Motivations for indirect reciprocity: Good deeds or good people?

Kieran Gibson
Lionel Page
Vera L. te Velde[*]

June 2, 2025

Abstract

We investigate the leading motivations for indirect reciprocity by experimentally studying the role of outcome-based, intentions-based, and type-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 more likely to be strategically motivated to build reputation, while only truly altruistic people behave altruistically 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 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. We thank Priscilla Man, Dorothea Kübler, Alessandra Cassar, Nick Netzer, Nicola Pavoni, Kenan Kalayci, David Smerdon, Zachary Breig, Antonio Rosato, Jamie Cross, Andy McLennan, and several seminar audiences for helpful conversations.

1 Introduction

The psychological motivations behind acts of kindness are multifaceted, ranging from genuine selflessness to covertly self-serving tactics. On one end of the spectrum, truly altruistic acts stem from empathy and genuine concern for the recipient, without any underlying expectation of reciprocation. On the other end of the spectrum, kind actions might be underpinned by a calculated self-interest, where the intent is to build a favorable reputation that could be beneficial in the future.[1] Distinguishing between these contrasting motivations is particularly critical in the context of *indirect reciprocity*, the phenomenon in which individuals tend to reward those who have demonstrated kindness previously. Extensive research has shown that indirect reciprocity can sustain cooperation in large groups (Mailath and Samuelson, 2006; Roberts et al., 2021a). The manner by which we judge others' records of kindness, and choose how to reciprocate, is however not well understood. In particular, it is not clear how strategically motivated acts of kindness–motivated by its associated reputational benefits–are assessed.

In the present 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 develop 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. Subsequently, the Observer decides whether to reward the Agent reciprocally. By varying several aspects of the information that the Observer has about the Agent and their actions, we can clarify 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 disentangle the role of the Agent's actions and their outcomes from the inner motivations behind those actions. To achieve this, we introduce experimental variation in the observability of the Agent's decision: the Observer witnesses the Agent's decision with a commonly known probability, p, which may be either high or low. This variation in the observability of the Agent's decision generates differences in the Agent's incentives to act kindly for reputational reasons. A helpful action that is likely to go unobserved is more likely to be driven by genuine altruism. On the other hand, a helpful action that is likely to be observed may also be driven by an instrumental concern to appear helpful in order to benefit from the Observer's indirect reciprocity

[1] This 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, Seinen and Schram (2006), Engelmann and Fischbacher (2009), and the results of this present paper.

later. Because the Observer is informed of the probability p, and therefore aware of whether the Agent expected their actions to be observed, the Observer can infer how genuinely altruistic the Agent is. In this way, we can explore whether the Observer's reciprocity is influenced by the perceived motives of the Agent or only by the actual choices made by the Agent.

In our second game, we add random variations in the outcome of the Agent's decision in order to disentangle the role of the Agent's actions (kind or unkind in intent, independent of the realized outcome) from resulting outcomes in the Observer's decision to reciprocate. The Agent makes choices about helping two different Recipients, but only one of the Agent's decisions is randomly implemented. The Observer sees one or both attempts and knows which choice was implemented, and so can condition their reciprocation on these factors separately. Furthermore, because observability of the Agent's choices varies as in the first game, this second game provides an opportunity to study how Observers react when Agents acts ”guilefully”, i.e. behaving kindly in public towards one recipient but unkindly towards another recipient when their action is unlikely to be observed.

Taken together, these two experimental games allow us to disentangle the motivations for indirect reciprocity. Surprisingly, we find that Observers do not care about the inner motives of the Agents – they reward good deeds regardless of whether they were done by genuinely ”good” people or not. That is, while Agents do strategically respond to the observability of their actions, Observers reward behavior no differently when done under high or low observability. Instead, reciprocity depends on both the outcome of the Agent's choices and also on the Agent's intended action, even when good intended actions do not lead to good outcomes.

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. 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 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 et al., 1986; Nowak and Sigmund, 1998). An important question in this approach is whether the reputational effect of failing to cooperate with others in the past depends on whether it was justified as a punishment of others' non-cooperation (Okada, 2020).

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 specific 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 and Nikiforakis, 2012).

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, among them (Kahneman, Knetsch and Thaler, 1986; Turillo et al., 2002; Eckel and Grossman, 1996; Güth et al., 2001; Stanca, 2009; Herne, Lappalainen and Kestilä-Kekkonen, 2013, e.g.). And even in one-shot games, third-party punishment has been extensively documented in the experimental literature (Balafoutas, Grechenig and Nikiforakis, 2014; Fehr and Fischbacher, 2004, e.g.). 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).

Economists have proposed 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 (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002, e.g.).[2] Second, intentions-based preferences, whereby people want to reciprocate with others who have been kind to them (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010, e.g.). Third, type-based preferences, whereby altruists are willing to act kindly with other altruists (Levine, 1998; Gul and Pesendorfer, 2016; Rotemberg, 2008, e.g.). 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.[3]

While type-based preferences are the leading model of indirect reciprocity (Engelmann and Fischbacher, 2009), they 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

[2] While 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.

[3] Experimental 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).

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, Sections 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.

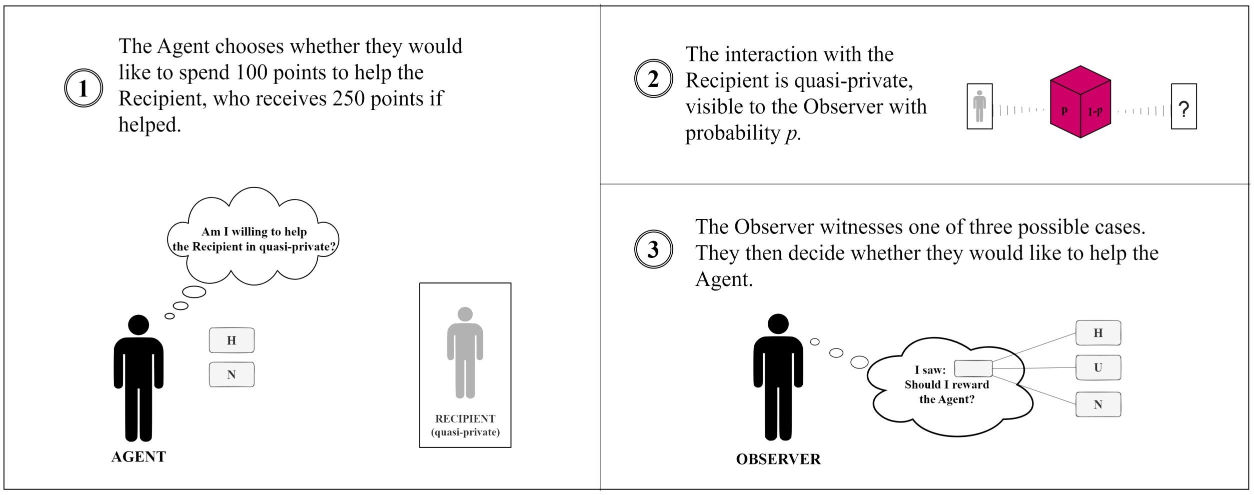

1
The Agent chooses whether they would like to spend 100 points to help the Recipient, who receives 250 points if helped.
Am I willing to help the Recipient in quasi-private?
H
N
AGENT
RECIPIENT (quasi-private)
2
The interaction with the Recipient is quasi-private, visible to the Observer with probability p.
p
1-p
?
3
The Observer witnesses one of three possible cases. They then decide whether they would like to help the Agent.
I saw:
Should I reward the Agent?
H
U
N
OBSERVER

Figure 1: The Three-Player Game

In addition, we generate random variations 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 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). 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 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, a 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.

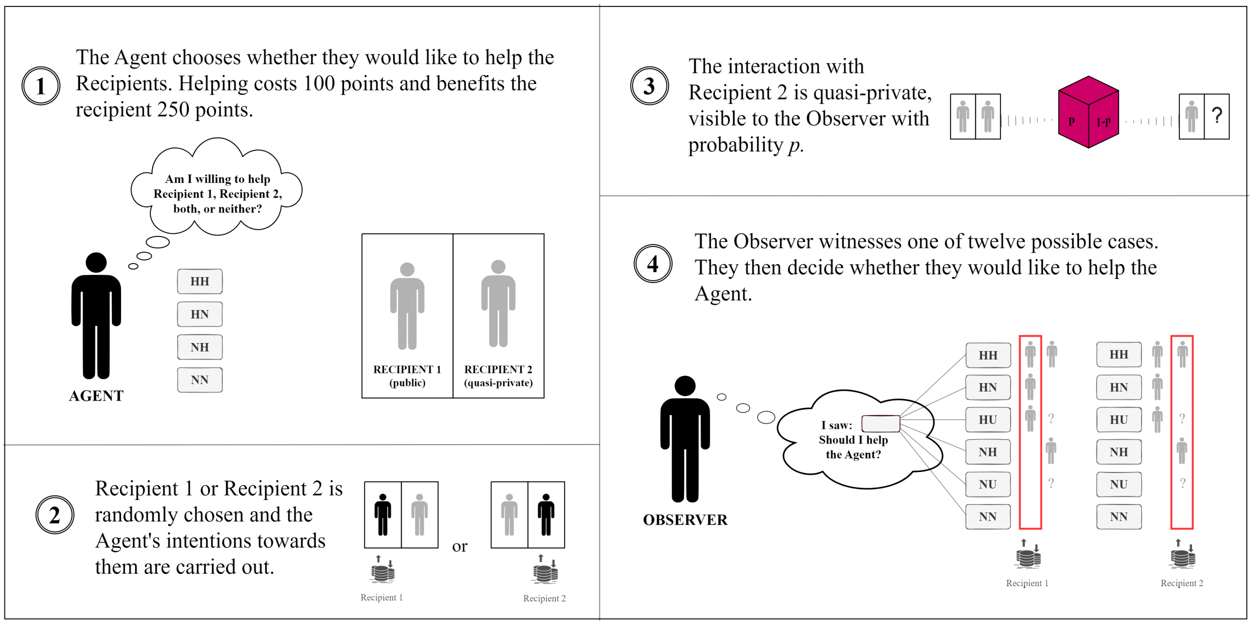

1
The Agent chooses whether they would like to help the Recipients. Helping costs 100 points and benefits the recipient 250 points.
Am I willing to help Recipient 1, Recipient 2, both, or neither?
HH
HN
NH
NN
AGENT
RECIPIENT 1 (public)
RECIPIENT 2 (quasi-private)
2
Recipient 1 or Recipient 2 is randomly chosen and the Agent's intentions towards them are carried out.
or
Recipient 1
Recipient 2
3
The interaction with Recipient 2 is quasi-private, visible to the Observer with probability p.
p
1-p
?
4
The Observer witnesses one of twelve possible cases. They then decide whether they would like to help the Agent.
OBSERVER
I saw:
Should I help the Agent?
HH
HN
HU
NH
NU
NN
?
?
Recipient 1
HH
HN
HU
NH
NU
NN
?
?
Recipient 2

Figure 2: The Four-Player Game

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 Schmidt, 2000; Bolton and Ockenfels, 2000; Charness and Rabin, 2002). Intentions-based preferences (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010) encapsulate the extent to which people react to kind intentions, independently from the outcomes those intentions lead to. 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 like inequality aversion are intuitively applicable to both the 3-player game and 4-player game. Concepts like inequality aversion, pure altruism, and similar models all predict that an Observer will be more inclined to opt for 'H' (help) when they witness an Agent's choice of 'H' being implemented. 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). 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 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, they do not directly apply to indirect reciprocity games: the presumption is that my inclination to show kindness to someone is dependent on their kindness towards me, not on their kindness towards anyone else. If anything, these models imply that if I perceive that someone's kindness towards another is meant to elicit indirect reciprocity from me, which ultimately leaves me worse off, I might interpret this as unkind. Therefore, we do not endeavor to formally adapt these models to our setting, but rather, we emphasize the underlying mechanism of reciprocating kindness with kindness. In our games, "intentions-based" indirect reciprocity involves 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. They will reciprocate least after witnessing NN. Reciprocation rates will not depend on p.*

2.2.3 Type-based Analysis

Type-based models of reciprocity, such as Levine (1998), are well suited to explain indirect reciprocity. 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. Accordingly, we adapt Levine's (1998) model to our experimental games.

Every player, denoted by i, is defined by their individual level of altruism, α_i. We assume α_i is weakly positive, reflecting a proclivity towards kindness (for simplicity we ignore the possibility of spite, $\alpha_i < 0$). The distribution of α_i is assumed to be uniform, represented as $\alpha_i \sim \phi = U[0, A]$. 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 refer to the Observer's altruism as α_O and the Agent's altruism as α_A.

Player i's altruism towards another player j is contingent on α_i and i's expectation of α_j according to the following equation:

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

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 $\lambda \in [0,1]$.[4]

In the 3-player game, we can derive the unique subgame perfect equilibrium via backwards induction. 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.

The Agent, anticipating this response by the Observer, strategically decides whether to help the Recipient based on their innate altruism and the likelihood that it will lead to indirect reciprocity from the Observer. As they understand that the Observer employs a cutoff strategy, they understand that their chances of being helped after the Observer witnesses H is equal to the chance that the Observer's altruism level α_O exceeds O_H (and similarly after the Observer witnesses N or U). The Agent's expected utility when choosing H is thus equal to:

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

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

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

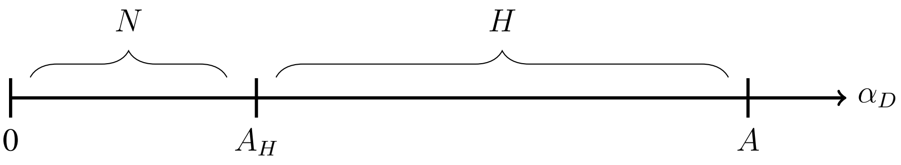
Similarly to Observers, we can define a cutoff value A_H for the Agent's altruism parameter α_A above which the Agent chooses to help. Given the cutoff strategy employed by the Agent, the Observer, upon witnessing H, can infer that the Agent's altruism level is at least A_H. Conversely, if the Observer 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. Sufficiently altruistic Observers will reciprocate, and the cutoff defining "sufficiently altruistic" depends on what they observe of the Agent. More precisely, the fewest Observers are altruistic enough to reciprocate after witnessing N and inferring that $\alpha_A < A_H$, while the most Observers are altruistic enough to reciprocate after witnessing H and inferring that $\alpha_A > A_H$. Proposition 1 describes this equilibrium:

[4] We 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.

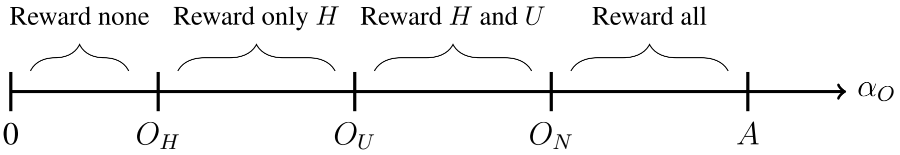
Proposition 1. *In the 3-player game with utility of the players determined by* (1)*, there is a unique subgame perfect equilibrium in which the Agent's and the Observer's decisions to help are determined by whether their respective altruism levels, α_A and α_O, are higher than thresholds A_H, O_U, O_N, O_H, specific to each situation:*

- *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 $\alpha_O \geq O_H$*
- *$O_H < O_U < O_N$*

Appendix A derives closed-form values of the cutoffs defined in Proposition 1. Figure 3 summarizes the cut-off values used to govern the Observer's decision. 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.



(a) The Agent's decision



(b) The Observer's decision

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 contrast, in the 4-player game some actions may not be supported in any subgame perfect equilibrium. Equilibrium 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 $\overline{p}$, 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 (**?**).

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 actions. But when p is large, HU is infrequent and the financial benefit 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, 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 choose 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.

We can therefore establish the following result (see Appendix A for further detail):

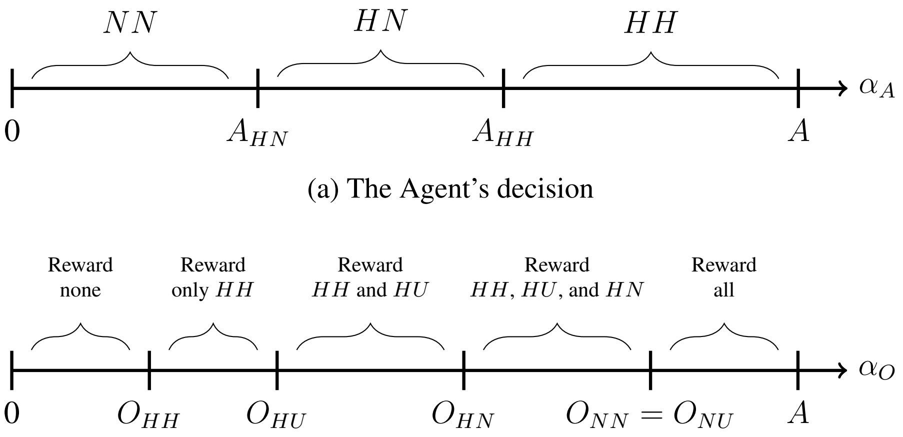
Proposition 3. *In the 4-player game with utility determined by* (1)*, there is a unique subgame perfect equilibria in which the Agent and the Observer decide to help in each specific situation if their altruism parameters are high enough. This equilibrium takes one of two forms. In both, the Observer reciprocates after witnessing $x \in \{HH, HU, HN, NH, NU, NN\}$ as long as $\alpha_O > O_x$.*

1. *Type 1 equilibrium occurs when p is below an upper limit $\overline{p}$. In this case, NN, HN, and HH are Agent strategies chosen on the equilibrium path. The Agent chooses NN if $\alpha_A < A_{HN}$; they choose HN if $A_{HN} < \alpha_A < A_{HH}$, and they choose HH if $\alpha_A > A_{HH}$. For the Observer, $O_{HH} < O_{HU} < O_{HN} < O_{NU} = O_{NN}$.*

2. *Type 2 equilibrium occurs when $p > \overline{p}$. In this case only HH and NN are supported in equilibrium. The Agent chooses HH so long as $\alpha_A > A_{2HH}$. For the Observer, $O_{HH} = O_{HU} < O_{NU} = O_{NN}$.*

Type 1 equilibrium is visually represented in Figure 4 which delineates the cutoff values for the Agent and the Observer.

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



(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

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 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 in which NN, HN, and HH are the three Agent strategies on the equilibrium path, if more Agents choose HH as p rises – that is if A_{HH} is decreasing in p – then fewer Observers will reciprocate after observing HH as p rises. That is, O_{HH} is increasing in p.*

2. *In a Type 2 equilibrium in which NN, HH are the two Agent strategies on the equilibrium path and $p > p^*$, changes in p do not change the equilibrium so long as p remains above 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 $\overline{p}$, 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 and 4-player game so long as Agents choose HH more often at higher p.*

3 Experimental Procedures

We conducted 9 experimental sessions between October 2021 and March 2022, totalling 168 participants. Two participants were dropped from analysis due to prior familiarity with the research project, leaving a sample of 166. Sessions were advertized to UQ's SONA participant database, comprised of several thousand of students and staff members from The University of Queensland. Sessions were conducted in person and lasted one hour, and paid an average of AUD $28.31. All tasks were computerized using oTree (Chen, Schonger and Wickens, 2016).

During each session, participants were re-matched anonymously and randomly with other subjects in the room between each of the 37 game rounds that they played. The complete protocol is shown in Online Appendix D. The sequence of tasks was as follows:

1. *Instructions*: Basic information about the experimental session was provided initially, without details about upcoming games.

2. *Mini-dictator game:* Participants were matched in pairs and each asked to choose between allocations of ($1,$1) and ($2,$0) for them and their partners respectively. One of the two partners' choices was randomly chosen to be implemented and they were informed of the result at the conclusion of the session.

3. *3-player game instructions and understanding check*: The rules of the 3-player game were introduced and all understanding check questions had to be answered correctly before proceeding.

4. *3-player game with direct method and feedback:* Each participant played six rounds of the 3-player game, once in each role with each of the two levels of observability (p=0.1 or 0.9). All participants learned the result of each round at its conclusion. One of the six rounds was randomly selected for payment, and the chosen round was reported at the conclusion of the session.

5. *3-player game with strategy method and no feedback:* Each participant played six rounds of the 3-player game, once in each role with each of the two levels of observability (p=0.1 or 0.9). Without any contemporaneous feedback, decisions could be made in any order, and so each participant first played as a Agent with one value of p, then as an Observer with the same value of p, and then as Agent and Observer with the other value of p, and lastly as Recipient with each p value. The ordering of p values was randomized. No feedback was provided and so no observational learning was possible between these rounds. One of the six rounds was randomly selected for payment, and the selected round was reported at the conclusion of the session.

6. *4-player game instructions and understanding check*: The rules of the 4-player game were introduced and all understanding check questions had to be answered correctly before proceeding.

7. *4-player game with direct method and feedback:* Each participant played 8 rounds of the 4-player game, once in each role with each of the two levels of observability (p=0.1 or 0.9). All participants learned the result of each round at its conclusion. One of the rounds was randomly selected for payment, and the chosen round was reported at the conclusion of the session.

8. *4-player game with strategy method and no feedback:* Each participant played 16 rounds of the 4-player game, once in each role with each of the two levels of observability (p=0.1 or 0.9). Ordering of decisions was as in part 5, with extra rounds as the second recipient and twice as many rounds as in part 7 in order for every game type to be experienced with each of Recipient 1 and Recipient 2 being randomly selected to have the Agent's choice towards them implemented. 2 of the 16 rounds were randomly selected for payment (at least one of

which was in the role of Agent or Observer) and the selected rounds were reported at the conclusion of the session.

9. *Belief elicitation*: Participants were asked to report their beliefs about how Agents played each of the games above at each level of p. They reported rates (integer instances out of 100) of choosing ($1,$1) in the mini-dictator game and rates of choosing H at each level of p in the 3-player game, along with rates of choosing each of HH, HN, NH, and NN at each level of p in the 4-player game. One of these 11 probabilities was randomly selected for payment according to the quadratic scoring rule.

10. *Results*: Results from rounds selected for payment were reported in full along with the resulting payments and total payment.

In half of the sessions, parts 3 through 5 were swapped with parts 6 through 8 in order to control for order effects, just as the ordering of $p = .1$ versus $p = .9$ was balanced in the strategy method parts described above. Online Appendix C investigates these order effects specifically.

Part 2 provides a rough baseline measure of altruism for each participant, which we use in Section 4 as an supplementary source of information in our investigation as to whether Observers are motivated by type-based reciprocity. Part 9 similarly provides the data we need to check whether Observers understand how Agents play the game, which is necessary in order to infer that they are basing their indirect reciprocity on that pattern of behavior.

Parts 4 and 7 are primarily intended to give participants practice and experience with the game in all scenarios prior to measuring their fully contingent plans using the strategy method. We want to minimize the learning that is taking place throughout parts 5 and 8 in order for those rounds to be as comparable as possible, and so we include the direct method games to support early learning and suppress all feedback during the strategy method games in order to prevent further learning.

Parts 5 and 8 are where we collect the core data analysed in the results. The strategy method pertains to the Observer, who reports whether they want to help the Agent in *each* of the 3 (in the 3-player game) or 6 (in the 4-player game) scenarios they might end up witnessing, conditional on the known value of p and (in the 4-player game) on which Recipient has been randomly selected to have their outcome implemented. Without this approach, we would observe very few reactions towards events that are witnessed with only a 10% probability, or towards choices that are rarely chosen by Agents. We exclusively analyze the strategy method data in the results below, and show in Online Appendix B that the differences between between strategy method and direct method data, where comparison is feasible, are minimal.

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. 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 5. 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. This finding strikingly contradicts part 3 of Hypothesis 3.

This result extends to the 4-player game as well. Figure 6 shows the choice rates for Agents and Observers in the 4-player game, contingent on observability.[5] 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 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.

Notice that in Figure 6, the decrease in Agents' choice of HH coincides with a large increase in the popularity of HN when observability is low, 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 incentive of a substantial number of Agents and brings into question the altruistic motivations of any HH-choosing Agent when observability is high. This also makes our evidence

[5] See Appendix B Figure 9 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.

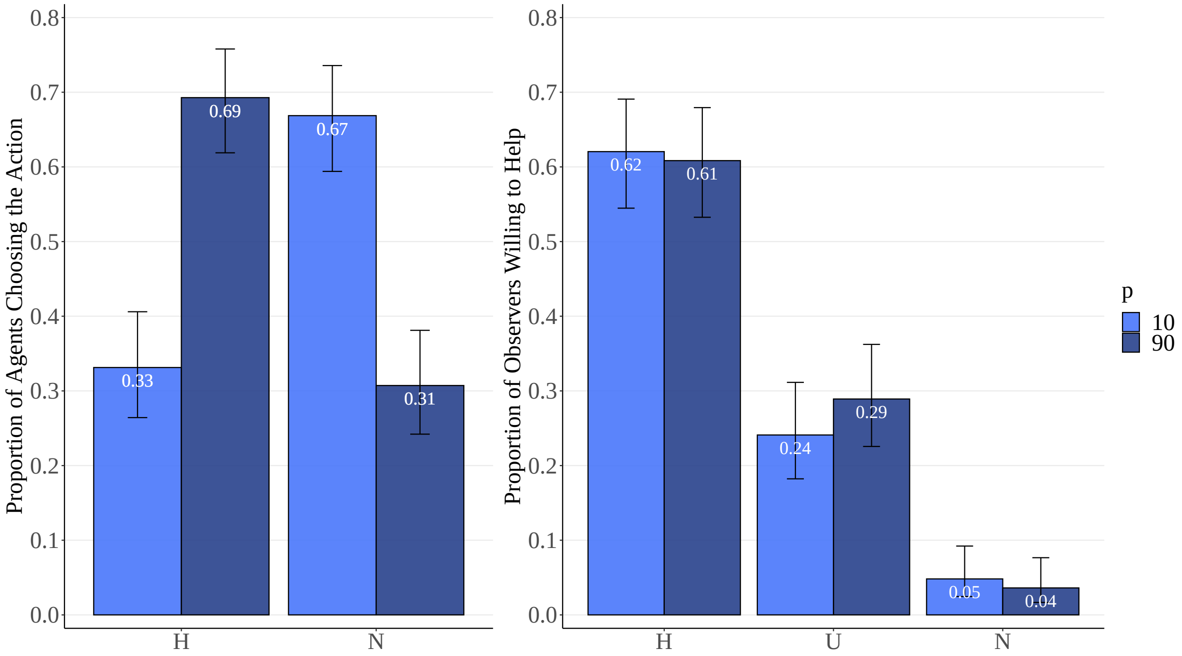


Figure 5: Observer and Agent choices in the 3-player game by observability, i.e. the probability that the Agent's quasi-private decision was witnessed by the Observer. The left panel shows the rates of Agents choosing to be helpful (H) or not (N), while the right panel shows rates of reciprocation in the three possible scenarios the Observer may witness (H, N, or Unobserved (U)). Wilson score 95% confidence intervals are indicated.

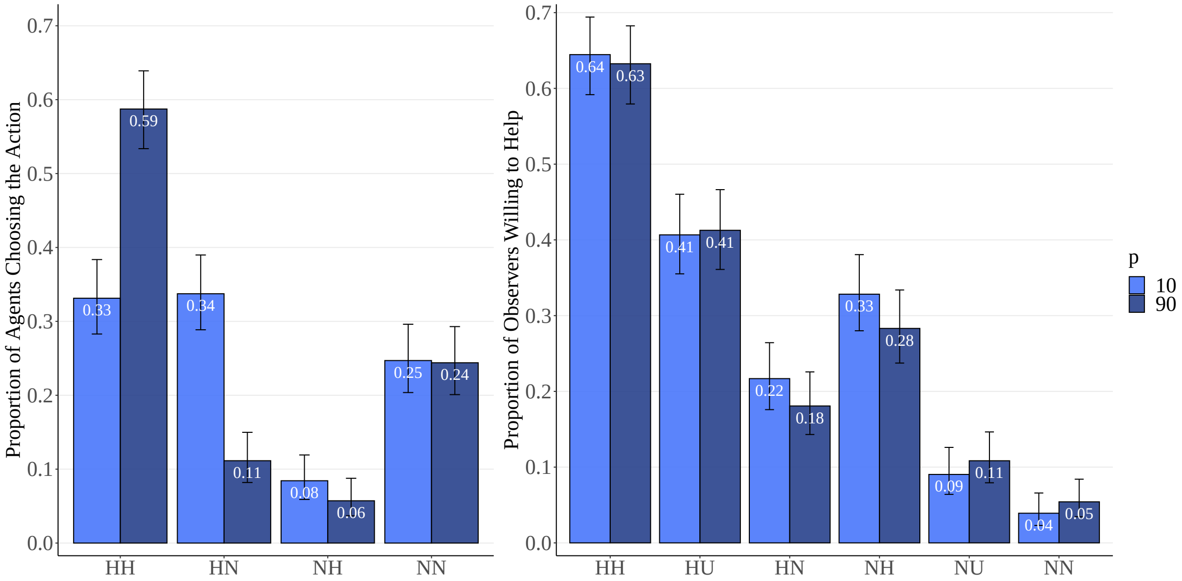


Figure 6: Observer and Agent choices in the 4-player game, by observability p of the Agent's second quasi-private choice. The left panel shows the rates of Agents choice combinations. The right panel shows Observers' rates of reciprocation contingent on the six possible scenarios they may have witnessed — 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.

against type-based preferences even starker — 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). 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 and more likely to choose HH in the 4-player game, consistent with Hypothesis 3 part 1.*

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 7 shows how Observers' reciprocation varies contingent on the Agent's decision x_A and the randomly chosen Recipient whose outcome is implemented.[6] 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, independently of the actual decisions made, appear to play a significant role in indirect reciprocity.

To summarize,

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 8 compares these rates of reciprocation. The darker bars to the left represent rates of reciprocity 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 (39% for HN and 26%

[6] For 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.

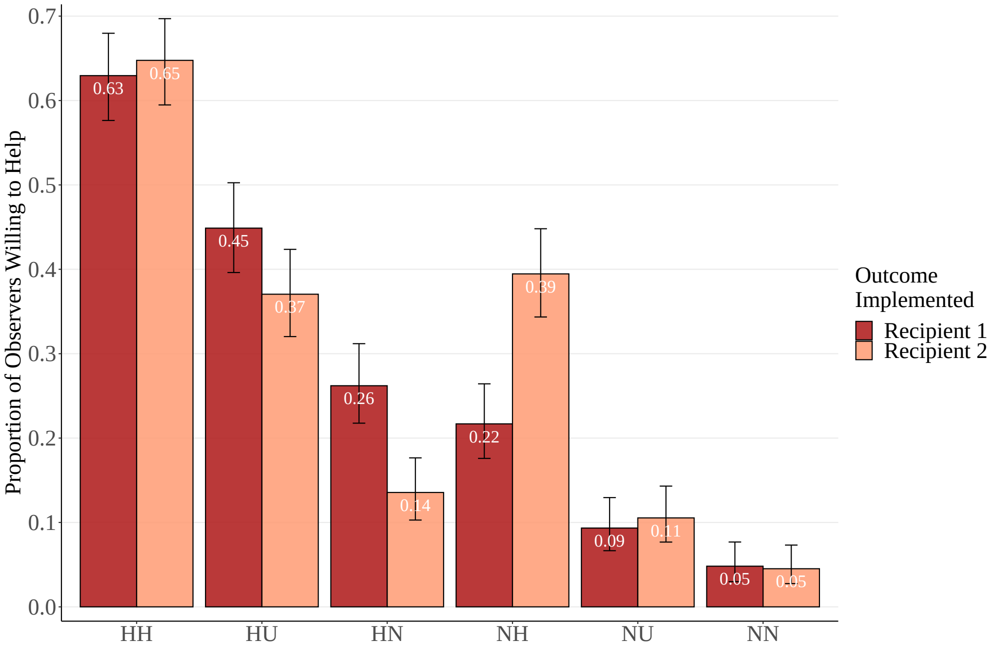


Figure 7: Observer rates of reciprocation in the 4-player game according to whether Recipient 1 or Recipient 2 was randomly selected to have the Agent's choice towards them implemented. Wilson score 95% confidence intervals are indicated.

for NH). These differences are highly statistically significant (both two-sample Z-test p-values < 0.001).[7]

The lighter bars on the right of Figure 8 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 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. These differences are also statistically significant (both two sample Z-test p-values < 0.01).

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 bear significance in and of themselves, rather than predominantly as an indicator of an underlying altruistic character type of the Agent, even when these intentions are directed towards someone else. This finding supports the general approach ofintentions-based models of reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010). However, intentions-based preferences were initially postulated to explain direct reciprocity specifically. It's not readily apparent

[7] Comparisons remain highly statistically significant when also controlling for observability.

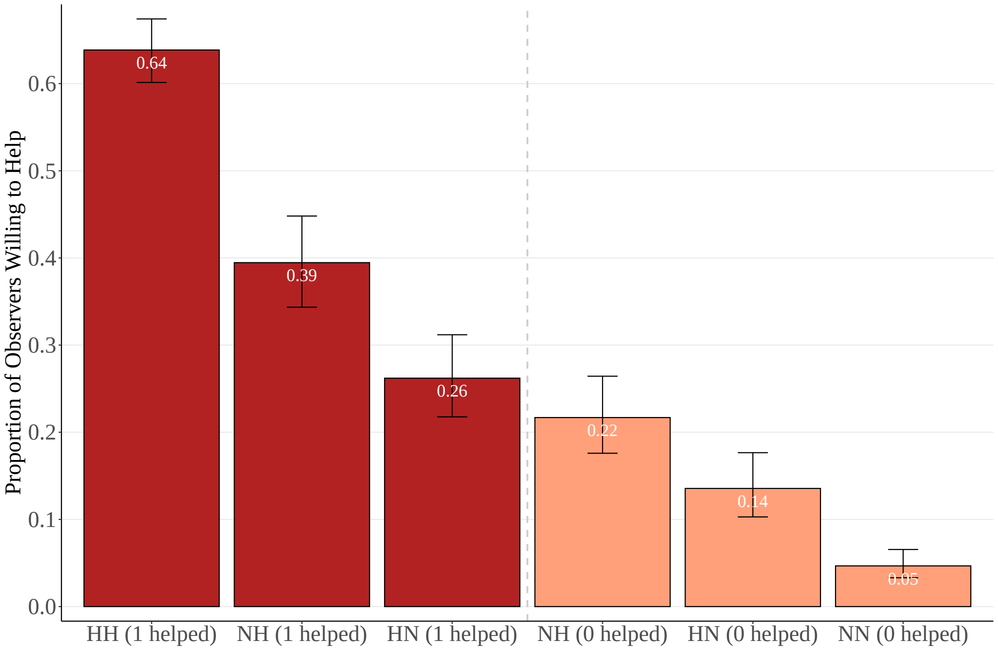


Figure 8: Observer rates of reciprocation in the 4-player game according to whether a helpful outcome was achieved as a result of the Agent's intentions. Wilson score 95% confidence intervals are indicated.

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 a 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.[8] Outcomes and intentions both separately strongly influence indirect reciprocity, in line with our results above.

[8] Including 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.

	Dependent Variable: Observer helped Agent?	
	Case Observed	
	(1)	(2)
Helpful Outcome?	0.4830***	0.5093***
	(0.0814)	(0.0877)
	[0.1404]	[0.1471]
Intentions	0.7825***	0.7756***
	(0.0712)	(0.0730)
	[0.2140]	[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, as a 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 Nonexistence of type-based reciprocity

Our experimental results indicate that indirect reciprocity is not type-based. 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 consider this when reciprocating. This suggests that type-based reciprocity isn't a strong motivation. Given the surprising nature of this result, three potential alternative explanations for our results are examined and ultimately dismissed.

The first alternative is that Observers might not recognize the strategic behavior of Agents (despite also acting as Agents themselves). However, our beliefs data robustly contradict this possibility. As shown in Table 2, participants believe that 32% of Agents are helpful when $p = 0.1$ and 66% are helpful when $p = 0.9$. Their perception of Agents being twice as helpful when observability is high, combined with the impressive accuracy of participants' average beliefs, demonstrates that they are well aware of Agents' strategic responses to changes in observability in the 3-player game. Beliefs in the 4-player game also clearly recognize that many more Agents are helpful when $p = 0.9$.

The second alternative explanation is that some types of players do have type-based preferences but this heterogeneity is concealed by focusing on aggregate behavior and beliefs. For example, it's possible that only strategic helpers realize that others will be strategic, or perhaps that strategic helpers are kinder to helpers likely also to be strategic (when $p = 0.9$) due to homophily or a similar motivation. 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),

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).

Strategists (who help only when observability is high), and Selfish (who never help).[9]

	N	N%	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.

Table 3 presents Observer behavior and beliefs for these types. All types clearly recognize that Agents are more helpful when $p = 90\%$. Furthermore, all types, separately, treat helpful and unhelpful Agents equally, regardless of observability (though differently from other types). These results reinforce our interpretation that Observers are not driven by type-based reciprocity.

[9] 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.

The last alternative explanation questions the assertion of our type-based reciprocity model that Agents choosing to Help when $p = 0.1$ are indeed more altruistic than those who help when $p = 0.9$. If this were 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 weakly statistically significant with a p-value of 0.06 (one-tailed two-sample Z-test).

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*b*). 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).

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 possibilities will require further research.

Another possibility is that type-based preferences matter in social interactions but only in specific settings. In our games, like in typical models of indirect reciprocity, the Observer only interacts with the Agent when deciding to help or not. Many social settings with repeated interactions are richer, giving people the choice of whether to trust others in further cooperative games. 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).

Regardless of which of these interpretations is correct, our results clearly indicate that future theoretical approaches to indirect reciprocity should move away from type-based preferences. We suggest instead that the intentions-based framework from the literature on direct reciprocity be adapted to account for the importance of intentions towards third parties that we observe.

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 6 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 is surprising. 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

	Dependent Variable: Observer helped Agent?		
	Case Observed		
	(1)	(2)	(3)
Helpful Outcome?	0.5093***	0.5289***	0.5284***
	(0.0877)	(0.0921)	(0.0924)
	[0.1471]	[0.1523]	[0.1520]
Intentions	0.7756***	0.7285***	0.7288***
	(0.0730)	(0.0692)	(0.0694)
	[0.2101]	[0.1950]	[0.1949]
Guile?		−0.2779***	−0.2235**
		(0.0727)	(0.0873)
		[−0.0740]	[−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.

by pure altruism, they should be rewarded less than a pure altruist.[10] 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 Table 4 Column 3 which shows that guile does not substantially vary with observability.

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

We explore a key question about the underlying motivations of indirect reciprocity: Do people reward good deeds or good people?

To answer this, we design an experimental framework that allows us to separate the effect of outcomes from an Agents' actions and motives on future indirect reciprocity. We observe extensive indirect reciprocity, but, surprisingly, find that type-based preferences are unable to explain it: Although the possible strategic motives of players who help in the hope of benefiting from indirect reciprocity can be detected by observers in our experiment, and even though these strategic motives seem well understood by the participants in our games, they do not affect the propensity of observers to reciprocate that helpful behavior.

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 rewarded good deeds rather than good people. This result is striking given that type-based preferences,

[10] 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 implied by intentions-based direct reciprocity.

based on the assumption that people aim to reward good people, have been seen as ideally suited to explain indirect reciprocity.

Besides discovering indirect reciprocity is driven by a desire to reward good deeds, we also identified 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 more privately. This effect, which we label as "guile aversion," raises a critical question: How should people react toward 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 results shed a new light on the psychological mechanisms sustaining an indirect reciprocity equilibrium in a large group. While type-based preferences seem naturally suited to sustain an indirect reciprocity equilibrium based on reputation, our results suggest that they may not be necessary. Instead, such an equilibrium may rely on people caring about others having a reputation for doing the right thing, without necessarily caring about whether this reputation reflects either inner prosocial motives 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, our results do not exclude the possible role of type-based preferences in all types of social interactions. In situations where people select partners for medium to long-term interactions (co-workers, friends, mates) selecting partners with pro-social types may be the best strategy. It may help ensure that they will do the right thing when, in some future situations, they may face little scrutiny and have the opportunity to gain from defecting from cooperation. Preferences over the reputation of one's social partners could be, in that sense, dependent on the type of social interactions. This possibility opens the way for future research to provide richer insight into the reasons why we care about others' reputations in different types of social contexts.

Preferences over the reputation of one's social partners could be, in that sense, dependent on the type of social interactions. For limited interactions, caring about the reputation of others to do the right thing may be enough to sustain cooperation, while for situations where people select others for a lasting partnership, it may be more important to care about the reputation of others as being good people. This possibility opens the way for future research to provide richer insight into the reasons why we care about others' reputations in different types of 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.

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.

Balafoutas, Loukas, and Nikos Nikiforakis. 2012. "Norm enforcement in the city: A natural field experiment." *European Economic Review*, 56(8): 1773–1785.

Balafoutas, Loukas, Kristoffel Grechenig, and Nikos Nikiforakis. 2014. "Third-party punishment and counter-punishment in one-shot interactions." *Economics Letters*, 122(2): 308–310.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2009. "Dynamic psychological games." *Journal of Economic Theory*, 144(1): 1–35.

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.

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.

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.

Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A theory of equity, reciprocity, and competition." *American Economic Review*, 90(1): 166–193.

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.

Charness, Gary, and David I. Levine. 2007. "Intention and stochastic outcomes: An experimental study." *The Economic Journal*, 117(522): 1051–1072.

Charness, Gary B. 2004. "Attribution and reciprocity in an experimental labor market." *Journal of Labor Economics*, 22(3): 665–688.

Charness, Gary B., and Matthew Rabin. 2002. "Understanding social preferences with simple tests." *Quarterly Journal of Economics*, 117(3): 817–869.

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.

Cox, James C. 2004. "How to identify trust and reciprocity." *Games and Economic Behavior*, 46: 260–281.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A theory of sequential reciprocity." *Games and Economic Behavior*, 47(2): 268–298.

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.

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.

Falk, Armin, and Urs Fischbacher. 2006. "A theory of reciprocity." *Games and Economic Behavior*, 54(2): 293–315.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2003. "On the nature of fair behavior." *Economic Inquiry*, 41(1): 20–26.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2008. "Testing theories of fairness - Intentions matter." *Games and Economic Behavior*, 62(1): 287–303.

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 Klaus M Schmidt. 2000. "Fairness, incentives, and contractual choices." *European Economic Review*, 44(4-6): 1057–1068.

Fehr, Ernst, and Urs Fischbacher. 2004. "Third-party punishment and social norms." *Evolution and Human Behavior*, 25(2): 63–87.

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.

Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological games and sequential rationality." *Games and Economic Behavior*, 1(1): 60–79.

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.

Gul, Faruk, and Wolfgang Pesendorfer. 2016. "Interdependent preference models as a theory of intentions." *Journal of Economic Theory*, 165: 179–208.

Güth, Werner, Manfred Königstein, Marchand Nadége, 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. Frankenhuis, and Tyler Moore. 2009. "Adaptive rationality: An evolutionary perspective on cognitive bias." *Social Cognition*, 27(5): 733–763.

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.

Johnsen, Åshild A., and Ola Kvaløy. 2016. "Does strategic kindness crowd out prosocial behavior?" *Journal of Economic Behavior & Organization*, 132: 1–11.

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.

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.

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.

Khadjavi, Menusch. 2017. "Indirect reciprocity and charitable giving: Evidence from a field experiment." *Management Science*, 63(11): 3708–3717.

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

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.

Levine, David K. 1998. "Modeling altruism and spitefulness in experiments." *Review of Economic Dynamics*, 1(3): 593–622.

Lin, Hong (Hannah), and David Ong. 2011. "Separating gratitude from guilt in the laboratory." Available at SSRN: https://ssrn.com/abstract=1943949 or http://dx.doi.org/10.2139/ssrn.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.

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.

Milinski, Manfred, Dirk Semmann, and Hans-Jürgen Krambeck. 2002. "Reputation helps solve the 'tragedy of the commons'." *Nature*, 415(6870): 424–426.

Nelson, William Robert. 2002. "Equity or intention: it is the thought that counts." *Journal of Economic Behavior & Organization*, 48(4): 423–430.

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.

Offerman, Theo. 2002. "Hurting hurts more than helping helps." *European Economic Review*, 46(8): 1423–1437.

Okada, Isamu. 2020. "A review of theoretical studies on indirect reciprocity." *Games*, 11(3).

Orhun, A. Yeşim. 2018. "Perceived motives and reciprocity." *Games and Economic Behavior*, 109: 436–451.

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*a*. "The benefits of being seen to help others: Indirect reciprocity and reputation-based partner choice." *Philosophical Transactions of the Royal Society B*, 376(1838): 20200290.

Roberts, Gilbert, Nichola Raihani, Redouan Bshary, Héctor M. Manrique, Andrea Farina, Flóra Samu, and Pat Barclay. 2021*b*. "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.

Rotemberg, Julio J. 2008. "Minimally acceptable altruism and the ultimatum game." *Journal of Economic Behavior & Organization*, 66(3-4): 457–476.

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.

Sebald, Alexander. 2010. "Attribution and reciprocity." *Games and Economic Behavior*, 68(1): 339–352.

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.

Semmann, Dirk, Hans-Jürgen Krambeck, and Manfred Milinski. 2004. "Strategic investment in reputation." *Behavioral Ecology and Sociobiology*, 56(3): 248–252.

Stanca, Luca. 2009. “Measuring indirect reciprocity: Whose back do we scratch?” *Journal of Economic Psychology*, 30(2): 190–202.

Stanca, Luca, Luigino Bruni, and Luca Corazzini. 2009. “Testing theories of reciprocity: Do motivations matter?” *Journal of Economic Behavior & Organization*, 71(2): 233–245.

Sugden, Robert, et al. 1986. *The economics of rights, co-operation and welfare.* Palgrave Macmillan UK.

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.

van Apeldoorn, Jacobien, and Arthur Schram. 2016. “Indirect reciprocity: A field experiment.” *PLOS ONE*, 11(4): e0152076.

Wedekind, Claus, and Manfred Milinski. 2000. “Cooperation Through Image Scoring in Humans.” *Science*, 288(5467): 850–852.

Wedekind, Claus, and Victoria A Braithwaite. 2002. “The long-term benefits of human generosity in indirect reciprocity.” *Current Biology*, 12(12): 1012–1015.

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 Propositions 1 and 2

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, $E[\alpha_A] = \overline{\alpha} = A/2$. 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 equilibrium:

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

The final equation that must be satisfied 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 cut-off values that define the Agent's and Observer's equilibrium strategies.[11] Solving this system yields the following result, characterizing equilibrium:

[11] We assume these values satisfy $0 < A_H, O_H, O_N, O_U < A$. Otherwise, some actions will never be taken in equilibrium resulting in trivial solutions.

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

These cutoff values are all, of course, functions of p, although we have not made this notationally explicit for the sake of readability. But this relationship between Agents' and Observers' choices and p is of primary importance. So let us consider, how does a change in p affect the Agents? $\frac{\partial}{\partial p}A_H = -\frac{\lambda}{2} < 0$. Intuitively, this is because the higher the observability the more opportunities there are for reciprocity, and so Agents with lower altruism are drawn to helping.

How does a change in p affect the Observers? $\frac{\partial}{\partial p}O_H = \frac{\partial}{\partial p}O_N = \frac{\lambda^2}{4} \geq 0$. Both partial derivatives are positive, which is because the average altruism of helpful and unhelpful Agents are both decreasing as p increases. Helpful Agents are less impressive as p increases because they are more likely to be helping with a hope of reciprocation, and unhelpful Agents are also less impressive as they still did not help the Recipient despite the added financial motivation. Thus, both H and N trigger a lower probability of reciprocation when p is high than when it is low.

A.2 Proof of Propositions 3 and 4

To first 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 small, and HH is preferred to all other options as long as α_A is sufficiently high[12], so any 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).

We can furthermore eliminate NH as a possibility by appealing to the D1 criterion. 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

[12] As in the 3-player game, we assume that the distribution of types is broad enough that both NN and HH are in fact chosen by some types in equilibrium.

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: Following the same approach as the 3-player game, the system of equations for the six cut-off values that characterize a Type 1 equilibrium are as follows:

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

These yield the following closed-form solutions for A_{HH} and A_{HN} (the remaining values will not be needed and are omitted for brevity):

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

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.

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

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

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

The Agent's strategy in a Type 2 equilibrium is characterized by a single cutoff parameter, A_{2HH}, at which Agents switch from choosing NN to HH. In this situation, the Observer can infer exactly what choice the Agent made even when the quasi-private choice is unobserved. Equilibrium is therefore characterized by the following three cut-off values:

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

(Closed-form solution omitted for brevity.) Note that this equilibrium characterization is not dependent on p.

B 4-player game reciprocity

Figure 9 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.

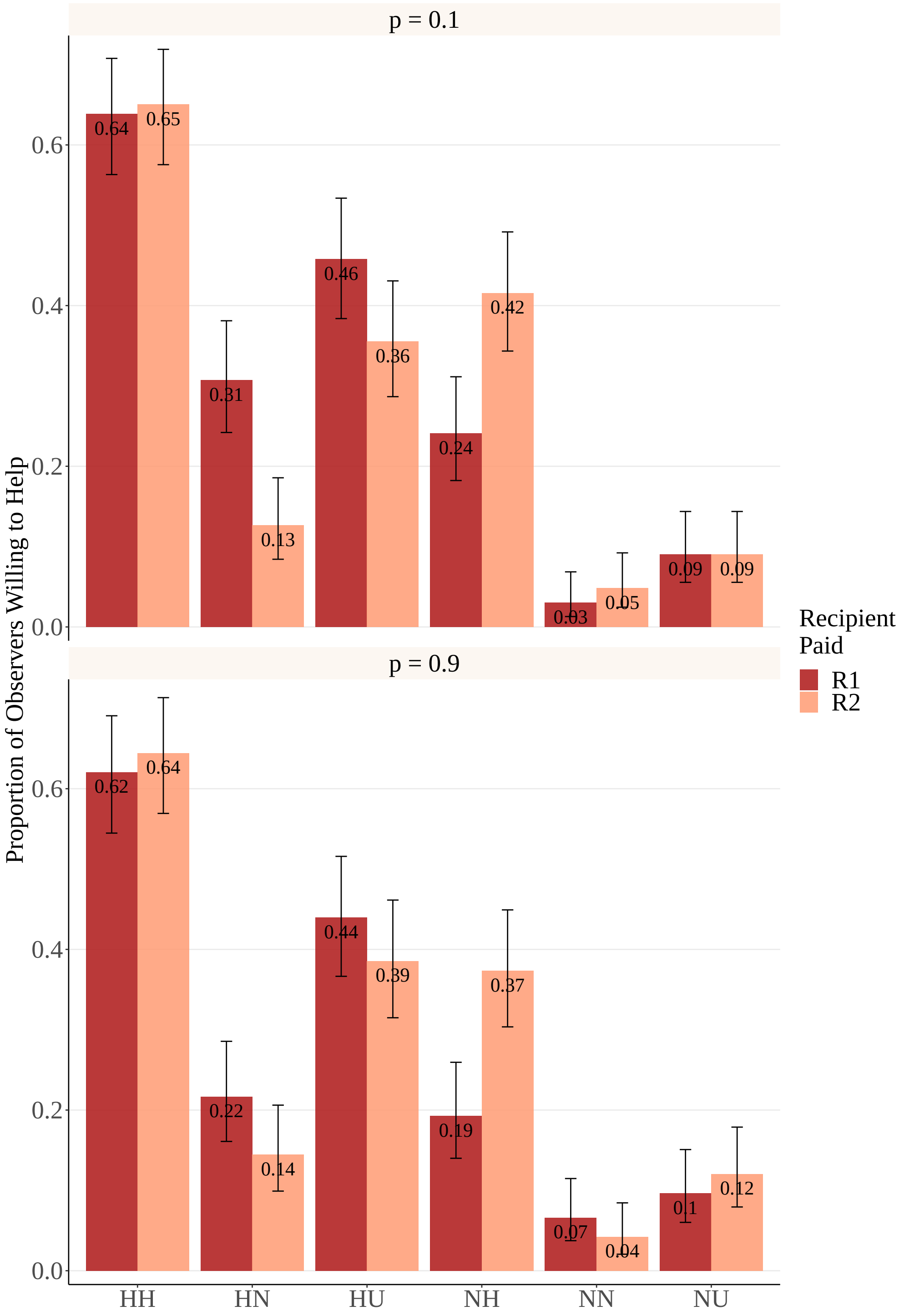


Figure 9: Full summary of Observer rates of reciprocation in the 4-player game, in each of the six possible scenarios witnessed, and according to the probability that the second Agent choice was witnessed and whether Recipient 1 or Recipient 2 was randomly selected to have the Agent's choice towards them implemented. Wilson score 95% confidence intervals are indicated.