://doi.org/10.1098/rstb.2020.0290>.
- Rotemberg, Julio J.** 2008. “Minimally Altruistic Wages and Unemployment in a Matching Model.” Working Paper.
- Rutte, Christel G., Henk A. M. Wilke, and David M. Messick.** 1987. “Scarcity or Abundance Caused by People or the Environment as Determinants of Behavior in the Resource Dilemma.” *Journal of Experimental Social Psychology*, 23(3): 208–216. DOI: [https://doi.org/10.1016/0022-1031\(87\)90032-1](https://doi.org/10.1016/0022-1031(87)90032-1).
- Sebald, Alexander.** 2010. “Attribution and Reciprocity.” *Games and Economic Behavior*, 68(1): 339–352. DOI: <https://doi.org/10.1016/j.geb.2009.07.004>.
- Seinen, Ingrid, and Arthur Schram.** 2006. “Social Status and Group Norms: Indirect Reciprocity in a Repeated Helping Experiment.” *European Economic Review*, 50(3): 581–602. DOI: <https://doi.org/10.1016/j.euroecorev.2004.10.005>.
- Semmann, Dirk, Hans-Jürgen Krambeck, and Manfred Milinski.** 2004. “Strategic Investment in Reputation.” *Behavioral Ecology and Sociobiology*, 56(3): 248–252. DOI: <https://doi.org/10.1007/s00265-004-0782-9>.
- Servátka, Maroš.** 2009. “Separating Reputation, Social Influence, and Identification Effects in a Dictator Game.” *European Economic Review*, 53(2): 197–209. DOI: <https://doi.org/10.1016/j.euroecorev.2008.04.001>.

- Simpson, Brent, Ashley Harrell, David Melamed, Nicholas Heiserman, and Daniela V. Negraia.** 2018. “The Roots of Reciprocity: Gratitude and Reputation in Generalized Exchange Systems.” *American Sociological Review*, 83(1): 88–110. DOI: <https://doi.org/10.1177/0003122417747290>.
- Stanca, Luca.** 2009. “Measuring Indirect Reciprocity: Whose Back Do We Scratch?” *Journal of Economic Psychology*, 30(2): 190–202. DOI: <https://doi.org/10.1016/j.joep.2008.07.010>.
- Stanca, Luca, Luigino Bruni, and Luca Corazzini.** 2009. “Testing Theories of Reciprocity: Do Motivations Matter?” *Journal of Economic Behavior & Organization*, 71(2): 233–245. DOI: <https://doi.org/10.1016/j.jebo.2009.04.009>.
- Sugden, Robert.** 1986. *The Economics of Rights, Co-operation and Welfare*. . 2nd ed., New York: Palgrave Macmillan.
- Sylwester, Karolina, and Gilbert Roberts.** 2010. “Cooperators Benefit through Reputation-Based Partner Choice in Economic Games.” *Biology Letters*, 6(5): 659–662. DOI: <https://doi.org/10.1098/rsbl.2010.0209>.
- Turillo, Carmelo Joseph, Robert Folger, James J Lavelle, Elizabeth E Umphress, and Julie O Gee.** 2002. “Is Virtue Its Own Reward? Self-sacrificial Decisions for the Sake of Fairness.” *Organizational Behavior and Human Decision Processes*, 89(1): 839–865. DOI: [https://doi.org/10.1016/S0749-5978\(02\)00032-8](https://doi.org/10.1016/S0749-5978(02)00032-8).
- van Apeldoorn, Jacobien, and Arthur Schram.** 2016. “Indirect Reciprocity: A Field Experiment.” *PLOS ONE*, 11(4): e0152076. DOI: <https://doi.org/10.1371/journal.pone.0152076>.
- Wedekind, Claus, and Manfred Milinski.** 2000. “Cooperation Through Image Scoring in Humans.” *Science*, 288(5467): 850–852. DOI: <https://doi.org/10.1126/science.288.5467.850>.
- Wedekind, Claus, and Victoria A Braithwaite.** 2002. “The Long-Term Benefits of Human Generosity in Indirect Reciprocity.” *Current Biology*, 12(12): 1012–1015. DOI: [https://doi.org/10.1016/S0960-9822\(02\)00890-4](https://doi.org/10.1016/S0960-9822(02)00890-4).
- Yoeli, Erez, N. Aygun Dalkiran, Bethany A. Burum, Martin A. Nowak, and Moshe Hoffman.** 2022. “Categorical Norms.” Working Paper.

A Proofs

A.1 Proof of Proposition 1

As described in Section 2.2, the Observer's net gain from helping is $(\alpha_O + \lambda E[\alpha_A])b - c$, and so they choose to help if $\alpha_O > c/b - \lambda E[\alpha_A]$. When nothing is known about the Agent, $E[\alpha_A] = \bar{\alpha} = A/2$. If the Observer observes H or N , they more precisely infer that α_A is above or below the threshold A_H , with expected values $\frac{A+A_H}{2}$ and $\frac{A_H}{2}$ respectively. This provides three equations that define the cutoff values of α_O above which Observers help given what they observe, which must be satisfied in a PBE:

$$\begin{aligned}O_U &= \frac{c}{b} - \lambda \frac{A}{2} \\O_H &= \frac{c}{b} - \lambda \frac{A + A_H}{2} \\O_N &= \frac{c}{b} - \lambda \frac{A_H}{2}\end{aligned}$$

Note that so long as $A > A_H > 0$, which we show below, it is true that $O_H < O_U < O_N$.

The final equation that must be satisfied in a PBE defines A_H by balancing the expected utilities of helping versus not that the Agent faces, as stated in Section 2.2:

$$A_H = \frac{c}{b} + p \frac{O_H - O_N}{A} - \lambda \frac{A}{2}$$

Altogether we have four equations in four unknown cutoff values that define the Agent's and Observer's equilibrium strategies. Solving this system yields the following result, characterizing the PBE:

$$\begin{aligned}
A_H &= \frac{c}{b} - \frac{\lambda}{2}(A + p) \\
O_H &= \frac{c}{b} - \frac{\lambda}{2b}(c + Ab) + \frac{\lambda^2}{4}(A + p) \\
O_N &= \left(1 - \frac{\lambda}{2}\right) \frac{c}{b} + \frac{\lambda^2}{4}(A + p) \\
O_U &= \frac{c}{b} - \frac{\lambda}{2}A
\end{aligned}$$

To be a PBE, this strategy profile must satisfy: 1) The Agent's strategy is a best response to the Observer's strategy, 2) The Observer's strategy is a best response to the Agent's strategy, given their beliefs about the Agent's type, and 3) The Observer's beliefs about the Agent's type are consistent with the Agent's strategy and are updated using Bayes' rule whenever possible. These conditions are satisfied by construction; however, the analysis above assumes that all cutoff values are strictly between 0 and A when calculating Bayesian posteriors and the Agent's expected return from reciprocation. We must show that this is indeed the case given our assumption (described in the main text) that $A_{\min} < A < A_{\max}$. In fact, as we have not yet defined A_{\min} and A_{\max} precisely, we can do so to ensure that it is the case.

On the lower end, we need $A_H > 0$ and $\min\{O_H, O_U, O_N\} = O_H > 0$. That is,

$$\frac{c}{b} - \frac{\lambda}{2}(A + p) > 0 \Leftrightarrow \frac{2c}{\lambda b} - p > A$$

and

$$\frac{c}{b} - \frac{\lambda}{2}(A + A_H) > 0 \Leftrightarrow \frac{c}{\lambda b} > A$$

due to the fact that we will require $A_H < A$ with the condition below.

On the upper end, we similarly need $A_H, O_N < A$. That is,

$$\frac{c}{b} - \frac{\lambda}{2}(A + p) < A \Leftrightarrow \frac{2}{2 + \lambda b} \frac{c}{b} - \frac{\lambda p}{2 + \lambda} < A$$

and

$$\frac{c}{b} - \lambda \frac{A_H}{2} < A \Leftrightarrow \frac{c}{b} < A$$

due to the fact that we require $A_H > 0$ with the condition above. Note that the latter condition

implies the former, and so altogether we need

$$\frac{c}{b} < A < \min \left\{ \frac{2c}{\lambda b} - p, \frac{c}{\lambda b} \right\}.$$

In our experimental context, $c = 1$ and $b = 2.5$ and $p \in \{0.1, 0.9\}$. We can therefore define $A_{\min} = \frac{c}{b}$ and $A_{\max} = \min \left\{ \frac{2c}{\lambda b} - p, \frac{c}{\lambda b} \right\}$ and A_{\max} will in fact exceed A_{\min} so long as λ is small enough, which is a reasonable assumption since we expect people to put a greater weight on their own altruism rather than others' altruism. Therefore we restrict our analysis to the situation in which $A_{\min} < A < A_{\max}$ and when λ is small enough that such an A exists, and then the PBE characterized in this proposition is accordingly predicted to exist.

A.2 Proof of Proposition 2

The comparative statics in Proposition 2 follow directly from the equilibrium cutoffs derived in the proof of Proposition 1. We have:

$$\begin{aligned} \frac{\partial A_H}{\partial p} &= -\frac{\lambda}{2} < 0 \\ \frac{\partial O_H}{\partial p} &= \frac{\partial O_N}{\partial p} = \frac{\lambda^2}{4} \geq 0 \end{aligned}$$

The intuition for these results is as described in Section 2.2. The higher the observability, the more opportunities there are for reciprocity, and so Agents with lower altruism are drawn to helping. As p increases, the average altruism of both helpful and unhelpful Agents decreases, leading to a lower probability of reciprocation for both H and N actions when p is high compared to when it is low.

A.3 Proof of Proposition 3

To establish the form of possible equilibria in the 4-player game, consider the Agent's choice between options HH , HN , NH , and NN . The expected utilities of each of these options are:

$$\begin{aligned}
v_A(NN) &= pbP(\alpha_O > O_{NN}) + (1-p)bP(\alpha_O > O_{NU}) \\
&= pb\frac{A - O_{NN}}{A} + (1-p)b\frac{A - O_{NU}}{A} = b\left(1 - p\frac{O_{NN}}{A} - (1-p)\frac{O_{NU}}{A}\right) \\
v_A(NH) &= pbP(\alpha_O > O_{NH}) + (1-p)bP(\alpha_O > O_{NU}) + \frac{(\alpha_A + \lambda E[\alpha_{R_2}])b - c}{2} \\
&= b\left(1 - p\frac{O_{NH}}{A} - (1-p)\frac{O_{NU}}{A} + \frac{\alpha_A + \lambda\frac{A}{2}}{2}\right) - \frac{c}{2} \\
v_A(HN) &= pbP(\alpha_O > O_{HN}) + (1-p)bP(\alpha_O > O_{HU}) + \frac{(\alpha_A + \lambda E[\alpha_{R_1}])b - c}{2} \\
&= b\left(1 - p\frac{O_{HN}}{A} - (1-p)\frac{O_{HU}}{A} + \frac{\alpha_A + \lambda\frac{A}{2}}{2}\right) - \frac{c}{2} \\
v_A(HH) &= pbP(\alpha_O > O_{HH}) + (1-p)bP(\alpha_O > O_{HU}) - c + \frac{(\alpha_A + \lambda E[\alpha_{R_1}])b}{2} + \frac{(\alpha_A + \lambda E[\alpha_{R_2}])b}{2} \\
&= b\left(1 - p\frac{O_{HH}}{A} - (1-p)\frac{O_{HU}}{A} + \alpha_A + \lambda\frac{A}{2}\right) - c
\end{aligned}$$

First note that the Agent prefers $HN \succ NH$ so long as $p(O_{HN} - O_{NH}) < (1-p)(O_{NU} - O_{HU})$, which is either true or false for all Agents. Therefore either HN or NH is chosen in equilibrium, but not both, except in knife-edge cases. Also, NN is preferred to all other options as long as α_A is sufficiently small¹⁴, and HH is preferred to all other options as long as α_A is sufficiently high, so if the support of α_i includes these high and low types, then equilibrium must consist of some types choosing NN , some choosing HH , and *potentially* some mid-level types choosing either HN or NH (but not both). Similarly to our analysis of the 3PG, we restrict our analysis to the distributions of α_i that satisfy this condition (see below for more details on what assumptions about A this requires.)

We can furthermore eliminate NH as a possibility by appealing to the D1 criterion, a standard equilibrium refinement in signaling games. Intuitively, this criterion assumes that when an observer sees an unexpected action, they should believe it was taken by the type of player who would be most likely to benefit from this deviation. In our setting, this means that if an observer sees an agent choose NH (which is not predicted in equilibrium), they should believe that this agent is most likely of the type that would gain the most from this action, compared to their equilibrium

¹⁴Or from another perspective, if O_{HH} is not too low.

payoff. Applying the D1 criterion allows us to rule out equilibria that rely on “unreasonable” beliefs and focus on the most plausible outcomes.

Suppose that some equilibrium did exist in which NN is chosen when $\alpha_A < A'_{NH}$, NH is chosen when $A'_{NH} < \alpha_A < A'_{HH}$, and HH is chosen when $\alpha_A > A'_{HH}$. The D1 criterion requires that, upon observing HN , the Observer must attribute it to the types who are tempted to deviate from the equilibrium to that option for the widest possible range of values $E_O[\alpha_A|HN]$.

First consider someone who, in this equilibrium, is choosing NN . In order to switch to HN they would require $v_A(HN) > v_A(NN)$, which according to the expressions above, is satisfied for the widest range of possible Observer inferences (i.e. O_{HN}) when α_A takes on the highest possible value, which in this case is exactly A'_{NH} . Similarly, someone who is choosing HH is tempted to switch to HN when $v_A(HN) > v_A(HH)$, which is true for the widest range of O_{HN} when α_A takes on the lowest possible value, i.e. A'_{HH} . Finally, someone who is choosing NH will switch to HN when $v_A(HN) > v_A(NH) \Leftrightarrow p(O_{HN} - O_{NH}) < (1-p)(O_{NU} - O_{HU})$, which is true for the same values of O_{HN} for all types who are choosing NH . Altogether, the Observer must infer that someone choosing HN is from exactly the set of types who choose NH in equilibrium, so that $E_O[\alpha_A|HN] = E_O[\alpha_A|NH]$. But if this is the case, Agents strictly prefer HN to NH , because the LHS of the previous inequality reduces to 0 and the RHS is strictly positive. This contradicts our assumption that NH is chosen in equilibrium.

Any equilibrium to the 4-player game therefore takes on the form of either a Type 1 or Type 2 equilibrium as described in Section 2.2.

Type 1 Equilibrium: In a Type 1 equilibrium, the cutoffs $(A_{HH}, A_{HN}, O_{HH}, O_{HN}, O_{HU}, O_{NN})$ must satisfy the following system of equations:

$$\begin{aligned}
A_{HH} &\equiv \frac{2p}{A}(O_{HH} - O_{HN}) + \frac{c}{b} - \frac{\lambda A}{2} \\
A_{HN} &\equiv \frac{2p}{A}O_{HN} + \frac{2(1-p)}{A}O_{HU} - \frac{2}{A}O_{NN} + \frac{c}{b} - \frac{\lambda A}{2} \\
O_{HH} &\equiv \frac{c}{b} - \frac{\lambda}{2}(A_{HH} + A) \\
O_{HN} &\equiv \frac{c}{b} - \frac{\lambda}{2}(A_{HH} + A_{HN}) \\
O_{HU} &\equiv \frac{c}{b} - \frac{\lambda}{2}(A_{HN} + A) \\
O_{NN} &\equiv \frac{c}{b} - \frac{\lambda}{2}A_{HN}
\end{aligned}$$

Note that so long as $A_{HH} < A_{HN}$, which must be true for this to be a Type 1 equilibrium, and so long as $A_{HN} < A$ which we show to be true given our assumption about A below, then it follows that $O_{HH} < O_{HU} < O_{HN} < O_{NN}$.

These yield the following closed-form solutions for A_{HH} and A_{HN} :

$$A_{HH} = \frac{A}{2b(A^2 + \lambda^2 p^2)} (2\lambda pc - 2b\lambda^2 p(1-p) - Ab\lambda^2 p - 2Ab\lambda p + 2Ac - \lambda A^2 b)$$

$$A_{HN} = \frac{A}{2b(A^2 + \lambda^2 p^2)} (2b\lambda^2 p^2 - 2\lambda pc + Ab\lambda^2 p - 2Ab\lambda(1-p) + 2Ac - \lambda A^2 b)$$

Closed form solutions for O_{NN} , O_{HN} , O_{HU} and O_{HH} are omitted for brevity but can be straightforwardly calculated using the formulae above and the values for A_{HH} and A_{HN} .

Type 2 Equilibrium:

By construction, and because we restrict attention to equilibria with NN and HH on the equilibrium path as we did in the 3-player game, this solution requires $0 < A_{HN} < A_{HH} < A$ to constitute a valid Type 1 equilibrium. When $A_{HN} \geq A_{HH}$, Type 2 equilibrium replaces Type 1, i.e. when

$$p < \frac{Ab}{(2 + \lambda)Ab + b\lambda - 2c} \equiv \bar{p},$$

When $p = 0$ $A_{HH} - A_{HN} = \lambda > 0$, so to summarize, Type 1 equilibrium exists when $0 < p < \bar{p}$ and Type 2 exists when $p > \bar{p}$.

In a Type 2 equilibrium, the Agent's behavior is defined by a single cutoff value A_{2HH} : if $\alpha_A > A_{2HH}$, the Agent will choose HH , and otherwise will choose NN . The Observer can therefore infer exactly what the Agent chose based on the first, public choice towards Recipient 1, and reciprocates accordingly. The equilibrium is characterized by the following cutoff values:

$$A_{2HH} = \frac{c}{b} - \frac{\lambda}{2}A + \frac{1}{A}(O_{HH} - O_{NN})$$

$$O_{HH} = \frac{c}{b} - \frac{\lambda}{2}(A + A_{2HH})$$

$$O_{NN} = \frac{c}{b} - \frac{\lambda}{2}A_{2HH}$$

This yields the following closed-form solution for A_{2HH} :

$$A_{2HH} = \frac{c}{b} - \frac{\lambda}{2}(A + 1)$$

The remaining caveat to handle is that we have assumed in our calculations of Bayesian posteriors and expected utilities that all derived cutoff values lie strictly between 0 and A in both types of equilibria, under the assumption that A is in an appropriate range of values that ensures this is true. As in the proof of Proposition 1, we need to define the bounds A_{\min} and A_{\max} so that these conditions hold whenever A is restricted to lie strictly between A_{\min} and A_{\max} . Unlike in the 3PG, however, closed-form expressions for these bounds are cumbersome to work with and not obviously descriptive of a valid range. Instead, we illustrate appropriate bounds numerically. Figure 10 serves to define these bounds numerically, as a function of λ . This figure illustrates whether Type 1 or Type 2 equilibrium exists as a function of A , λ , and p , for the setting we use experimentally with $b/c = 2.5$. For small values of λ , A can fall within a wide range of values for either Type 1 or Type 2 equilibrium to exist at any value of p . This range shrinks and disappears when λ rises. Thus, as in the 3PG, we assume that λ is sufficiently small and that A falls within the compatible range.

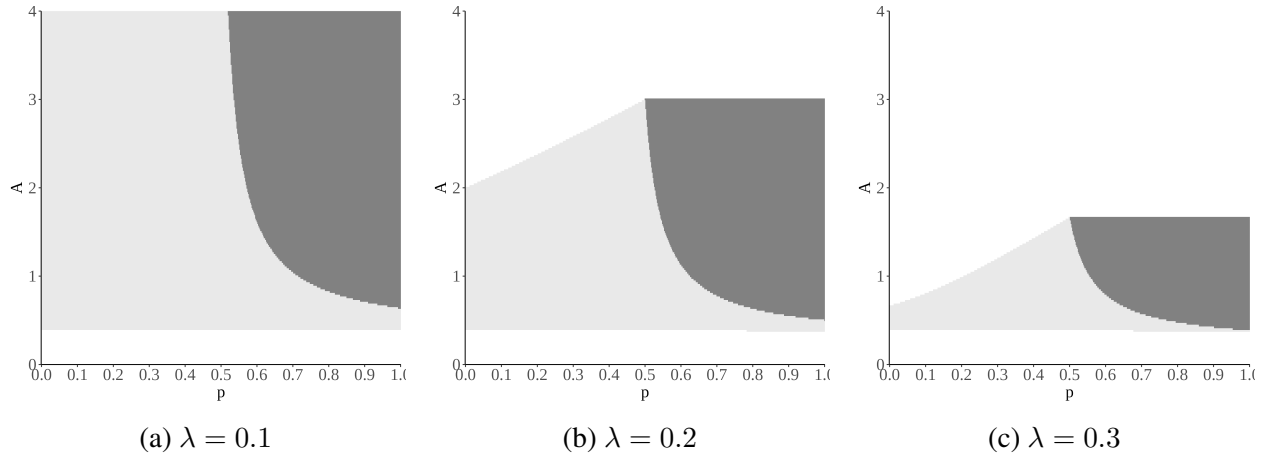


Figure 10: Three cross sections of the regions in (p, A, λ) space for which Type 1 (light grey regions), or only Type 2 (dark grey regions), equilibria exist. When λ is small (left), A can fall within a wide range and one of these two equilibria will always exist (at any level of p). As λ increases, A must fall within a smaller range.

A.4 Proof of Proposition 4

The comparative statics in Proposition 4 follow from the equilibrium cutoffs derived in the proof of Proposition 3.

In a Type 1 equilibrium, reciprocity towards HH is decreasing when HH becomes more prevalent because O_{HH} rises (making it a threshold that is met for fewer Observers) when A_{HH} falls.

In a Type 2 equilibrium, the equilibrium characterization is not dependent on p , so changes in p do not affect the equilibrium as long as p remains above \bar{p} .

B 4-player game reciprocity

Figure 11 shows the complete breakdown of Observer behavior in the 4-player game, broken down by the scenario witnessed by the Observer, observability, and which Recipient was randomly selected for payment.

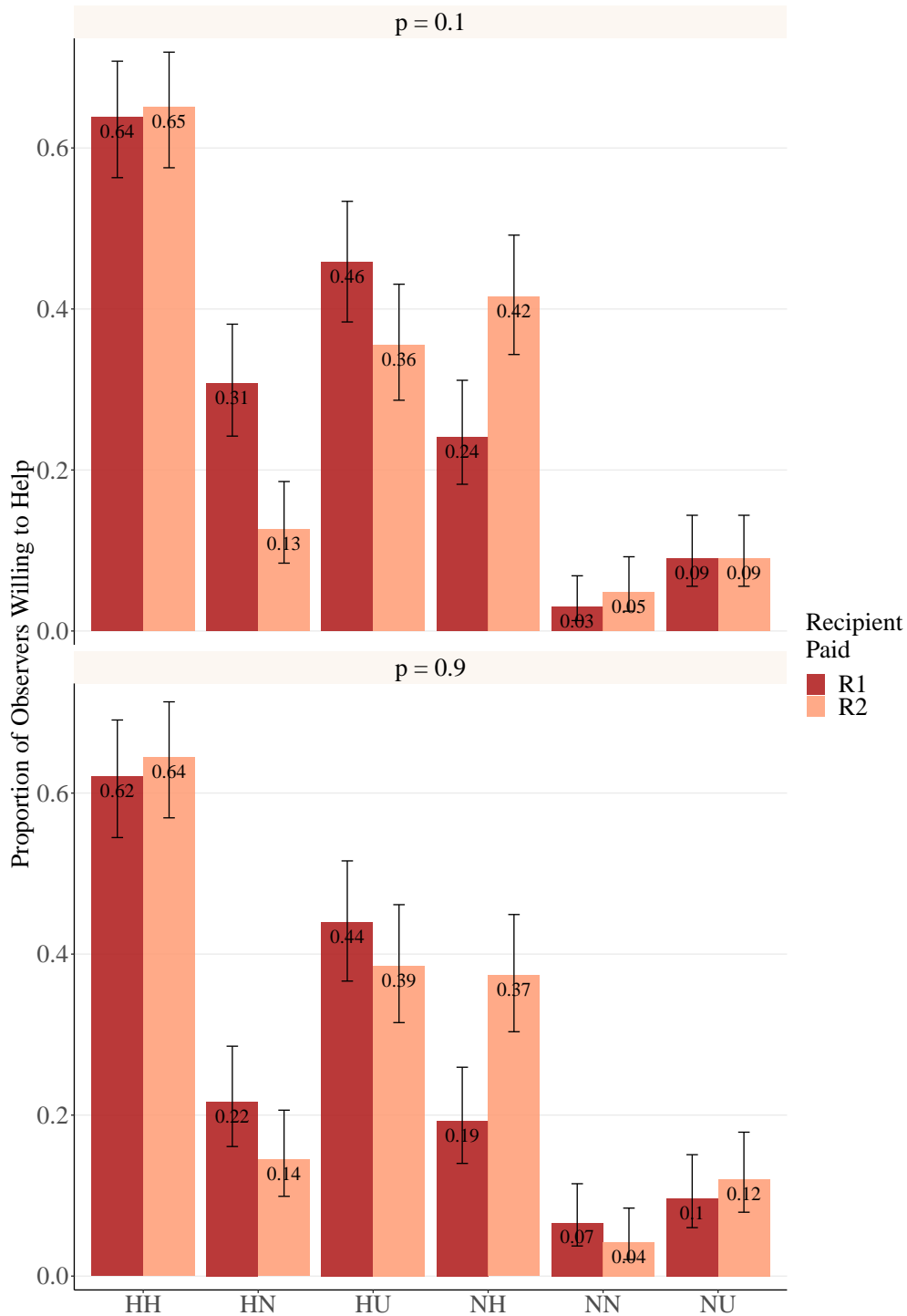


Figure 11: Full summary of Observer rates of reciprocation in the 4-player game, in each of the six possible scenarios witnessed, and according to the probability that the second Agent choice was witnessed and whether Recipient 1 or Recipient 2 was randomly selected to have the Agent's choice towards them implemented. Wilson score 95% confidence intervals are indicated.