working paper
2024

Reprocity and Uncertainty: When Do People Forgive?

Andrés Gago

October 2020

CENTRO DE ESTUDIOS
MONETARIOS Y FINANCIEROS
Casado del Alisal 5, 28014 Madrid, Spain
www.cemfi.es

CEMFI Working Paper No. 2024
October 2020

Reprocity and Uncertainty: When Do People Forgive?

Abstract

A sizable proportion of individuals act reciprocally. They punish and reward depending on the (un)kindness of those with whom they interact. In this paper, I explore whether individuals still reciprocate intentions when others lack full control over the consequences of their actions. By means of a dictator game with punishment opportunities, I show that unkind intentions are enough to trigger punishments, irrespectively of the outcome. By contrast, accidents are forgiven. To isolate how uncertainty over the result of an action affects the assessment of intentions, I control for other possible departures from self-profit maximization, such as distributional concerns or efficiency maximization. I find that the former also plays a role in respondents' behavior.

JEL Codes: C91, D63, C79.

Keywords: Reciprocity, uncertaint,; blame, intentions, dictator, punishment.

Andrés Gago
Universidad Torcuato Di Tella
agago@utdt.edu

1. Introduction

The literature on experimental economics has shown that around one third of people act reciprocally. They are willing to forgo wealth in order to reward those who have been kind to them and punish those who have been unkind.[1] Nevertheless, people are not always successful when they intend to be either kind or unkind, as the result of their actions is usually subject to uncertainty. In this paper, I study how lack of full control over outcomes affects reciprocity.

Conventional wisdom tells us that when we judge what others do to us, *it is the thought that counts*. Nevertheless, some evidence also suggests that subjects could be confounded when they assess actions with unintended consequences. In an influential psychology paper, Baron and Hershey (1988) find, by means of vignettes, that people differently evaluate decisions that are ex-ante identical in probabilistic terms, but lead to different outcomes. Similarly, in a recent paper, Brownback and Kuhn (2019) show that luck affected principals' inference about agent types even when effort was perfectly observable. If subjects suffer a cognitive bias when they judge someone else's actions, it is not clear how they would reciprocate unintended outcomes. This paper examines whether subjects react to intentions when these are not congruent with outcomes.

Using an experiment, I observe that subjects punish and reward according to the will behind an action, even if its result is the opposite to what was intended. When people intend to be unkind and fail, they are punished. Similarly, if they intend to be kind but do not succeed, they are forgiven. In my experiment, adding randomness to human action does not change the fact that reciprocity is intention-based.

To reach this conclusion, I run a dictator game with punishment opportunities. The dictator chooses between two options to split €20 between herself and a respondent. In one treatment, she chooses between a certain allocation of €10-€10 or a lottery that allocates €16 to the dictator and €4 to the respondent with a high probability. In the other treatment, she chooses between a certain option that allocates €16-€4 (favoring the dictator), or a lottery that allocates €10-€10 with a high probability. In both treatments, when the dictator chooses the lottery there is a small probability that the final outcome coincides with that of the certain allocation. The respondent then observes the dictator's choice (and the outcome of the lottery, if applicable) and decides whether she wants to assign any punishment (or reward) to the dictator at a cost. This enables me to compare the reaction to choices that have the same intentions but different consequences.

To identify the effect that partial control over outcomes has on reciprocity, I run two additional treatments. First, a treatment in which outcomes are decided by nature -a random device- from the very beginning, which captures inequity aversion. This allows me to disentangle whether subjects might suffer a cognitive bias or whether they have distributional concerns. Second, I run another reciprocal treatment in which both options are

[1] For a discussion on this see Fehr and Schmidt (2006) and Cooper and Kagel (2016).

certain. This allows me to explore whether choosing a certain/uncertain option vis-a-vis an uncertain/certain option modifies responses in any other way that could be orthogonal to intentions.

I observe that when intentions are unkind and outcomes are bad, respondents are willing to forgo wealth to punish dictators and these punishments go beyond the level that can be explained solely by inequity aversion. This is in line with previous results in McCabe, Rigdon and Smith (2003), Falk, Fehr and Fischbacher (2008), Blount (1995), and Gächter and Thoni (2010). Moreover, and this is the novelty of the paper, I observe that in situations in which outcomes and intentions are not aligned, the latter still drive reciprocity. When intentions are unkind, but outcomes are good, respondents still punish dictators. Conversely, when outcomes are bad, but intentions are kind, the amount of money that respondents subtract is indistinguishable from the amount subtracted on the grounds of distributional concerns. Furthermore, when I compare responses in the certain and uncertain reciprocal treatments, I observe that punishments are statistically the same. Altogether, this reveals that when dictators have partial control over outcomes, respondents still reciprocate based on intentions. They do not suffer any cognitive bias when they judge actions with unintended consequences and uncertainty plays no role in their decisions.

These conclusions are partially at odds with previous literature. In a psychology paper that is closely related to mine, Gino, Shu and Bazerman (2010) find that participants tend to punish more when the same action leads to a negative result, arguing that people are biased by outcomes. Nonetheless, while Gino, Shu and Bazerman (2010) assess the effect of outcomes on recipient responses altogether, in my design I compare reciprocal responses to responses in the nature treatment. This makes it possible to distinguish between a cognitive bias and distributional concerns, which allows me to reconcile the notion that it is the thought that counts with the fact that respondents subtract money from dictators when outcomes are uneven. People do not suffer a cognitive bias, they judge others by their intentions. However, at the same time, they are willing to redress inequality, no matter how it arises. This distinction enhances our understanding of social preferences and sheds light on what to expect in situations in which distributional concerns play no role.[2]

The results of this paper are also at odds with the no-harm-no-foul hypothesis proposed in Bartling and Fischbacher (2012). In an interesting study about attribution of responsibility, they conclude, among other things, that when the outcome derived from an action is positive, people will not punish the subject even if her intention was unkind. To come to this conclusion, they observe behavior in two situations. In the first one, dictators can choose a fair allocation or delegate the choice to someone else. In the second one,

[2]There are also two important methodological differences. First, in their study, punishments are either costless or entail low average cost (0.1$), while in this paper the cost is always 1 to 3. Second, their study involves deception, as there are no participants in the role of dictators, while in this paper, there are.

dictators can choose between a fair allocation, an unfair allocation, or delegate the choice to a random device with a known probability distribution. In both cases, the authors find that respondents do not punish dictators if the outcome after delegation is good. In contrast with my experiment, in their first scenario respondents do not know the probability of each outcome after delegation (there is ambiguity), while in the second, delegation is the intermediate (and arguably neither kind nor unkind) choice. In my design, dictators choose between a certain outcome and a lottery with known probability and lower expected outcome. Arguably, when intentions are markedly unkind and unambiguous, the no-harm-no-foul hypothesis does not hold any longer.

My conclusions are aligned with what Charness and Levine (2007) find for positive reciprocity and Bartling, Fischbacher and Schudy (2015) for collective decision-making. Charness and Levine (2007) conclude that people reward good intentions in a gift-exchange game with uncertainty. As Offerman (2002) and Dohmen et al. (2008) show, positive and negative reciprocity are uncorrelated motivations that different individuals can have. Focusing on negative reciprocity enables me to test the hypothesis that no harm implies no foul and observe whether individuals forgive unintended negative outcomes.[3] Bartling, Fischbacher and Schudy (2015) run an experiment with a *board of dictators* deciding upon the distribution of a pie. They find that respondents held dictators responsible for their votes, no matter the final result of the voting. Introducing randomness through a lottery allows me to have a closer look into the settings in which psychologists have found judgments could be biased.

The fact that respondents are not influenced by an outcome bias and punish dictators for what they intend to do might explain why Friehe and Utikal (2018) observe that hiding intentions is punished. Dana, Weber and Kuang (2007) show that dictators more often choose the selfish option when their choice is not directly observable. The authors rationalize this behavior as a desire of dictators to avoid looking unfair, which might erode their social and self-image. Moreover, I document that when receivers actually have the opportunity to punish dictators for their choices, they judge them solely by their intentions. The absence of a nature treatment in Friehe and Utikal (2018) does not allow to establish, analyzing their data, whether the reason for a significant difference in punishments after good/bad outcomes is due to inequity aversion or to the type of cognitive bias documented in Baron and Hershey (1988). Nevertheless, the fact that hiding intentions is punished, is congruent with the hypothesis that when dictators lack full control over outcomes, respondents use intentions to judge dictators and punish them both monetarily and socially.

[3] Rubin and Sheremeta (2015) also study how uncertainty affects behavior in a gift-exchange game. Unlike Charness and Levine (2007) they use a game with three decisions: salary, effort and bonuses, and find that, with random shocks to the effort level, individuals get further from the optimum. To disentangle strategic considerations from reciprocity, in my experiment I look at a simpler setting and avoid framing the situation as an investment opportunity (Stanca, Bruni and Corazzini, 2009)

In the last years many other authors have explored the intersection between social preferences and uncertainty.[4] However, none of them tackles the specific question that is central to this paper: Whether negative reciprocity is still intention-based when intentions and outcomes are incongruous. The same is true for papers that study the relative importance of intentions versus outcomes by studying how responses are affected when the choice set varies (see Brandts and Sola, 2001; Bolton, Brandts and Ockenfels, 1998; Sutter, 2007).

The results of this paper inform how to extend theories of social preferences to a context of uncertainty. Fudenberg and Levine (2012) and Saito (2013) show that only applying expected utility theory to classical models of other regarding preferences might be an unsatisfactory solution. Hence, to build sound models of social preferences under uncertainty, it is essential to provide theorists with data that disentangles and correctly characterizes all different motivations that drive behavior. The decisions taken by respondents in this experiment are consistent with the extension that Sebald (2010) proposes to the theory of sequential reciprocity in Dufwenberg and Kirchsteiger (2004). He proposes that respondents evaluate the kindness or unkindness of dictators by looking at expected outcomes and decide punishments in accordance. This is exactly what the experiment shows. In addition, I find this coexists with a preference for equality, absent in his model, that leads subjects to balance payoffs whenever they are uneven.

These results apply to situations in which reciprocity has been found to be relevant and individuals do not have perfect control over the consequences of their actions. A classic example are labor market relationships, in which the decision to go on strike (or pay a bonus) would be closely related to the intentions of the management (workers).[5] Following the results of this paper, the CEO of a firm that is struggling due to an external economic shock, like a pandemic, would have more support from her employees if she has to make unpopular decisions than the CEO of a firm that wants to relocate production to China to save costs and earn a higher bonus. Even if in both cases the ultimate result is the closure of a factory, the workers would analyze the intentions of the management to decide their response.

[4] Rand, Fudenberg and Dreber (2015), Bereby-Meyer and Roth (2006), Xiao and Kunreuther (2016), Markussen, Putterman and Tyran (2016), and Klempt (2012) study how uncertainty affects the ability to cooperate, Cappelen et al. (2013), Cettolin and Riedl (2017), Trautmann and van de Kuilen (2016) and Andreoni et al. (2020) study whether inequity-averse individuals care about people having equal chances or equal outcomes, Gurdal, Miller and Rustichini (2013) let an agent decide whether to invest in a safe or a risky lottery on behalf of a principal and Krawczyk and Le Lec (2010), and Brock, Lange and Ozbay (2013) examine the behavior of dictators after introducing noise into their decisions.

[5] The relevance of reciprocity in labor market relationships is shown in Akerlof (1982), Dufwenberg and Kirchsteiger (2000), Krueger and Mas (2004), Dohmen et al. (2009), Brandes and Franck (2012), Kube, Marëchal and Puppe (2012), Cohn, Fehr and Goette (2014), or Gilchrist, Luca and Malhotra (2016) among others. Moreover, evidence has shown that when effort is involved and punishments are conducted personally, as it frequently occurs in the workplace, punishments are harsher than in the classical lab dictator games (see Dankova and Servatka, 2015 and Duersch and Müller, 2015 respectively).

The results of this paper also suggest what to expect in situations in which reciprocity has a bite but distributional concerns play no role. An example of this are consumer reviews. Reviews are in essence a reciprocal effort in which customers invest time to reward or punish the service provider. They have gained a lot of relevance in consumer choice and most companies put in a great effort to improve them. According to the results of this paper, consumers might forgive an issue with a service (e.g. a mistake in the check at a restaurant) if they think it was unintended, but they will punish it (writing a bad review), if they felt it was a scam. The perceived intention will determine the customer's reaction, who will punish the scam tentative even if it was not successful. Likewise, patient-physician relationships might be affected by the same logic. Feeling mistreated could be a reason to write a complaint after an intervention, even if the final outcome was positive. Conversely, a diligent attitude and a good relationship with the patient could protect physicians from bad unexpected outcomes. In light of my results, signaling good intentions could be as important for service providers as minimizing mistakes to avoid customer retaliation, that could go from poor reviews to more formal claims.

2. Experiment design

The experiment is designed as a dictator game with punishment opportunities. There are seven different treatments. In the first two treatments (uncertain reciprocal treatments 1 and 2) individual A (the dictator) chooses how to split €20 between herself and individual B (the respondent), with the particularity that the result of her choice may not be deterministic (one alternative is a lottery). In the next three treatments (nature treatments) a random device determines the allocation from the very beginning. Finally, in the last two treatments (certain reciprocal treatments), the dictator chooses among two deterministic outcomes. In all treatments, after observing the outcome and the choice of A (if there is any), individual B decides whether to add or subtract money to individual A, paying a cost.

Changing the lotteries that are available for the dictator in uncertain reciprocal treatments 1 and 2 enables me to test what happens when the outcome is unintendedly good, and what happens when it is unintendedly bad. The nature treatments makes it possible to distinguish distributional or efficiency concerns from the cognitive biases individual B might suffer when she judges individual A's actions (see Section 3 for the details). The certain treatments work as a baseline that allow verifying if uncertainty affects responses in any other manner.

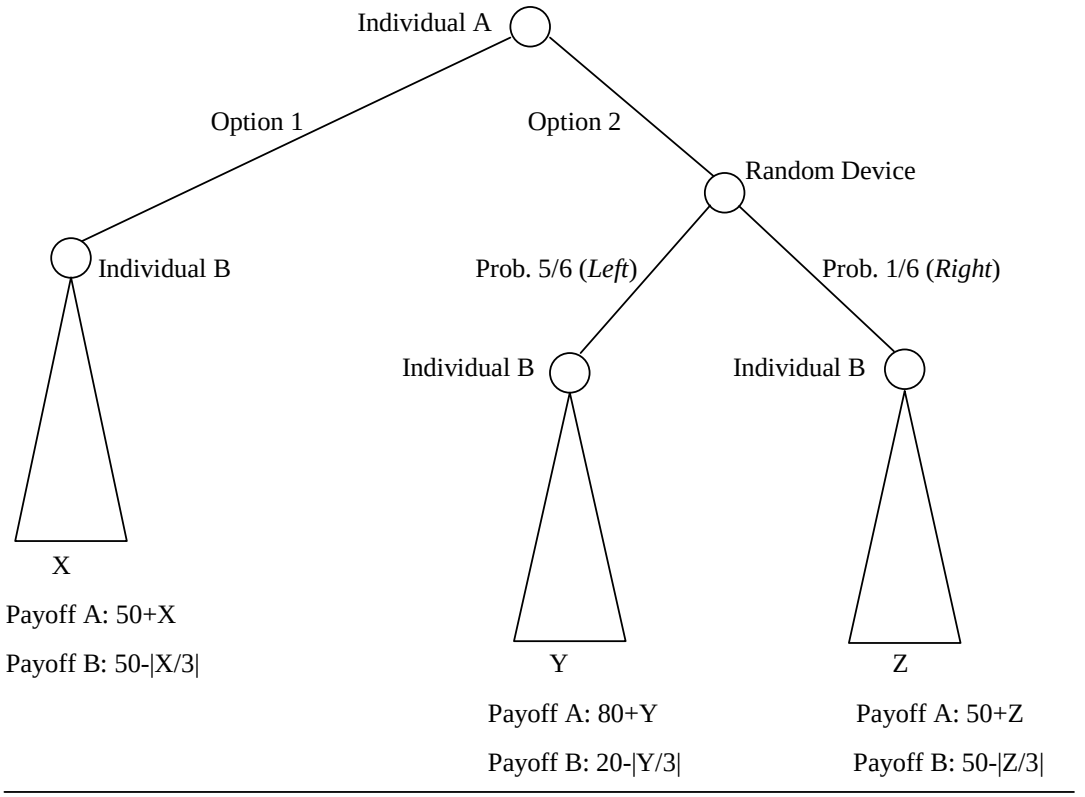
I conducted the experiment in the LEE UC3M lab at Carlos III University (Madrid, Spain). Subjects who participated in the experiment were undergraduate students of various degrees, ranging from engineering to journalism. They were recruited using ORSEE (Greiner, 2015). To run the experiment I used z-Tree experimental software (Fischbacher, 2007). Treatments were presented and explained to participants, one at a time. In every

treatment, pairs were randomly matched and the role of each participant was private information. Subjects were paid a show-up fee of €5 plus €0.20 for each experiment point earned during the game.

2.1. Uncertain reciprocal treatment 1

In the first treatment, individual A (the dictator) must decide between two options to split 100 points between herself and individual B (the respondent) with whom she has been paired. At a cost, individual B then chooses how many positive or negative points she wishes to assign to individual A contingent on this decision.

FIGURE 1
UNCERTAIN RECIPROCAL TREATMENT 1



X, Y and Z represent points awarded by B to A in each situation.

Figure 1 shows how this treatment works. If individual A chooses Option 1, 50 points go to herself and 50 to individual B. If she chooses Option 2, a die is rolled by the computer. There is a probability of 5/6 that the die will place them in Option 2 Left, where 80 points go to herself and 20 points to individual B. There is a probability of 1/6 that the die will place them in Option 2 Right, where 50 points go to herself and 50 points to individual B. This makes Option 1 the kind option and Option 2 the unkind option.

Contingent on this decision, individual B chooses how many points she wants to assign to individual A. Throughout the experiment, assigning 3x positive points has a cost

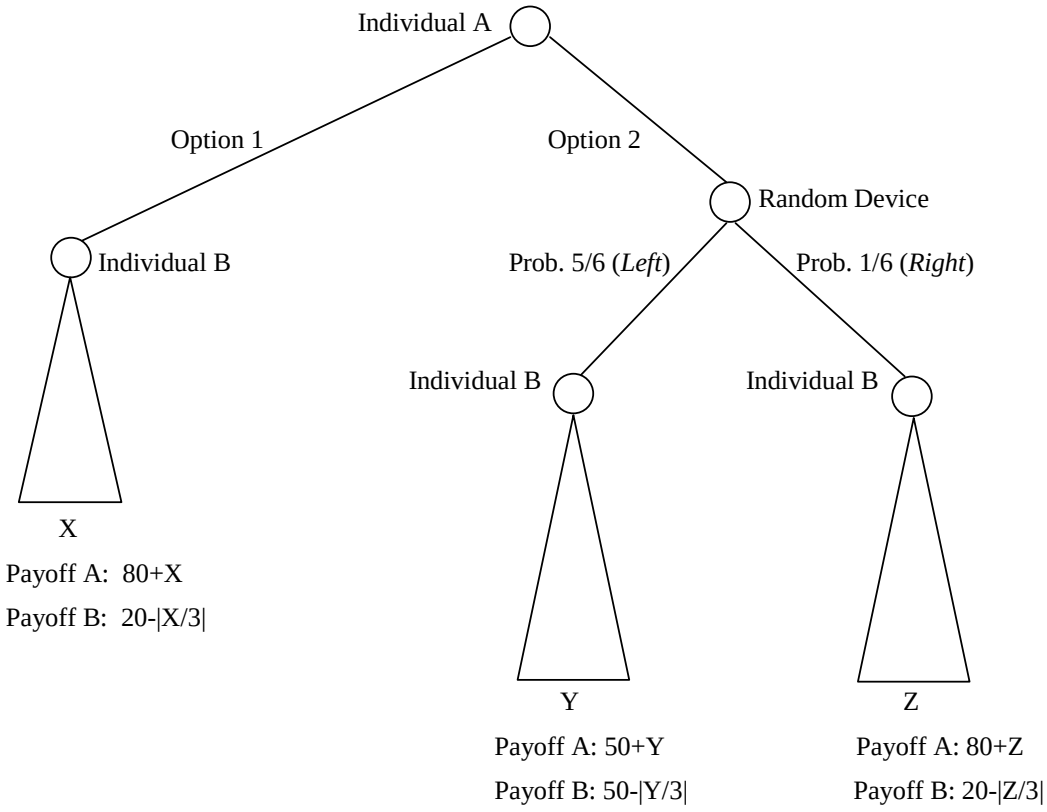
for individual B of 1x point. Assigning 3x negative points has a cost for individual B of 1x point. This means that both adding and subtracting points to/from A is costly for B. Allowing B to allocate positive and negative points avoids any experimenter demand effect for punishments.

The maximum number of positive points that individual B can assign to A is +48 and the maximum number of negative points is -48. Imposing these upper- and lower-bounds ensures that nobody ends up having negative payoffs. These limits hold for all treatments.

2.2. Uncertain reciprocal treatment 2

This treatment is similar to *uncertain reciprocal treatment 1*, but the options available to individual A are now different (Fig. 2). Randomness is introduced in the kind option rather than in the unkind. This enables additional predictions to be tested (see Section 3).

FIGURE 2
UNCERTAIN RECIPROCAL TREATMENT 2



X, Y and Z represent points awarded by B to A in each situation.

If Individual A chooses Option 1, 80 points go to herself and 20 to individual B. If she chooses Option 2, a die is rolled. There is a probability of 5/6 that the die will place them in Option 2 Left, where 50 points go to herself and 50 to individual B. There is a probability of 1/6 that the die will place them in Option 2 Right, where 80 points go to

herself and 20 to individual B. Then individual B chooses how many points she wants to assign to individual A.

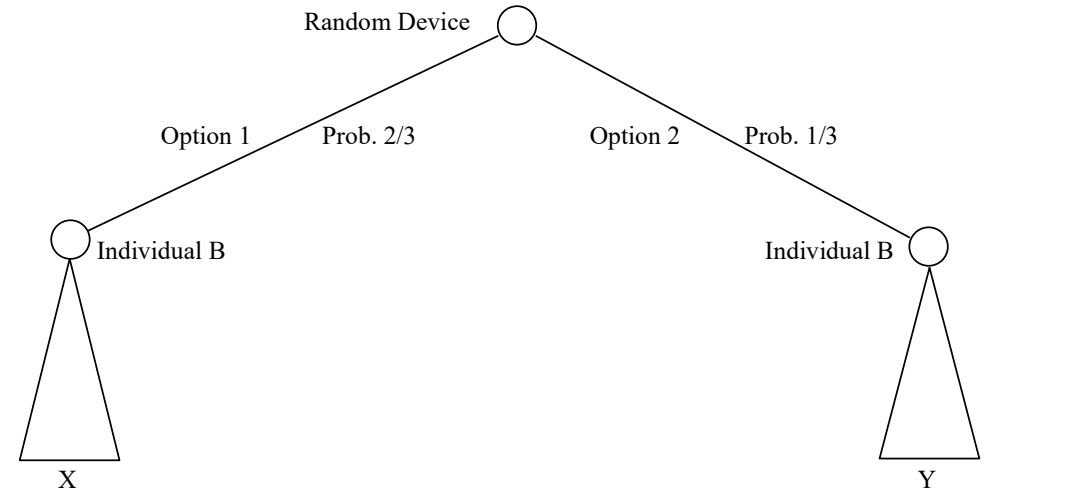
2.3. Nature treatments

In nature treatment 1, 100 points are again divided between A and B, but in this case a die determines how points are split without player A's participation (see Fig. 3). The die is rolled by the computer from the very beginning. There is a probability of 2/3 that individuals will be placed in Option 1, where 80 points go to individual A and 20 to individual B. There is a probability of 1/3 of them being placed in Option 2, where 50 points go to A and 50 points to B. Following Falk, Fehr and Fischbacher (2008), these probabilities were chosen so as to roughly mimic the decisions taken by dictators in some initial pilots that I ran, so the random device is perceived as neutral.[6] After the die is rolled, individual B has the possibility of adding or subtracting points to/from individual A at a cost.

Figure 3

Nature Treatment 1

Figure 3: Nature treatment



X and Y represent points awarded by B to A in each situation.

Nature treatment 2 maintains everything as nature treatment 1 except that in Option 1, individual A receives 75 points and individual B receives 25 points. Conversely, nature

[6] Bolton, Brandts and Ockenfels (2005) describe how people react to procedural fairness. The authors show that people judge random devices as "fair" or "unfair" and that this has an impact on the decisions of players B (for more information on procedural fairness see also Mertins, Egbert and Könen, 2013 and Mertins, 2008). Following Falk, Fehr and Fischbacher (2008) I assume that people would judge as neutral a random device that imitates reality. Another possible solution would have been to choose a 50-50 random device, a plausible focal point for a neutral device.

treatment 3 maintains everything as nature treatment 1 except Option 2, where individual A gets 55 points and individual B gets 45.

Notice that in these treatments there are no kind or unkind options, given that player A is making no decision. Hence, points assigned by B should not be thought of as punishments or rewards. Instead, they capture other departures from self-profit maximization (see Section 3).

2.4. Certain reciprocal treatments

In certain reciprocal treatments 1 and 2, individual A chooses between two options to split 100 points. In certain reciprocal treatment 1, if she chooses Option 1, she gets 75 points and B gets 25, if she chooses Option 2, both get 50. In certain reciprocal treatment 2, if she chooses Option 1, she gets 80 points and B gets 20, if she chooses Option 2, she gets 55 and B gets 45. As in previous treatments, after observing A's choice, B can add or subtract points to/from A at a cost. Notice that the outcomes of these treatments coincide with the expected outcomes of the uncertain reciprocal treatments.

2.5. Experiment procedure

186 people participated in the experiment. In the first five sessions, 66 participants played the two uncertain reciprocal treatments and nature treatment 1. In the next four sessions, 120 participants played all 7 treatments.

Before the game started, each subject was randomly assigned to role A or B. Participants maintained the same roles throughout the experiment and played once in each treatment. This is a key point in the design, as the impact of reciprocity is measured as deviations relative to behavior in the nature treatment (see Section 3 for details). Using a within-subject design makes it possible to control for individual fixed effects.

Contingent choices for each treatment were given one at a time. In order to avoid learning effects or reputation building, participants were not told the result of any treatment or any decision made by any player until the experiment was over.[7] In the same spirit as Brandts et al. (2015), to control for possible order effects I used a counterbalanced design. I randomized whether participants play first nature or reciprocal treatments, whether they play first certain or uncertain treatments, and for the subset that only played three treatments, whether they play first uncertain reciprocal treatment 1 or 2. I found no evidence of any order effects. Being aware of the existence of the nature (reciprocal) treatment does not introduce any demand effects on the reciprocal (nature) treatment, and the same applies to certain and uncertain treatments.

[7] In a survey, Brandts and Charness (2011) find evidence consistent with the strategy method having no or modest impact on results. Still, some authors think it elicits a cold instead of a hot response. This feature might make it inadequate to test theories of anger as Battigalli, Dufwenberg and Smith (2019), but it should be less problematic to explore reciprocity. Moreover, if it had an impact, it would affect all treatments, which would mitigate any confound effects.

To avoid wealth effects, participants were paid for one treatment chosen at random. Every treatment had the same probability of being the payment treatment, and that was public knowledge. After playing all treatments, I elicited the risk preferences of participants in sessions 6 to 9 using the incentivized game proposed in Holt and Laury (2002). After completing the risk elicitation experiment, participants took a short survey that included the streamline module developed by Falk et al. (2016) to measure self-reported altruism, trust, positive reciprocity, negative reciprocity and risk preferences.[8]

Individuals were seated in front of their terminals and given the instructions, which were read aloud so that there was no doubt that they were common to everyone. Instructions also included illustrative examples. Once the instructions had been read and any doubts clarified, but before the experiment actually started, they had to correctly answer a number of control questions to make sure that everybody had understood the experiment rules. The instructions for the main experiment and the control questions can be found in appendices A and B.

3. Predictions

When other-regarding preferences are studied, the same observed behavior can often be explained by different motivations. Hence, to ensure that the role of intentions in reciprocal behavior is pinned down correctly, it is necessary to control for competing explanations. In this section, I list the predictions of intention-based negative reciprocity together with what competing theories in the literature would predict in all possible scenarios. Then, comparing the behavior of individuals in different situations, I propose the tests to identify whether respondents suffer any cognitive bias when they judge dictators' choices.

Table 1 presents the full game. In each situation, different motivations lead to different optimal strategies for the respondent. Even though a subject's preferences might actually be a mixture of some of these motivations, for the sake of clarity in the exposition, I group them under three categories and describe them individually. Below, when the tests are presented, I account for the possibility that some of them might jointly affect the decisions of the respondent. Considering that assigning positive or negative points reduces one's own profits, a traditional self-profit-maximizer respondent would always assign zero points to the dictator. Nature Treatment 2 is omitted in the table because predictions in Nature Treatment 2 coincide with those in Nature Treatment 1.

The first category covers mutually exclusive preferences that always prescribe the same behavior. I refer to them as unconditional preferences. In both cases, the reason that explains behavior is the same: Positive and negative points have a cost of 1 to 3. This

[8]The existence of the risk preference elicitation experiment was unannounced. Earnings in this experiment were added to those of the main experiment. The instructions of the Holt and Laury game and the Falk et al. (2016) preference module can be found in appendices C and D.

	Uncertain R. Treat. 1			Uncertain R. Treat. 2			Nature Treat. 1		Nature Treat. 3		Certain R. Treat. 1		Certain R. Treat. 2	
	Option 1 (Kind)	Option 2 (Unkind) Pr. 5/6 *Left*	Pr. 1/6 *Right*	Option 1 (Unkind)	Option 2 (Kind) Pr. 5/6 *Left*	Pr. 1/6 *Right*	Option 1 Pr. 2/3	Option 2 Pr. 1/3	Option 1 Pr. 2/3	Option 2 Pr. 1/3	Option 1	Option 2	Option 1	Option 2
Individual A	50	80	50	80	50	80	80	50	80	55	75	50	80	55
Individual B	50	20	50	20	50	20	20	50	20	45	25	50	20	45
PREDICTIONS														
Self-Profit Maximization	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Spitefulness	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
Welfare Concerns	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)	(+)
Distributional Concerns	0	(-)	0	(-)	0	(-)	(-)	0	(-)	(-)	(-)	0	(-)	(-)
Intention-based Negative Reciprocity	0	(-)	(-)	(-)	0	0	0	0	0	0	(-)	0	(-)	0
Outcome-biased Negative Reciprocity	0	(-)	0	(-)	0	(-)	0	0	0	0	(-)	0	(-)	0
Label	NL50K	L20U	L50U	NL20U	L50K	L20K	NT1_20	NT1_50	NT3_20	NT3_45	CT25	CT50	CT20	CT45

means that respondents can actually harm others more than they harm themselves by always assigning negative points and can benefit others more than they harm themselves by always assigning positive points. Hence, if they are guided only by spitefulness (Klempt, 2012; Brañas-Garza et al., 2014) they should always give negative points to individual A. Similarly, if their goal is to maximize welfare (Kamas and Preston, 2012; Engelmann and Strobel, 2004, 2006; Charness and Rabin, 2002; Dufwenberg and Gneezy, 2000; Brandts et al., 2015), they should always give maximum positive points.[9]

The second category covers distributional preferences: Envy (Kirchsteiger, 1994; Cobo-Reyes and Jiménez, 2012), inequity aversion (Cox, 2004; Bolton and Ockenfels, 2006; Fehr, Naef and Schmidt, 2006; Falk, Fehr and Fischbacher, 2003; Yang, Onderstal and Schram, 2016; Brandts et al., 2015; Xiao and Bicchieri, 2010), and maximin preferences (Engelmann and Strobel, 2004, 2006). Respondents guided by distributional concerns maximize their utility by taking points away from A whenever they are behind and assigning zero points when they are even. They also rely on the 1 to 3 cost for negative points, but in this context it becomes a mechanism for reducing payoff inequality. Distributional preferences are conditional on outcomes, but as in the case of unconditional preferences, they do not depend on the dictator's participation.

Finally, the last category covers reciprocity (Brandts and Sola, 2001; Bolton, Brandts and Ockenfels, 1998; Falk, Fehr and Fischbacher, 2003; Nelson, 2002; Cox, 2004; Blount, 1995; Falk, Fehr and Fischbacher, 2008). Reciprocity does depend on the dictator's participation, as it prescribes a different reaction depending on her actions. It relies on the 1 to 3 cost as a punishment device. If negative reciprocity is intention-based, respondents will maximize their utility by punishing dictators after unkind choices and giving zero points otherwise. However, if it is true that people do not punish unkind choices after good outcomes (no-harm-no-foul), and/or it is true that they punish others after accidental bad outcomes, then once uncertainty is introduced reciprocal individuals become not intention-based but outcome-biased.[10]

In the next sub-sections I present the tests that support the exploration of the nature of other-regarding-behavior. I refer to each possible final situation using the labels indicated in Table 1. Labels are constructed using the following criterion. First, they indicate treatment, and in the case of the uncertain reciprocal treatment, they also indicate the choice: CT stands for certain reciprocal treatment, L stands for the option of the uncer-

[9]For the sake of clarity, I am describing the utility-maximizing behavior of individuals who are guided only by spitefulness/welfare concerns. However, in the literature these are never the sole motivations of agents. Later in the tests, I account for the fact that non-mutually exclusive motivations could appear jointly in the utility functions of participants.

[10]Whether reciprocal individuals responding to outcomes are really outcome-biased or consequence-based is a rather philosophical question. In the latter case, they would consciously disregard intentions and reward and punish the other people's actions depending on outcomes. In the former, they would do so by mistake. Psychology papers such as Baron and Hershey (1988) or Gino, Shu and Bazerman (2010) portray this behavior as a cognitive bias. I stick to this interpretation.

tain reciprocal treatment that involves a lottery, NL for the option that does not involve a lottery and NT1, NT2 and NT3 for nature treatments 1, 2 and 3. Then, they indicate how many points the option chosen by A assigns to B. Finally, in the case of the uncertain reciprocal treatment, they indicate whether the intention was kind (K) or unkind (U).

3.1. Does reciprocity matter?

First, I want to test whether the results in this experiment are in line with previous evidence. L20U, NL20U, CT25 and CT20 represent the situation that previous papers studying negative reciprocity have addressed, i.e. how people react to bad outcomes that result from unkind choices. Following McCabe, Rigdon and Smith (2003), Falk, Fehr and Fischbacher (2008), Blount (1995), and Gächter and Thoni (2010), I disentangle reciprocity from distributional preferences by testing whether punishments in reciprocal treatments are harsher than punishments in the nature treatment when the roll of the die leads to 80-20 or 75-25.

$H_0 : L20U, NL20U = NT1\_20$
$H_0 : CT20 = NT3\_20$
$H_0 : CT25 = NT2\_25$

If L20U, NL20U, CT20 and CT25 turn out to be more negative than NT1_20, NT3_20 and NT2_25 respectively, this would be evidence that negative reciprocity is making B punish A more aggressively, over and above any possible distributional concerns.[11]

3.2. Is intention-based reciprocity robust when outcomes and intentions are incongruous?

Once I confirm if negative reciprocity is relevant, I test whether it is still intention-based when outcomes and intentions are not congruent. This is the main contribution of the paper. I test whether subjects are forgiven after accidental bad outcomes, and whether they are punished if they intend to be selfish and fail.

First, I compare L50U and NT1_50.

$H_0 : L50U = NT1\_50$

If points assigned in a 50-50 outcome after an unkind choice are more negative than those assigned in the nature treatment, it means that when player A tries but fails to be greedy, she is still punished. This would reject the no-harm-no-foul hypothesis proposed in Bartling and Fischbacher (2012). Notice that NT_50 captures welfare maximizing concerns (if any). Therefore, if L50U is below NT_50, it means that unkind intentions of the dictator make those players B who are not efficiency maximizers punish, and those who are assign lower or negative points.

[11] NT2_25 corresponds to Nature Treatment 2 (75,25).

Likewise, if L20K is not below NT_20, then, when people try to be nice but unwittingly harm others, they are forgiven. Unintended bad outcomes are not punished beyond distributional concerns.

$H_0 : L20K = NT1\_20$

Both things would indicate that adding observable uncertainty does not change the fact that unkind intentions are a necessary and sufficient condition to explain the decision to reciprocate. Once we control for other motivations, the outcome bias documented by psychologists would not affect reciprocity when imposing punishments is costly.

3.3. Do consequences affect the intensity of punishments?

After establishing whether intentions are a sufficient and/or necessary condition for reciprocal punishments, I study the effect of outcomes on punishment intensity. If respondents evaluate unkind choices that lead to a bad outcome as worse, they may punish more severely. To test this hypothesis, I compare the difference between L20U and NT_20 with the difference between L50U and NT_50.

$H_0 : L20U - NT\_20 = L50U - NT\_50$

Notice that the expected outcomes of the dictator's choice after a kind or unkind action are not the same in uncertain reciprocal treatments 1 and 2. This can be regarded as a difference in the kindness/unkindness of these actions. Hence, to keep intentions constant, I fix the expected outcome comparing L20U with L50U. Moreover, to control for distributional and welfare concerns, I do a dif-in-dif subtracting NT_20 and NT_50 from L20U and L50U. If the difference between L20U and NT_20 turns out to be the same as the difference between L50U and NT_50, then, we cannot reject that subjects suffer no outcome bias when they decide the intensity of their punishments.

3.4. Do individuals have distributional concerns?

Finally, to explore if subjects have distributional concerns, I test whether points assigned in the nature treatments are negative when outcomes are unequal.

$H_0 : NT\_20, NT\_25 = 0$

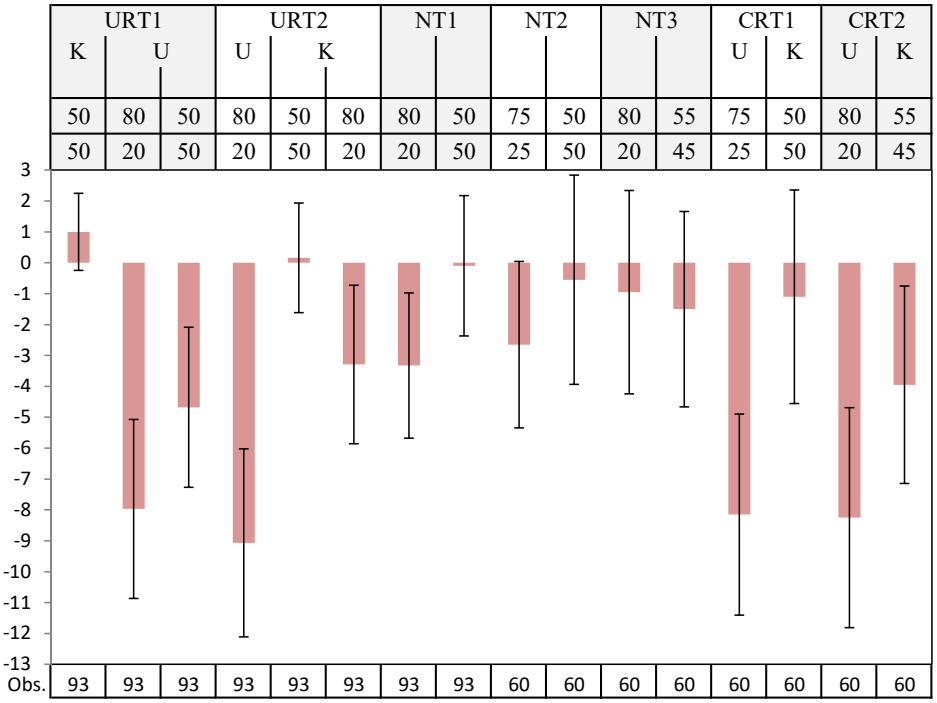
Several studies claim that individuals are willing to forgo wealth to rectify inequality (see for example Xiao and Bicchieri, 2010; Bolton and Ockenfels, 2006; Fehr, Naef and Schmidt, 2006). This is fundamentally different from suffering a cognitive bias when judging someone else's actions. People can judge actions only by their intentions and, at the same time, dislike inequality, irrespectively of whether someone can be held responsible for it (as proposed in the models of Ockenfels and Bolton (2000) and Fehr and Schmidt (1999)). The nature treatment allows me to capture this.

4. Results

In this section, I explore the robustness of intention-based reciprocity under uncertainty using non-parametric techniques and regression analysis. Figure 4 summarizes the average points allocated by B to A in every possible situation.[12] Looking at the graph, it is apparent that unkind intentions always lead to punishments, even when the final outcome is good for B. Moreover, it seems that when intentions are kind, negative points are not greater than those assigned solely on the grounds of distributional concerns.

FIGURE 4

AVERAGE POINTS ASSIGNED BY B TO A



Note: URT stands for Uncertain Reciprocal Treatment, NT for Nature Treatment, CRT for Certain Reciprocal Treatment, K for kind and U for unkind. Bands show 95% confidence intervals

In Table 2, I test formally the hypotheses introduced in Section 3. The Wilcoxon tests make a non-parametric comparison of the distributions across treatments, accounting for dependence within individual. They are considered to be well suited for experimental data, where sample size is often limited. I also report paired t-tests for completeness. Altogether, the tests provide a clear picture. Intention-based negative reciprocity matters and is robust to uncertainty. If outcomes and intentions are incongruous, reciprocal individuals still look at the latter when they judge a choice, not showing any cognitive biases. This coexists with some evidence of people being concerned about inequality.

[12] Overall, individual A chooses the kind option 29% times. Figure 5 in the appendix shows the statistics detailed by treatment.

In the four versions of test 1, I replicate results from previous studies (McCabe, Rigdon and Smith, 2003; Falk, Fehr and Fischbacher, 2008; Blount, 1995; Gächter and Thoni, 2010). I find strong evidence that negative reciprocity matters. When someone is unkind and the outcome is bad, responses go beyond those triggered by distributional concerns. To explore how reciprocity works when outcomes and intentions are misaligned, I use tests 2 and 3. Test 2 documents that points assigned by B in reciprocal treatment 1 when the unkind option is chosen and 50-50 is the outcome are significantly lower than those assigned in the nature treatment when 50-50 is the outcome. This rejects no-harm-no-foul hypothesis: Actions with unkind intentions do trigger punishments, even when the outcome is good for B. Conversely, test 3 shows that points assigned after a bad outcome when the intention is kind are not below points assigned to correct for payoff inequality in the nature treatment. Thus, the notion that people are forgiven when they try to be kind but do not succeed cannot be rejected. Test 2 and test 3 taken together indicate that when outcomes and intentions are incongruous, the former still explains reciprocity.

TABLE 2

TESTS

	Wilcoxon	t-test
Tests 1: Baseline Negative Reciprocity		
1A: $L20U = -7.97 < -3.32 = NT1\_20$	0.015	0.008
1B: $NL20U = -9.06 < -3.32 = NT1\_20$	0.017	0.001
1C: $CT20 = -8.25 < -0.95 = NT3\_20$	0.000	0.004
1D: $CT25 = -8.15 < -2.65 = NT2\_25$	0.002	0.005
Test 2: No-harm-no-foul		
$L50U = -4.68 < -0.1 = NT1\_50$	0.000	0.002
Test 3: Unintended Damage		
$L80K = -3.29 \not< -3.32 = NT1\_20$	0.628	0.98
Test 4: Outcomes and Punishment Intensity		
$L20U - NT1\_20 = -4.65 \not< -4.58 = L50U - NT1\_50$	0.471	0.973
Test 5: Distributional Concerns		
5A: $NT1\_20 = -3.32 < 0$	0.001	0.007
5B: $NT2\_25 = -2.65 < 0$	0.007	0.058
5C: $NT3\_20 = -0.95 < 0$	0.214	0.573

Note: Labels refer to the instances in which B has to make a decision. CT stands for certain reciprocal treatment, L for the option that involves a lottery in the uncertain reciprocal treatment, NL for the option that does not involve a lottery and NT1, NT2 and NT3 for nature treatments 1, 2 and 3. They are followed by the points awarded to B and by the intention -(K)ind or (U)nkind- (full description of the tests in Section 3). The column on the left indicates average points assigned by B to A in each situation. The two columns on the right show the p-value of Wilcoxon and paired t-test respectively.

The existence of a cognitive bias that could affect punishment intensity when outcomes are negative is assessed in test 4. It compares responses in the two situations that

could follow option 2 in reciprocal treatment 1, which holds intentions constant while varying outcomes. To control for distributional and efficiency concerns, it subtracts points assigned by B in the nature treatment. Results show that once unconditional preferences are controlled for, there is no evidence of a cognitive bias and outcomes have no effect on how B judges A. This seems to coexist with a preference for equality that is apparent when we compare L20U with L50U and L20K with L50K, as in both cases the difference coincides almost exactly with the difference between NT1_20 and NT1_50. To assess whether this might actually be a consequence of distributional concerns, test 5 looks at behavior in nature treatments, where reciprocal motivations are absent. Results in tests 5A and 5B support the relevance of these concerns. In nature treatments 1 and 2, individuals B assign negative points to individual A when the outcome is unfavorable to them and partially correct for payoff inequality. In nature treatment 3 points are also negative, even if they are not significantly different from zero (test 5C).[13]

Taken together, the non-parametric analysis provides a clear picture. Negative reciprocity matters and is driven by intentions. When outcomes and intentions are incongruous, the latter fully explains reciprocal responses. This seems to coexists with a preference for redistribution, which is independent of player A's taking any action.

4.1. Robustness Checks

To check the robustness of these results, I do a regression analysis. This allows me to test whether there are any order effects and to control for individual fixed effects. I also split the sample, analyzing separately the first 5 sessions and the next 4, to see if adding new treatments to the experiment makes a difference. Results are robust across specifications and coherent with those found using non-parametric methods.

$$\begin{aligned} Points_{ij} &= \alpha + \gamma Unven_{ij} + \beta_1 High_{ij} * Unkind_{ij} + \beta_2 High_{ij} * Kind_{ij} \\ &+ \beta_3 Low_{ij} * Unkind_{ij} + \beta_4 Low_{ij} * Kind_{ij} + \theta X_i + \varepsilon_{ij} \end{aligned} \tag{1}$$

I estimate coefficients in equation 1 using an OLS with errors clustered at the individual level. On the left-hand side, are the points assigned by individuals B to individuals A in each possible situation. On the right-hand side, there is a set of dummy variables: *Uneven* takes value 1 if the final outcome is not 50-50. In every treatment, there is a more unequal and a less unequal outcome. *High* takes value 1 if the final outcome is the more unequal, *Low* takes value 1 if it is the lesser. *Kind* takes a value of 1 if the action taken

[13] The p-value of the t-test in 5B is slightly above the level of 5%. However, the p-value of the Wilcoxon test, better suited for experimental data, is well below any conventional threshold. Something that distinguishes nature treatment 3 from 1 and 2 is that in the former, both options lead to an outcome that is unfavorable to player B. According to classical models of distributional concerns (Ockenfels and Bolton, 2000; Fehr and Schmidt, 1999), this should not make a difference. Why it seems to make a difference in my experiment is something that would require further investigation.

by individual A has kind intentions, while *Unkind* takes a value of 1 if it has unkind intentions -notice that the intention of an action can be kind or unkind, but can also be non-existent when there is no human intervention (in nature treatments)-. X_{it} controls for order and individual fixed effects.

TABLE 3
MAIN RESULTS WITH ORDER EFFECTS & FIXED EFFECTS

	(1) Sessions 1-5	(2) Sessions 6-9	(3) All	(4) Fixed Effects
Uneven Out.	-2.818**	-2.314**	-2.691***	-2.505***
	(1.041)	(0.908)	(0.853)	(0.876)
High*unkind	-7.136***	-5.724***	-5.881***	-6.022***
	(2.408)	(1.709)	(1.491)	(0.851)
High*kind	0.091	-1.662	-0.774	-1.085
	(2.290)	(1.584)	(1.387)	(1.256)
Low*unkind	-5.273**	-4.726**	-4.852***	-4.977***
	(2.330)	(1.849)	(1.498)	(1.327)
Low*kind	0.909	-0.560	-0.289	-0.281
	(1.552)	(1.216)	(1.043)	(0.877)
Certain First		1.206		
		(1.761)		
RT2 First	0.591			
	(2.262)			
Nature First	4.381	2.194	2.656	
	(2.399)	(1.761)	(1.472)	
Constant	-1.284	-1.924	-1.140	
	(1.220)	(1.854)	(1.169)	
Observations	264	960	1224	1224

Standard errors clustered at the individual level in parentheses
** p<0.05, *** p<0.001

The estimates of the coefficients should be interpreted as follows. γ captures distributional concerns. If it is negative and significant it means that respondents assign negative points to correct for payoff inequality. β_1 captures negative reciprocity. If it is negative it means that when there is a human being responsible for an action, the outcome is bad, and the intention is unkind, points allocated by the respondent are more negative than those assigned only on the grounds of distributional concerns. β_2 captures the effect of accidental bad outcomes. If it is not negative, then, those who try to be kind and fail are

not punished. β_3 tests the no-harm-no-foul hypothesis. If it is smaller than zero, then, when intentions are unkind player B still punishes player A, even if outcomes are good. Lastly, $Low * Kind$ is included to keep the specification econometrically sound and all the coefficients easy to interpret and X_i controls for order effects or individual fixed effects.

The estimated coefficients are presented in Table 3: The first column analyzes data from sessions 1 to 5, where participants only took uncertain reciprocal treatments 1 and 2 and nature treatment 1; the second column analyzes data from sessions 6 to 9, where participants played the full game; and columns 3 and 4 pool all observations together. I find results are robust across specifications. Players B hold distributional preferences and correct for payoff inequality, assigning negative points whenever the outcome is uneven. Moreover, when there is someone making a decision, they punish her if she is unkind, irrespectively of the outcome, but forgive her if she tries to be kind and fails. Estimates are robust to the inclusion of individual fixed effects (column 4) and order effects.[14] Altogether, conclusions from the regression analysis are aligned with those of the non-parametric tests: Negative reciprocity matters, it is driven by intentions, and when they are incongruous with outcomes respondents do not show any cognitive bias judging dictators. To explore if outcomes affect the intensity of punishments, I use equation 2.

Equation 1 does not allow to verify if a cognitive bias might have any effect on punishment intensity. To check this and assess the robustness of the result in test 4, I run the following specification:

$$Points_i = \alpha + \tau_u neven + \theta_1 Unkind_i + \theta_2 Kind_i + \phi High * Unkind_i \tag{2}$$

Column 2 in Table 4 presents the estimated coefficients of equation 2. A negative and significant τ shows participants have distributional concerns, while a negative θ_1 shows they punish unkind intentions. As in the non-parametric analysis, I find no evidence of negative outcomes having any impact on the intensity of punishments: ϕ is not significantly different from zero. This result is again robust to the inclusion of fixed effects (column 3).

4.2. Further Analysis

Once the robustness of results has been checked, I complement the analysis exploring if it makes any other difference in B responses whether A has full control over outcomes. Moreover, I assess how B's individual characteristics, including risk aversion, could explain her behavior.

Test 6 in Table 5 compares points assigned after an unkind choice when the result

[14] In sessions 1 to 5 (column 1), I randomize whether participants played first the nature treatment and then the reciprocal treatments as well as the order of uncertain reciprocal treatments 1 and 2. According to the results, neither of such manipulations had any significant effect on results. In sessions 6 to 9 (column 2), I randomize whether they played the nature treatment first and whether they played the certain or uncertain treatment first. Again, none of these manipulations had any impact on results.

Table 4

Outcomes, Intentions, and Cognitive Biases

	(1) Baseline	(2) Cog. Bias	(3) Cog. Bias+FE	(4) Uncertainty	(5) Uncer.+FE
Uneven Out.	-3.071***	-2.832***	-2.749***	-2.683***	-2.637***
	(0.906)	(0.820)	(0.759)	(0.822)	(0.707)
Unkind	-5.597***	-4.950***	-5.122***	-5.158***	-5.367***
	(1.335)	(1.520)	(1.301)	(1.386)	(0.941)
Kind	-0.504	-0.442	-0.526	-0.400	-0.609
	(0.971)	(0.969)	(0.759)	(1.089)	(0.941)
High*unkind		-0.883	-0.809		
		(1.283)	(1.456)		
Certain game				-2.184	-1.824
				(1.689)	(1.516)
Certain option				1.223	1.246
				(0.653)	(1.380)
Cer.-Game*unkind				3.048	3.071
				(2.400)	(2.082)
Cer.-Opt.*unkind				-2.624	-2.669
				(1.431)	(2.015)
Constant	0.426	0.273		0.177	
	(1.123)	(1.091)		(1.122)	
Observations	1224	1224	1224	1224	1224

Standard errors clustered at the individual level in parentheses

** p<0.05, *** p<0.001

Columns 3 and 5 include fixed effects

of the unkind action is certain and when it is not, holding intentions constant. Remember CT25 corresponds to points assigned after option 1 in certain reciprocal treatment 1, where A gets 75 points and B gets 25, and LU25 corresponds to average points assigned after option 2 in uncertain reciprocal treatment 1, where the expected outcomes are the same.[15] I find that as long as intentions coincide, choosing an option that entails uncertainty makes no difference in responses. Similarly, test 7 compares points assigned after a certain unkind choice, depending on whether the alternative is certain or random, holding intentions constant. Remember that in certain reciprocal treatment 2, A can choose keeping 80 points for herself and 20 for B (CT20) vis-à-vis keeping 55 for herself and 45 for B, while in uncertain reciprocal treatment 2, A can choose 80-20 (NL20U) vis-à-vis a lottery with expected outcome 55-45. The comparison shows that adding uncertainty to the alternative does not make a significant difference either. I supplement this analysis with column 4 in Table 4, which adds equation 2 a dummy equal 1 for certain treatments (that takes value 1 for certain reciprocal treatments one and two) and a dummy equal 1 for options that do not involve a lottery (that takes value 0 for L20U, L20K, L50U, L50K). Again, results show no significant effect and no interaction with intentions for any of them. Altogether, this suggests that uncertainty has no impact at all on reciprocal responses.

TABLE 5
UNCERTAINTY

	Wilcoxon	t-test
Test 6: Certain Unkind vs. Uncertain Unkind		
$CT25 = -8.15 < -7.87 = L25U$	0.357	0.837
Tests 7: Unkind with Certain Alternative vs Unkind with Uncertain Alternative		
$CT20 = -8.25 < -8.2 = NL20U$	0.905	0.970

Note: Labels refer to the instances in which B has to make a decision. CT stands for certain reciprocal treatment, L for the option that involves a lottery in the uncertain reciprocal treatment, NL for the option that does not involve a lottery. They are followed by the points awarded to B and by the intention -(K)ind or (U)nkind-. The column on the left indicates average points assigned by B to A in each situation. L25U is the average of L20U and L50U weighted by their respective probability: 5/6 and 1/6. The two columns on the right show the p-value of Wilcoxon and paired t-test respectively. The table only includes observations from sessions 6 to 9, where players played all treatments.

I finish the analysis with a correlational study of individual characteristics and decisions in the experiment (see Table 6 in the appendix).[16] Column 1 uses the risk aversion

[15] To compute the average, I multiply points assigned in L20U and L50U by 5/6 and 1/6 respectively.

[16] The relation between social preferences and individual characteristics have been addressed in several studies. Müller and Rau (2016) document a positive relation between risk aversion and inequity aversion, Croson and Gneezy (2009) conduct a survey of the literature and find mixed results for the relation between gender and generosity and some papers starting with Marwell and Ames (1981) have observed that eco-

measure of the incentivized experiment, while column 2 uses the self-reported measure on the survey. Both regressions include a dummy that indicates whether B is a female, a dummy indicating if he/she is an economist and self-reported measures of trust, altruism, positive reciprocity, and negative reciprocity. The only variable that explains points assigned in the experiment is the self-reported measure of negative reciprocity, which has a negative and significant coefficient. To explore if risk aversion could have an incidence only in games or in options that involve uncertainty, I replicate the regression in column 4 Table 4 including a measure of risk aversion and an interaction between risk aversion and the variables *certain game* and *certain option*. I find no significant effect nor for the incentivized measure of risk aversion neither for the self-reported measure (not shown). The rest of the results remain the same.

5. Discussion

This paper sets out to study the robustness of intention-based negative reciprocity in a context of uncertainty. Using a dictator game with punishment opportunities, I show that reciprocity is still intention-based when dictators only have partial control over the consequences of their actions. By means of non-parametric tests and regression analysis, I reject the no-harm-no-foul hypothesis and observe forgiveness after unintended bad outcomes. In line with the extension by Sebald (2010) of the model of sequential reciprocity in Dufwenberg and Kirchsteiger (2004), respondents identify expected outcomes as intentions and decide punishments in accordance. This coexists with a preference for redistribution that leads participants to forgo earnings to make final allocations more equitable.

Probably as a consequence of the experiment design, I find no evidence of rewards. Although I allow subjects to assign positive and negative points to avoid any demand effects, the kind option never favors the respondent (fifty-fifty is as good as it can get). Arguably, this is not enough to trigger positive reciprocity, even when participants in the survey declare to follow both positive and negative reciprocity rules. To learn how rewards are affected by observable uncertainty, we can look at results in Charness and Levine (2007). They gather evidence on the existence of positive reciprocity in a gift exchange game where proposers have only partial control over outcomes and observe that rewards are decided looking at intentions, also in line with Sebald (2010). I find their evidence complementary to mine, as gift exchange games are better suited to study positive reciprocity while my design targets negative reciprocity.

The decision to reward and punish looking at intentions in one-shot interactions is consistent with the dominant strategy in repeated games. Rand, Fudenberg and Dreber (2015) show that in a repeated interaction setting, where agents do not have perfect control over outcomes but intentions are observable, cooperation is achieved punishing

nomists tend to be more self-profit oriented.

selfish intentions. This coincides with what we observe in formal legal systems, where intentions play a major role -a citizen can be charged and imprisoned for an attempted murder, even if no harm was done to the victim, but she would be absolved if she drives over an oil spill, loses control of her vehicle, and runs over someone-. Interestingly, I find that subjects use this same logic to decide their punishments even when dynamic considerations are absent.

The results of this paper apply to economic problems where intentions are observable but not contractible and where reciprocity is a relevant force. They could describe the essence of consumer reviews and characterize important features of employer-employee and physician-patient relationships. Moreover, they could also explain why politicians strive to attribute the failure of government policies to external forces (Gasper and Reeves, 2011). Following the conclusions of this paper, if subjects perceive that a crisis is a result of the politician's unkind intentions, their reaction would go beyond that triggered simply by distributional concerns. Indeed, unkind intentions alone, without the mediation of a crisis, would be enough motive for electors to punish political leaders. If we extrapolate the results of the experiment, we could conclude that when others can retaliate, signaling virtue could be as important as signaling competence.

References

Akerlof, George. 1982. "Labor Contracts as Partial Gift Exchange." *The Quarterly Journal of Economics*, 97(4): 543–569.

Andreoni, James, Deniz Aydin, Blake Barton, B Douglas Bernheim, and Jeffrey Naecker. 2020. "When Fair Isn't Fair: Understanding Choice Reversals Involving Social Preferences." *Journal of Political Economy*, 128(5): 1673–1711.

Baron, Jonathan, and John Hershey. 1988. "Outcome bias in decision evaluation." *Journal of personality and social psychology*, 54(4): 569.

Bartling, Björn, Urs Fischbacher, and Simeon Schudy. 2015. "Pivotality and responsibility attribution in sequential voting." *Journal of Public Economics*, 128: 133–139.

Bartling, Björn, and Urs Fischbacher. 2012. "Shifting the Blame: On Delegation and Responsibility." *Review of Economic Studies*, 79(1): 67–87.

Battigalli, Pierpaolo, Martin Dufwenberg, and Alec Smith. 2019. "Frustration, aggression, and anger in leader-follower games." *Games and Economic Behavior*, 117: 15–39.

Bereby-Meyer, Yoella, and Alvin Roth. 2006. "The Speed of Learning in Noisy Games: Partial Reinforcement and the Sustainability of Cooperation." *American Economic Review*, 96(4): 1029–1042.

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, and Axel Ockenfels. 2006. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment." *American Economic Review*, 96(5): 1906–1911.

Bolton, Gary, Jordi Brandts, and Axel Ockenfels. 1998. "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game." *Experimental Economics*, 1(3): 207–219.

Bolton, Gary, Jordi Brandts, and Axel Ockenfels. 2005. "Fair Procedures: Evidence from Games Involving Lotteries." *Economic Journal*, 115(506): 1054–1076.

Brañas-Garza, Pablo, Antonio M Espín, Filippos Exadaktylos, and Benedikt Herrmann. 2014. "Fair and unfair punishers coexist in the Ultimatum Game." *Scientific reports*, 4: 6025.

Brandes, Leif, and Egon Franck. 2012. "Social preferences or personal career concerns? Field evidence on positive and negative reciprocity in the workplace." *Journal of Economic Psychology*, 33(5): 925–939.

Brandts, Jordi, and Carles Sola. 2001. "Reference Points and Negative Reciprocity in Simple Sequential Games." *Games and Economic Behavior*, 36(2): 138–157.

Brandts, Jordi, and Gary Charness. 2011. "The strategy versus the direct-response method: a first survey of experimental comparisons." *Experimental Economics*, 14(3): 375–398.

Brandts, Jordi, Enrique Fatas, Ernan Haruvy, and Francisco Lagos. 2015. "The impact of relative position and returns on sacrifice and reciprocity: an experimental study using individual decisions." *Social Choice and Welfare*, 45(3): 489–511.

Brock, J. Michelle, Andreas Lange, and Erkut Ozbay. 2013. "Dictating the Risk: Experimental Evidence on Giving in Risky Environments." *American Economic Review*, 103(1): 415–37.

Brownback, Andy, and Michael A Kuhn. 2019. "Understanding outcome bias." *Games and Economic Behavior*, 117: 342–360.

Cappelen, Alexander, James Konow, Erik Ø. Sørensen, and Bertil Tungodden. 2013. "Just Luck: An Experimental Study of Risk-Taking and Fairness." *American Economic Review*, 103(4): 1398–1413.

Cettolin, Elena, and Arno Riedl. 2017. "Justice under uncertainty." *Management Science*, 63(11): 3739–3759.

Charness, Gary, and David Levine. 2007. "Intention and Stochastic Outcomes: An Experimental study." *Economic Journal*, 117(522): 1051–1072.

Charness, Gary, and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests." *The Quarterly Journal of Economics*, 117(3): 817–869.

Cobo-Reyes, Ramón, and Natalia Jiménez. 2012. "The dark side of friendship: envy." *Experimental Economics*, 15(4): 547–570.

Cohn, Alain, Ernst Fehr, and Lorenz Goette. 2014. "Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment." *Management Science*, 61(8): 1777–1794.

Cooper, David, and John Kagel. 2016. "Other Regarding Preferences: A Selective Survey of Experimental Results." In *Handbook of Experimental Economics vol. 2*. Vol. 2, , ed. John Kagel and Alvin Roth, Chapter 4, 217–289. Princeton University Press.

Cox, James. 2004. "How to identify trust and reciprocity." *Games and Economic Behavior*, 46(2): 260–281.

Croson, Rachel, and Uri Gneezy. 2009. "Gender differences in preferences." *Journal of Economic literature*, 47(2): 448–74.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang. 2007. "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness." *Economic Theory*, 33(1): 67–80.

Dankova, Katarina, and Maros Servatka. 2015. "The house money effect and negative reciprocity." *Journal of Economic Psychology*, 48: 60–71.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2008. "Representative Trust and Reciprocity: Prevalence and Determinants." *Economic Inquiry*, 46(1): 84–90.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2009. "Homo Reciprocans: Survey Evidence on Behavioural Outcomes." *Economic Journal*, 119(536): 592–612.

Duersch, Peter, and Julia Müller. 2015. "Taking punishment into your own hands: An experiment." *Journal of Economic Psychology*, 46: 1–11.

Dufwenberg, Martin, and Georg Kirchsteiger. 2000. "Reciprocity and wage undercutting." *European Economic Review*, 44(4): 1069–1078.

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

Dufwenberg, Martin, and Uri Gneezy. 2000. "Measuring beliefs in an experimental lost wallet game." *Games and economic Behavior*, 30(2): 163–182.

Engelmann, Dirk, and Martin Strobel. 2004. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments." *American Economic Review*, 94(4): 857–869.

Engelmann, Dirk, and Martin Strobel. 2006. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Reply." *American Economic Review*, 96(5): 1918–1923.

Falk, Armin, Anke Becker, Thomas J Dohmen, David Huffman, and Uwe Sunde. 2016. "The preference survey module: A validated instrument for measuring risk, time, and social preferences."

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." *The Quarterly Journal of Economics*, 114(3): 817–868.

Fehr, Ernst, and Klaus M. Schmidt. 2006. "Chapter 8 The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories." In *Foundations*. Vol. 1 of *Handbook of the Economics of Giving, Altruism and Reciprocity*, , ed. Serge-Christophe Kolm and Jean Mercier Ythier, 615 – 691. Elsevier.

Fehr, Ernst, Michael Naef, and Klaus Schmidt. 2006. "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments: Comment." *American Economic Review*, 96(5): 1912–1917.

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

Friehe, Tim, and Verena Utikal. 2018. "Intentions under cover–Hiding intentions is considered unfair." *Journal of behavioral and experimental economics*, 73: 11–21.

Fudenberg, Drew, and David K Levine. 2012. "Fairness, risk preferences and independence: Impossibility theorems." *Journal of Economic Behavior & Organization*, 81(2): 606–612.

Gächter, Simon, and Christian Thoni. 2010. "Social comparison and performance: Experimental evidence on the fair wage-effort hypothesis." *Journal of Economic Behavior & Organization*, 76(3): 531–543.

Gasper, John T, and Andrew Reeves. 2011. "Make it Rain? Retrospection and Attentive Electorate in the Context of Natural Disasters." *American Journal of Political Science*, 55(2): 340–355.

Gilchrist, Duncan S, Michael Luca, and Deepak Malhotra. 2016. "When 3+ 1> 4: Gift structure and reciprocity in the field." *Management Science*, 62(9): 2639–2650.

Gino, Francesca, Lisa L Shu, and Max H Bazerman. 2010. "Nameless+ harmless= blameless: When seemingly irrelevant factors influence judgment of (un) ethical behavior." *Organizational Behavior and Human Decision Processes*, 111(2): 93–101.

Greiner, Ben. 2015. "Subject pool recruitment procedures: organizing experiments with ORSEE." *Journal of the Economic Science Association*, 1(1): 114–125.

Gurdal, Mehmet, Joshua B. Miller, and Aldo Rustichini. 2013. "Why Blame?" *Journal of Political Economy*, 121(6): 1205 – 1247.

Holt, Charles A, and Susan K Laury. 2002. "Risk aversion and incentive effects." *American economic review*, 92(5): 1644–1655.

Kamas, Linda, and Anne Preston. 2012. "Distributive and reciprocal fairness: What can we learn from the heterogeneity of social preferences?" *Journal of Economic Psychology*, 33(3): 538–553.

Kirchsteiger, Georg. 1994. "The role of envy in ultimatum games." *Journal of economic behavior & organization*, 25(3): 373–389.

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

Krawczyk, Michal, and Fabrice Le Lec. 2010. "Give me a chance! An experiment in social decision under risk." *Experimental Economics*, 13(4): 500–511.

Krueger, Alan, and Alexandre Mas. 2004. "Strikes, Scabs, and Tread Separations: Labor Strife and the Production of Defective Bridgestone/Firestone Tires." *Journal of Political Economy*, 112(2): 253–289.

Kube, Sebastian, Michel Marëchal, and Clemens Puppe. 2012. "The Currency of Reciprocity: Gift Exchange in the Workplace." *American Economic Review*, 102(4): 1644–62.

Markussen, Thomas, Louis Putterman, and Jean-Robert Tyran. 2016. "Judicial error and cooperation." *European Economic Review*, 89: 372–388.

Marwell, Gerald, and Ruth E Ames. 1981. "Economists free ride, does anyone else." *Journal of public economics*, 15(3): 295–310.

McCabe, Kevin, Mary Rigdon, and Vernon Smith. 2003. "Positive reciprocity and intentions in trust games." *Journal of Economic Behavior & Organization*, 52(2): 267–275.

Mertins, Vanessa. 2008. "The effects of Procedures on Social Interaction: A Literature Review." Institute of Labour Law and Industrial Relations in the European Union (IAAEU) IAAEG Discussion Papers until 2011 200806.

Mertins, Vanessa, Henrik Egbert, and Tanja Könen. 2013. "The effects of individual judgments about selection procedures: Results from a power-to-resist game." *Journal of Behavioral and Experimental Economics (formerly The Journal of Socio-Economics)*, 42: 112–120.

Müller, Stephan, and Holger A Rau. 2016. "The relation of risk attitudes and other-regarding preferences: A within-subjects analysis." *European Economic Review*, 85: 1–7.

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

Ockenfels, Axel, and Gary Bolton. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90(1): 166–193.

Offerman, Theo. 2002. "Hurting Hurts More than Helping Helps." *European Economic Review*, 46(8): 1423–1437.

Rand, David G., Drew Fudenberg, and Anna Dreber. 2015. "It's the thought that counts: The role of intentions in noisy repeated games." *Journal of Economic Behavior & Organization*, 116: 481–499.

Rubin, Jared, and Roman Sheremeta. 2015. "Principal–agent settings with random shocks." *Management Science*, 62(4): 985–999.

Saito, Kota. 2013. "Social preferences under risk: Equality of opportunity versus equality of outcome." *American Economic Review*, 103(7): 3084–3101.

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

Stanca, Luca, Luigino Bruni, and Luca Corazzini. 2009. "Testing Theories of Reciprocity: Do Motivations Matter?" *Journal of Economic Behavior & Organization*, 71(2): 233–245.

Sutter, Matthias. 2007. "Outcomes versus intentions: On the nature of fair behavior and its development with age." *Journal of Economic Psychology*, 28(1): 69–78.

Trautmann, Stefan T, and Gijs van de Kuilen. 2016. "Process fairness, outcome fairness, and dynamic consistency: Experimental evidence for risk and ambiguity." *Journal of Risk and Uncertainty*, 53(2-3): 75–88.

Xiao, Erte, and Cristina Bicchieri. 2010. "When equality trumps reciprocity." *Journal of Economic Psychology*, 31(3): 456–470.

Xiao, Erte, and Howard Kunreuther. 2016. "Punishment and cooperation in stochastic social dilemmas." *Journal of Conflict Resolution*, 60(4): 670–693.

Yang, Yang, Sander Onderstal, and Arthur Schram. 2016. "Inequity aversion revisited." *Journal of Economic Psychology*, 54: 1–16.

6. Appendix

6.1. Tables and Graphs

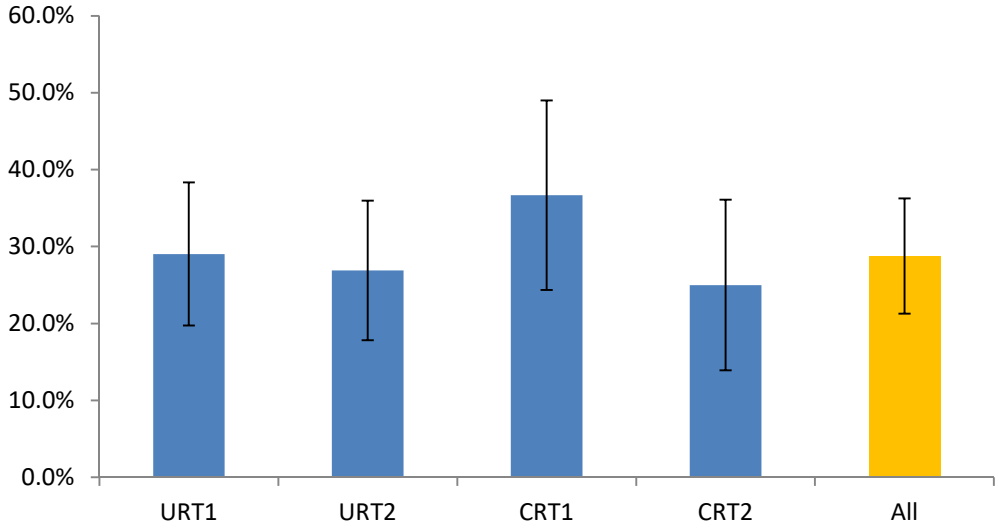
TABLE 6
INDIVIDUAL CHARACTERISTICS

	(1) Incentivized RA	(2) Survey RA
Incentivized RA	0.647	
	(0.356)	
Survey RA		-0.038
		(0.408)
Survey Trust	-0.057	-0.018
	(0.393)	(0.364)
Survey Pos Rec	0.340	0.302
	(0.464)	(0.479)
Survey Neg Rec	-0.979***	-0.926***
	(0.217)	(0.263)
Survey Altruism	-0.274	-0.242
	(0.328)	(0.370)
Female	-3.270	-1.918
	(1.639)	(1.789)
Economist	1.918	2.922
	(1.645)	(1.620)
Constant	-4.500	-0.980
	(6.845)	(6.520)
Observations	928	960

Note: Out of 60 respondents, 2 made an inconsistent choice in the incentiviced risk elicitation phase. I drop them from the sample of regression in column 1.
Standard errors clustered at the individual level in parentheses.
** p<0.05, *** p<0.001

Figure 5

Frequency of Kind Option



Note: URT1 and URT2 stand for Uncertain Reciprocal Treatment 1 and 2, CRT1 and CRT2 for Certain Reciprocal Treatment 1 and 2, 'All' is the average across treatments. There are 93 observations for URT1 and URT2 and 60 for CRT1 and CRT2. Bands show 95% confidence intervals.