Intention-Based Reciprocity and Signaling of Intentions*

Séverine Toussaert†

March 2, 2017

Intentions count in your actions. Abu Bakr

Abstract

Many experiments find that trust intentions are a key determinant of prosociality. If intentions matter, then prosociality should depend on whether trust intentions can be credibly conveyed. This conjecture is formalized and tested in a noisy trust game where I vary the extent to which trust can be credibly signaled. I find that the introduction of noise threatens the onset of trust relations and induces players to form more pessimistic beliefs. Therefore policies that increase transparency of the decision-making environment may foster prosociality. However, the potential impact of such policies could be limited by a large heterogeneity in how individuals respond to changes in their information environment.

JEL classification: C91, D63, D64

Keywords: Trust; Intentions; Reciprocity; Noise; Signaling; Experimental economics

*I am indebted to David Cesarini and Guillaume Fréchette for their guidance and continued support in the development of this project. I also thank Anwar Ruff for teaching me how to program experiments, Sevgi Yüksel and Elliot Lipnowski for numerous discussions, as well as Andrew Schotter, Tore Ellingsen, Magnus Johannesson, Sebastian Fehrler, Gary Charness, Pierpaolo Battigalli, two anonymous referees and participants at the 2013 North American ESA Conference. Finally, I would like to acknowledge financial support from the National Science Foundation via grant SES-1260891 as well as from the Center for Experimental Social Science of New York University. All remaining errors are mine.

†London School of Economics, Department of Social Policy. Email: S.Toussaert@lse.ac.uk

1. Introduction

Trust is often perceived as an essential element of social capital (Putnam 1995) as well as a "lubricant" for economic transactions (Arrow 1974). A central feature of trust relations is their reliance on an informal agreement rather than on an explicit contract whose terms are costly to enforce. Due to their implicit nature, relationships of trust are subject to all sorts of uncertainty, which may threaten the very establishment of trust. Understanding the conditions under which trust can arise seems therefore of particular importance.

Trust relations have been widely studied in the context of two-person experimental trust games between a Sender and a Receiver. In this class of games, relationships of trust are characterized by the following three conditions: (i) both parties can mutually benefit from the relationship, (ii) the Sender takes a risk by trusting the Receiver, (iii) the Receiver incurs a monetary cost by reciprocating the Sender's trust (McCabe et al. 2003). It is a robust finding that a large proportion of players deviate from the subgame perfect equilibrium of no trust predicted by standard non-cooperative theory (Berg et al. 1995, McCabe et al. 1998, Camerer 2003).

One mechanism proposed to explain these findings is that Receivers like to reciprocate the risk taken by the Sender for it signals his intention to trust. Many studies provide credence to this interpretation by reporting evidence of intention-based reciprocity in a variety of settings, including trust games and other well-known social dilemma games.[1] These studies typically compare an Intention Treatment, where the first mover's action was chosen intentionally, to a No-Intention treatment, where the first mover's action was either implemented by a random device or was the only option available.[2] For instance, McCabe et al. (2003) find that the percentage of Receivers who share is twice as high when the Sender's decision to trust was voluntary rather than involuntary.

Although there is evidence that intentions matter, almost all existing studies restrict attention to environments of full information, where actions perfectly reflect intentions. By doing so, they abstract away from the uncertainty inherent in trust relationships, where decisions often account for many implicit considerations and constraints, which render intentions hard to read. In this respect, one could worry that reciprocity considerations might disappear as soon as actions become noisy signals of trust, which in turn might deter trust.

To illustrate this point, consider the hiring decisions of a private company or university research

[1] See Blount (1995) in an ultimatum game, McCabe et al. (2003) and Cox et al. (2006) in trust games, Charness (2004) and Charness and Levine (2007) in gift exchange games, Rand et al. (2015) in a repeated prisoner's dilemma, or Offerman (2002), Falk et al. (2008) and Cox et al. (2008) in sequential games allowing for both positive and negative reciprocal behavior.

[2] Alternatively, some papers test for intention-based reciprocity by comparing treatments which differ in the alternative(s) forgone by first movers, as varying their outside options may lead second movers to make different inferences about their partner's intentions (see Brandts and Solà 2000, Falk et al. 2003 or Charness and Rabin 2002, 2005).

department. After several rounds of interviews, the preferred candidate receives a job offer; if he/she turns down the offer, the second best candidate on the list is offered the job and the process continues until a candidate accepts. Under those circumstances, one might expect the motivation of the new recruit to be higher if the job offer clearly signals an act of trust in his/her ability to perform successfully. The strength of this signal will generally depend on the level of uncertainty around the selection process, including the number of candidates who previously turned down the offer. If successful candidates feel less invested when the selection procedure is opaque, then hiring committees may in turn lack trust in their own recruits.

The purpose of this paper is to assess the extent to which the onset of trust relations might depend on the credibility with which intentions to trust can be conveyed. To study this question, I consider a noisy binary trust game where I vary the likelihood with which the Sender's action is implemented: with some common knowledge probability p, the Receiver faces the Sender's decision, while with probability $(1-p)$, the Sender's decision is replaced by the random choice of a computer. In the experiment, subjects make decisions under the strategy method for all values of p in $\{0, 0.1, ..., 1\}$, which allows to analyze the sensitivity of individual strategies to the amount of noise. For each value of p, I also elicit the Receiver's belief about the Sender's choice and the Sender's belief about the guess of his partner.[3]

If intentions matter, then prosocial behavior and beliefs should be affected by how much control the Sender possesses over the outcome. Intuitively, when p is low, the Sender's decision is unlikely to matter; in this case, no credible signal of trust can be sent and the Receiver should feel little inclined to act prosocially. On the other hand, when p is high, the Receiver is more likely to be facing the Sender's choice, which should increase his propensity to reciprocate. In turn, whether the Sender decides to trust should be affected by the possibility to credibly signal trust, which is more likely if p is high. I formalize this intuition in a simple model in which I isolate the behavioral implications of intention-based reciprocity relative to other concerns.

The present paper speaks to a burgeoning literature connecting prosociality to intentions in environments of imperfect information. Most connected to this study, two papers consider noisy binary games where the intended action is implemented with some probability p, and reversed otherwise.[4] In a binary trust game, Cox et al. (2006) find that the Receiver's degree of prosociality does not significantly decrease when the Sender's action is only implemented with probability $p = \frac{3}{4}$

[3]Besides this baseline treatment, the original design comprised two other treatments, which are not presented in this paper. In those treatments, the decision environment was more complex and artificial, making results harder to interpret. See Section 3 and the Online Appendix for more details about those two treatments.

[4]A few other papers link prosocial behavior to intentionality in games of *incomplete* information, i.e. games where there is a lack of common knowledge about other players' preferences or payoffs (Attanasi et al. 2015, McCabe et al. 1998). For instance, McCabe et al. (1998) find that when players only know their own payoffs, behavior closely follows Nash equilibrium predictions for rational and selfish agents.

and remains higher compared to when a coin flip decides for the Sender (i.e. $p = 0$); however, Senders do not anticipate this and trust significantly less. In a noisy repeated prisoner's dilemma with $p = \frac{7}{8}$, Rand et al. (2015) compare cooperation rates when intended actions are observable versus unobservable; they find that the cooperation rate is higher when intentions are observable and matches the rate observed when $p = 1$.

The present study complements the above papers in several ways. First, while both papers restrict attention to small departures from the perfect information case, this paper analyzes how prosocial behavior responds to changes in the amount of noise $(1 - p)$ over the entire spectrum. Second, this paper augments the binary choice data with players' beliefs in order to gather additional insights into how subjects interpret the noise. Finally, unlike previous studies, this paper exploits within-subject variation to identify the effect of noise on prosociality, thus allowing to assess whether individual heterogeneity exists in how subjects respond to changes in their decision-making environment.

In this experimental setting, I find that prosocial behavior and beliefs are affected by how much control the Sender has over the outcome. At the aggregate level, the trust-reciprocity outcome is less likely to emerge as the value of p decreases; furthermore, beliefs become progressively more pessimistic. Thus, whether trust intentions can be credibly communicated through actions matters for prosociality. While intentions matter, they do not seem to matter for everyone. The individual-level analysis indeed reveals substantial heterogeneity in how subjects respond to the noise, with a large fraction of players being irresponsive to p. Furthermore, if intentions seem to be a key determinant of prosociality, social image concerns appear to be almost as equally important to explain aggregate behavior.

The remainder of the paper is organized as follows. Section 2 introduces the noisy trust game on which the experiment is based and discusses the role of intentions in a simple model. Section 3 describes the experimental design, while Section 4 presents the main results. Section 5 concludes. Instructions and supplemental material can be found in an Online Appendix (henceforth OA) available on the author's website.

2. Intentions in a noisy trust game

In this section, I introduce the noisy trust game on which the experiment is based (Section 2.1) and discuss in a simple model the role of intentions relative to other prosocial concerns (Section 2.2).

2.1 The noisy binary trust game

Consider the noisy trust game represented in Figure 1 by the tree $\Gamma(p)$, where p parametrizes the amount of noise in the game. Payoffs are in dollars.

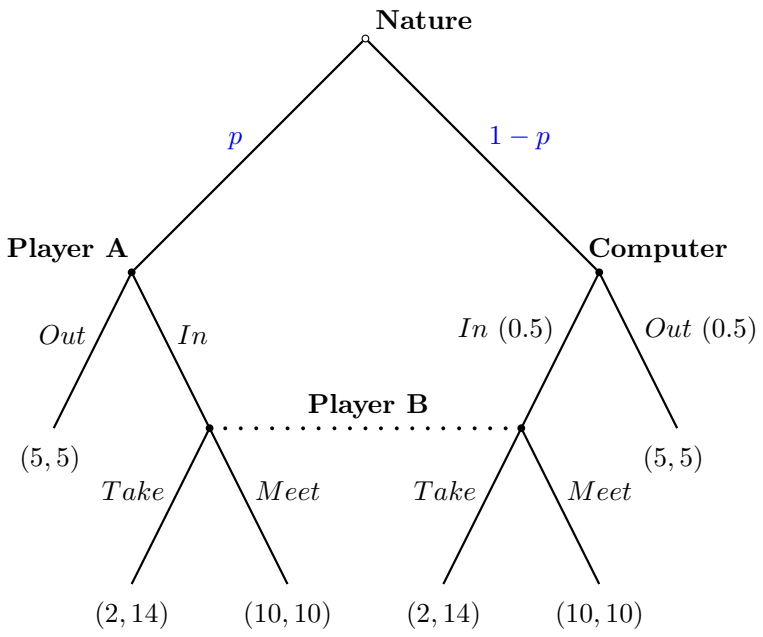


Figure 1: The game form $\Gamma(p)$

To understand this game, it is useful to consider the two special cases where $p = 0$ and $p = 1$, which have been widely studied in the literature. When $p = 1$ and this is common knowledge, the game simplifies to a standard mini trust game: A (the Sender) can choose to invest in a relationship with B (the Receiver) or to stay *Out* of the relationship; if A chooses In, B can choose to meet A's investment by sharing gains, or take most of the benefits for himself. At the other extreme, when $p = 0$, the game reduces to a mini dictator game: in this case, a computer randomly makes a choice for A by playing a mixed strategy 50/50 between In and *Out*. If B only cared about his own material payoffs, observed behavior would be identical in those two games: B would choose *Take* and by backward induction, A would go *Out* in the trust game. However, one typical finding is that a substantial fraction of A and B players exhibit prosocial behavior in those games. For instance, in the trust game with the most similar environment and payoff structure, Charness and Dufwenberg (2006) find that 44% of B players share and 56% of A players go In. Furthermore, it is a fairly robust result that B's propensity to share tends to be higher in the trust game than in the corresponding dictator game. For instance, Cox et al. (2006) find that 63.6% of Receivers cooperate

when $p = 1$, while this number drops to 35% when $p = 0$.[5] These observed behavioral differences are usually interpreted as reflecting B's reciprocity concerns in the trust game: by risking to go In, A signals his intention to enter in a relationship of trust with B, an act that B rewards by choosing $Meet$.

What would happen if some noise were introduced regarding the nature of the game? Consider the case where $p \in (0, 1)$ is common knowledge but players do not observe Nature's move. Furthermore, assume that B receives no information about the choice of A or the computer's draw. In this case, upon observing the realization of In, B cannot directly attribute this outcome to a trusting move of A. Under such circumstances, one would expect B's propensity to share to increase with p for two reasons. First, when p is low, B is less likely to be facing A's choice.[6] Second, even if B were facing A's action for certain, B would discount A's choice since choosing In presented a low risk for A in the first place: whatever A chooses, his fate will most likely be determined by the computer. In other words, when p is low, A's choice to go In is a weaker signal of trust than when p is high. Consequently, one would expect A's propensity to go In to depend on the credibility of the signal associated with this choice: if A believes in B's trustworthiness, A should be more confident about choosing In as p increases. This intuition is discussed more formally in the next section.

2.2 Intentions and prosociality in the noisy trust game

2.2.1 Strategies and beliefs

Let $\sigma_A \in \triangle(S_A)$ be player A's propensity to choose In in $\Gamma(p)$ where $S_A := \{In, Out\}$. Similarly, let $\sigma_B \in \triangle(S_B)$ be player B's propensity to choose $Meet$ in $\Gamma(p)$ where $S_B := \{Meet, Take\}$. Finally, let $\sigma_C = \left(\frac{1}{2}, \frac{1}{2}\right)$ denote the computer's mixed strategy and assume that σ_C is common knowledge.

In the following analysis, I assume that individuals only play pure strategies ($\sigma_i \in \{0, 1\}$, $i \in \{\text{A,B}\}$) and choose the prosocial action (In for A and $Meet$ for B) whenever indifferent. A mixed strategy $\sigma_i \in (0, 1)$ is then interpreted as coming from a statistical distribution of pure strategies played by individuals in role i who were drawn at random.[7]

In imperfect information environments such as $\Gamma(p)$, beliefs play an important role. Let $\sigma_A^* := \mathbb{E}_B\left[\sigma_A \mid \Gamma(p)\right]$ denote player B's prior belief about σ_A, referred to as B's *first-order belief*; in words, σ_A^* captures B's *confidence* at the start of the game that A will choose In. In turn, let $\sigma_A^{**} :=$

[5] See also Camerer (2003), Cox (2004) and McCabe et al. (2003).

[6] Notice that given their noise structure, this intuition cannot be as straightforwardly captured by Cox et al. (2006) and Rand et al. (2013). Indeed, in their environment, A's action is either implemented or reversed, so A can still retain control over the outcome as p tends to 0 by simply flipping her choice.

[7] Thus, $\sigma_i(s_i) \in (0, 1)$ refers to the fraction of individuals in role i playing s_i, as well as to the objective probability with which s_i is played. In this respect, I follow Nash's mass action interpretation (Weibull 1996), which was also adopted in recent work (see for instance Attanasi et al. 2016).

$\mathbb{E}_A[\sigma_A^* \mid \Gamma(p)]$ be A's prior belief about σ_A^*, or A's *second-order belief*; that is, σ_A^{**} measures A's *confidence perception.*

2.2.2 Preferences

For simplicity and to minimize departures from the standard theory, I assume that player A is a standard expected utility maximizer with selfish preferences:

$$u_A(\sigma) = m_A(\sigma)$$

where $m_i(\sigma)$ denotes i's material payoffs ($i \in \{\text{A,B}\}$) under strategy profile σ.[8] On the other hand, assume that the preferences of B take the following form:

$$u_B(\sigma) = m_B(\sigma) + [\alpha + 1_{\{In\}}\theta(p, \sigma_A^*)]m_A(\sigma)$$

where $\alpha \in [0,1]$, $\theta(.,.)$ is a weakly positive and continuous function discussed below, and $1_{\{In\}}$ is an indicator variable for whether In was realized. Here, B's decision between $Meet$ ($\sigma_B = 1$) and $Take$ ($\sigma_B = 0$) may depend on two types of prosocial concerns. First, B may care about A's material payoffs out of pure altruism; this is captured by the altruistic parameter $\alpha \geq 0$, which does not depend on A's action.[9] Second, upon observing In, B may care about A's payoffs to the extent that he believes In was A's intention to trust, as modeled by the reciprocity function $\theta(p, \sigma_A^*)$. As argued in Section 2.1, perceived trust should depend not only on B's belief that A choose In, σ_A^*, but also on the probability p with which A's action is implemented, since choosing In cannot be perceived as an act of trust when p is low. I therefore assume that the reciprocity function $\theta(p, \sigma_A^*)$ has the following properties:

$A1:$ $\theta(0, \sigma_A^*) = \theta(p, 0) = 0$ for all $p < 1$

$A2:$ $(\frac{\partial \theta}{\partial \sigma_A^*}, \frac{\partial \theta}{\partial p}) \gneqq 0$

$A3:$ $\theta(1, \sigma_A^*) = \theta \leq 1$

In words, a reciprocal B player only cares about A's payoffs if he believes that A chose In with some positive probability and if A's decision has a positive chance of being implemented ($A1$). In

[8] This assumption about the first mover is fairly common in the literature on belief-dependent motivations; see for instance, Geanakoplos et al. (1989), Dufwenberg (2002), Charness and Dufwenberg (2006), Battigalli and Dufwenberg (2009) or Tadelis (2011). The Online Appendix of a previous version of this paper (available on the website on the author) contains an extension where A is allowed to exhibit social preferences; the main message is preserved.

[9] I assume that B puts a weakly lower weight on A's payoffs than on his own payoffs ($\alpha \leq 1$) to be consistent with the experimental literature; however, this restriction has no consequences in the discussion below.

those cases, B's sensitivity to A's payoffs is increasing in his prior belief that A chose In and in A's level of agency in determining the outcome ($A2$). Finally, when A bears the full consequences of his decision ($p = 1$) and In is realized, B must interpret the outcome as an act of trust from A, regardless of his initial beliefs ($A3$); the restriction $\theta \leq 1$ ensures that B puts (weakly) less weight on A's payoffs than on his own payoffs. It is worth noting that the above properties are compatible with many specifications of the reciprocity function, which captures B's concern for A's intentions in a reduced-form manner as in Charness and Rabin (2002).[10]

2.2.3 Optimal strategies as a function of p

In the following, I derive comparative statics with respect to p, depending on the value of the preference parameters (α, θ). As in Charness and Dufwenberg (2006), I analyze the game without making equilibrium assumptions, which are particularly restrictive in one-shot interactions with no learning. Under the above specification of preferences, B will choose $Meet$ over $Take$ when In is realized if and only if

$$u_B(\sigma_A, 1) \geq u_B(\sigma_A, 0)$$

$$\Leftrightarrow \; 10 + 10[\alpha + \theta(p, \sigma_A^*)] \geq 14 + 2[\alpha + \theta(p, \sigma_A^*)]$$

$$\Leftrightarrow \; \alpha + \theta(p, \sigma_A^*) \geq \frac{1}{2} \; (*)$$

There are 3 cases depending on the value of α and θ:

Selfish case: $\boldsymbol{\alpha + \theta < \frac{1}{2}}$. In this case, condition $(*)$ is never satisfied and $\sigma_B(p) = 0$ for all p, that is, B chooses $Take$ irrespective of the amount of noise. Therefore, A's optimal response is to choose Out, i.e. $\sigma_A(p) = 0$ for all p.

Pure altruism case: $\boldsymbol{\alpha \geq \frac{1}{2}}$. Here, $(*)$ is always satisfied, regardless of p and σ_A^*. Thus $\sigma_B(p) = 1$ for all p, that is, B chooses $Meet$ irrespective of the amount of noise. As a response, A always chooses In, i.e. $\sigma_A(p) = 1$ for all p.

Intentions matter: $\boldsymbol{\alpha + \theta \geq \frac{1}{2}}$ and $\boldsymbol{\alpha < \frac{1}{2}}$. In this case, B chooses $Meet$ if and only if $\theta(p, \sigma_A^*) \geq \frac{1}{2} - \alpha$. Notice that by assumptions $A1$ and $A3$, B chooses $Take$ if $p = 0$ and $Meet$ if $p = 1$. Let $\bar{\sigma}(p)$ be such that $\theta(p, \bar{\sigma}(p)) = \frac{1}{2} - \alpha$. Then B chooses $Meet$ if and only if $\sigma_A^* \geq \bar{\sigma}(p)$ where

[10] For instance, a previous version of this paper assumed that $\theta(p, \sigma_A^*) = \theta\mu(p, \sigma_A^*)$ where $\mu(p, \sigma_A^*) := \frac{p\sigma_A^*}{p\sigma_A^* + \frac{1}{2}(1-p)}$ is B's posterior belief after observing In that the outcome comes from A (assume $\mu = 1$ if $p = 1$ and $\sigma_A^* = 0$). In this case, $\frac{\partial \theta}{\partial \sigma_A^*} > 0$ for all $p \in (0, 1)$, and $\frac{\partial \theta}{\partial p} > 0$ for all $\sigma_A^* > 0$. These assumptions are also compatible with the predictions made by existing psychological game theories of intention-based reciprocity (Rabin 1993, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher 2006); for instance, it can be shown that optimal strategies are monotone increasing in p in the framework of Dufwenberg and Kirchsteiger (2004).

$\frac{\partial \bar{\sigma}}{\partial p} = -\frac{\partial \theta / \partial p}{\partial \theta / \partial \bar{\sigma}} \leq 0$ (by the implicit function theorem and assumption $A2$). In turn, A chooses In if and only if $\sigma_A^{**} \geq \bar{\sigma}(p)$. Therefore, $\sigma_B(p)$ and $\sigma_A(p)$ are increasing in p.

In summary, rates of prosociality should be monotone increasing in p if players believe that trust intentions matter in the noisy trust game; on the other hand, behavior will be irresponsive to p for selfish and purely altruistic agents.

3. Experimental design and procedures

The experiment was conducted at the Center for Experimental Social Science (CESS) of New York University with a regular student subject pool. Sessions lasted about 45 minutes. At the start of the session, subjects were randomly assigned a role, either A or B, and matched with one person in the other role for the entire session. All interactions took place through computer terminals; the experiment was programmed and conducted with the software z-Tree (Fischbacher 2007). To establish common knowledge of the rules of the game, paper instructions were distributed to all subjects and read collectively. Each session was divided in two parts corresponding to the elicitation of subjects' strategies (Part 1) and the elicitation of their beliefs about their matched partner (Part 2). Instructions for the second part were distributed only after completion of Part 1 and subjects received no feedback between the two parts. At the end of the session, subjects answered a short questionnaire in order to assess their understanding of the experiment (see OA-C.4). In addition to their incentive payments for Parts 1 & 2, subjects received a $5 show-up fee.

The key experimental variable was the probability p that A's action would be implemented. In Part 1, subjects in role A (B) made a series of choices between In and Out ($Meet$ and $Take$) for 11 values of p from 0 to 1 in increments of 0.1, in order to elicit $\sigma_A(p)$ and $\sigma_B(p)$.[11] The probability values were presented in ascending order and framed as a percentage chance. The boundary cases of $p = 0$ (dictator game) and $p = 1$ (standard trust game) were included in order to facilitate comparisons with the existing literature. The choices of B were elicited using the strategy method (Selten 1967); that is, B made choices without knowing whether In was realized and was asked to behave assuming this was the case.[12] Finally, subjects were told that their decisions would be implemented for one randomly selected value of p.

In Part 2, B was asked to guess the likelihood that A chose In for each value of p, corresponding to B's first-order beliefs $\sigma_A^*(p)$. In turn, A was asked to guess B's answer for each p, corresponding

[11] The instructions used more neutral names for B's actions, replacing $Take$ with Up and $Meet$ with $Down$.

[12] This method is widely used in experimental economics, for it allows to elicit the complete strategy profile of a given player. Although its effects are not fully understood, the strategy method often appears to trigger less emotional responses than the direct response method (Brandts and Charness 2011, Casari and Cason 2009). Thus, if anything, one would a priori expect a downward bias on the effect of perceived intentions in the noisy trust game.

to A's second-order beliefs $\sigma_A^{**}(p)$. Subjects were paid according to their guess for the randomly selected value of p. B's first-order beliefs were incentivized using a method similar in spirit to the Becker-Degroot-Marschak (1964) mechanism, which provides a dominant strategy to reveal correct beliefs independently of the subject's risk attitudes.[13] More precisely, B was asked to assess the percentage chance that A chose In by choosing a number x among the set of options {0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}. Secondly, an integer n between 0 and 100 was drawn at random. If $x \geq n$, then B received 5 dollars if A chose In and 0 otherwise; if $x < n$, B received 5 dollars with a n % chance (and 0 otherwise). Subjects were given an example explaining why it was in their best interest to report their true beliefs. To minimize the cognitive demand on the A players who were asked for their second-order beliefs, the latter were incentivized by offering 5 dollars for correctly guessing B's answer (out of the 11 possible options) and 0 otherwise.

Besides this baseline treatment (called *No Information* in a previous version), I ran two additional treatments, which are not analyzed in this paper: *Exogenous Information* (EI) where the Receiver is exogenously informed of the Sender's choice; *Costly Communication* (COM) where the Sender can pay a cost to inform the Receiver of his choice. In both treatments, Receivers made decisions under the strategy method for each value of p and possible choice of the Sender. In COM, in addition to the baseline beliefs, B guessed A's likelihood of paying the signaling cost if he chose In (Out) and A guessed B's answers. In those two treatments, the use of the strategy method appears more artificial since B could not observe A's decisions; furthermore, subjects had to take a large number of decisions (actions and beliefs). As a consequence, the findings for EI and COM are harder to interpret. A summary of the main findings can be found in Section B of the Online Appendix as well as in a previous version of this paper, both available on the website of the author.

4. Results

This experiment was conducted with a total of 76 subjects (38 pairs) spread over 5 sessions; for each role, the dataset therefore contains 418 ($= 11 \times 38$) binary choices of an action and 418 belief guesses in $\{0, 0.1, ..., 1\}$ made by a given subject for a specific value of p. Except if specified otherwise, the statistical results of this section are based on OLS regressions with standard errors clustered at the subject level (or at the level of a match when appropriate) to account for within-subject correlation across decisions; results are shown both with and without observations for $p = 0$, since A's decisions were inconsequential in this case. Significance is assessed with one-sided t-tests when there is a directional hypothesis and two-sided t-tests otherwise. Section 4.1 studies the aggregate response

[13] For a formal analysis of this procedure, see Schlag and van der Weele (2013).

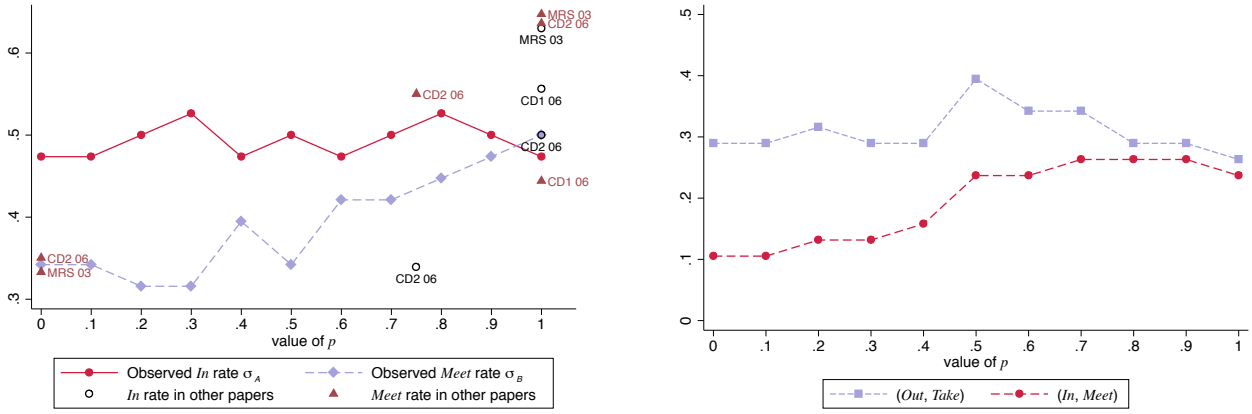
of behavior and beliefs to the value of p, while Section 4.2 studies individual heterogeneity.

4.1 Aggregate behavior and beliefs

4.1.1 Rates of prosocial behavior as a function of p

Figure 2: Aggregate behavior as a function of p

(a) Rate of prosociality by p (b) Proportion of No Trust and Trust profiles by p



Notes: In Panel (a), MRS 03 refers to McCabe et al. (2003), CD1 06 refers to Charness and Dufwenberg (2006) and CD2 06 refers to Cox and Deck (2006).

Figure 2 Panel (a) shows the fraction of A (resp. B) players who chose In (resp. $Meet$) for each value of p; as a benchmark, I also report the rates of prosociality observed in the 3 studies most connected to this paper, McCabe et al. (2003, MRS 03), Charness and Dufwenberg (2006, CD1 06) and Cox and Deck (2006, CD2 06), for the values of p that were tested in each paper.[14] Combining the behavior of A and B, Panel (b) shows the proportion of realized Trust and No Trust profiles, $(In, Meet)$ and $(Out, Take)$, for each value of p. If intentions matter, then players should become more prosocial as the value of p increases; as a result, the Trust profile $(In, Meet)$ should be more likely to emerge at higher values of p. Statistical tests of the relationship between action choices (or profiles) and the value of p are reported in Table A1 at the end of this paper.

Despite being embedded in a more complex setting, behavior at the extreme values of p is very much in line with previous findings, with similar rates of prosociality for both A and B.[15]

[14] This includes $p \in \{0, 1\}$ for B and $p = 1$ for A in McCabe et al. (2003); $p = 1$ for A and B in Charness and Dufwenberg (2006); $p \in \{0, 0.75, 1\}$ for B and $p \in \{0.75, 1\}$ for A in Cox and Deck (2006). To facilitate comparisons, a summary table of the relevant experimental features and findings of each study can be found in OA-A.4.

[15] The proportion of B players who share when $p = 0$ (dictator game) is almost identical across studies; for $p = 1$, it is very close to CD1 06 (the game closest in terms of payoff structure) and somewhat smaller to the other two papers (although not significantly so). Relative to the other studies, A's level of prosociality when $p = 1$ is slightly lower, but only significantly so relative to McCabe et al. (2003) (p-value = 0.065); unlike Cox and Deck (2006), the In rate does not drop for p in the range 0.7-0.8.

Importantly, B's degree of prosociality rises with p, with an increase in the $Meet$ rate from 34% at $p = 0$ to 50% at $p = 1$ (p-value = 0.057, one-sided t-test). Although the positive linear trend misses statistical significance ($\beta = 0.177$, p-value = 0.132 for the full sample), B's prosociality is somewhat sensitive to whether p takes a low, medium or high value, where $p_L \in \{0, 0.1, 0.2, 0.3\}$, $p_M \in \{0.4, 0.5, 0.6, 0.7\}$ and $p_H \in \{0.8, 0.9, 1\}$.[16] Given the observed $Meet$ rate $\sigma_B(p)$, the best response of a risk neutral A player is almost perfectly monotone increasing in p.[17] However, A's behavior appears to be insensitive to the amount of noise, with rates of prosociality around 50% across all values of p. As will be shown in the next section, this flat pattern in the aggregate hides substantial heterogeneity at the individual level. Putting together the behavior of A and B, the frequency of $(In, Meet)$ profiles more than doubles as p increases from 0 to 1 (10.5% vs 23.7%, p-value = 0.066 on a one-sided t-test). The positive linear trend is significant ($\beta = 0.184$, p-value = 0.057 for the full sample), although most of the increase seems to occur around $p = 0.5$. On the other hand, the frequency of $(Out, Take)$ is relatively stable across values of p at around 30%; it is significantly higher than the proportion of $(In, Meet)$ for $p \leq 0.3$ (p_L category) and almost identical to it for $p \geq 0.8$ (p_H category).

4.1.2 Distribution of beliefs as a function of p

Another way to study the effect of p on individual decisions is to look at subjects' belief patterns. If intentions matter and players form consistent beliefs, then B's first-order belief σ_A^* and A's second-order belief σ_A^{**} should be increasing in p. Figure 3(a) contrasts the mean beliefs of A and B with the observed In rate σ_A at each value of p, while Figure 3(b) shows kernel density estimates of the distribution of beliefs for low, medium and high values of p (see OA-A.1 for a complete breakdown by value of p). Table A2 at the end of the paper presents statistical tests of the effect of p on mean beliefs (Panel A) and median beliefs (Panel B).[18]

Mean beliefs exhibit a modest positive trend. On average, B players expect A to be 10 percentage points more likely to choose In when $p = 1$ relative to when $p = 0$ (61.8% vs 51.8%, p-value = 0.113, one-sided t-test). In turn, the average guess of A players about B's belief increases by 13 percentage points from $p = 0$ to $p = 1$ (44.4% vs 57.6%, p-value = 0.097, one-sided t-test.) Although the positive linear trend is insignificant in most specifications, the average guess of A and

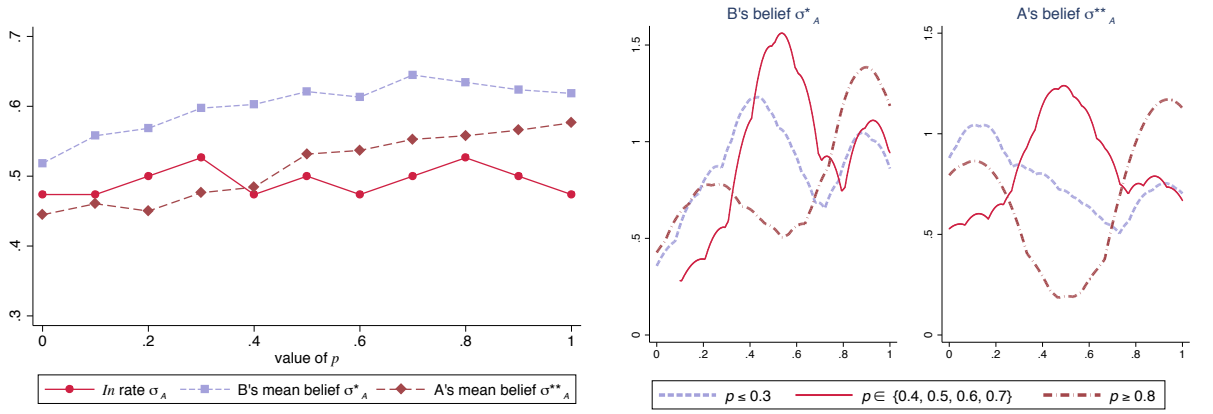
[16] A sensitivity analysis to different cutoff values of p_L, p_M and p_H is presented in OA-A.3; results are qualitatively similar, although somewhat weaker, with alternative breaks of the data.

[17] It is easy to verify that A's optimal strategy under risk neutrality is to choose In provided that $\sigma_B \geq 0.375$; given the observed $Meet$ rate $\sigma_B(p)$, A's best response is $\sigma_A(p) = 0$ for all $p \leq 0.5$ except 0.4 and $\sigma_A(p) = 1$ otherwise. Allowing for small levels of risk aversion (for instance, by assuming $u(x) = \frac{x^{1-\alpha}}{1-\alpha}$ with $\alpha \geq 0.2$) guarantees that σ_A is fully monotone increasing in p. See OA-A.5 for more details

[18] The findings discussed here are robust to alternative definitions of the cutoff values for p_L, p_M and p_H; see sensitivity analysis in OA-A.3.

Figure 3: Relationship between beliefs and value of p

(a) Mean belief of A and B by value of p (b) Distribution of beliefs for A and B for p_L, p_M and p_H



B is somewhat sensitive to whether p takes a low or a medium value. While the mean beliefs of A and B present a common trend, they differ in levels. The B players on average overestimate the In rate (significantly so for all $p \geq 0.3$), with a mean deviation of 10.7 percentage points across all values of p. The average guess of the A players closely follows the observed In rate, with a mean absolute deviation of 4.5 percentage points across all values of p. As a result, A underestimates B's guess (but only significantly so for $p \in \{0.3, 0.4\}$), with a mean deviation of 8.8 percentage points. While the effect of p on mean beliefs is relatively modest, it is more apparent for other measures of central tendency; in particular, the modal and median beliefs increase in an almost perfectly monotone fashion for A and B. Figure 3(b) shows a progressive shift of the mode from about 0.4 for B (0.2 for A) when p is low, to about 0.9 (for both A and B) when p is high. Similarly, quantile regressions show a strong positive effect of p on the median belief of A and B, which increases from 0.4 for A (0.5 for B) at low values of p to 0.8 (for both A and B) at high values of p.

Conclusion 1*: At the aggregate level, a weakening of A's level of agency as measured by a lower value of p has a negative effect on prosocial behavior and beliefs. The introduction of noise leads to fewer realizations of the Trust profile and more pessimistic beliefs. However, behavior appears somewhat less responsive to p than beliefs; in particular, A's behavioral response is essentially flat.*

4.2 Individual heterogeneity

The study of aggregate behavior only gives a partial view of players' sensitivity to changes in the amount of noise. For instance, if two groups of subjects of the same size respond to p in opposite manner, their aggregate behavior will appear to be irresponsive to p. In this section, I take a closer

look at individual strategies and beliefs in order to further investigate the extent to which perceived intentions may affect prosociality.

4.2.1 A typology of behavioral types

Subjects can be broadly classified into 5 behavioral types depending on whether and how their chosen action changes with p. The first two types correspond to subjects whose behavior is insensitive to p, meaning that they either always play the selfish action $Out/Take$, or always play the prosocial action $In/Meet$. The third and fourth types refer to subjects who switch their action exactly once as p increases, either from the selfish to the prosocial action (monotone increasing type) or from the prosocial to the selfish action (monotone decreasing type). The last type refers to the non monotone subjects who switch action multiple times.[19]

While the first 3 types can be rationalized by the theory presented in Section 2.2.3 (selfish play, pure altruism, case where intentions matter), the monotone decreasing and non monotone types cannot. As discussed below, non monotone behavior can be largely explained by random play. For the monotone decreasing types, an examination of subjects' answers to the exit questionnaire sheds some light on their motivation (see OA-C4 & -E). The prosociality of these subjects appears to be inversely related to B's chances of determining the final outcome, which happens only if In is realized. In other words, B is more likely to be prosocial when the cost of prosociality is low i.e. when his decision is unlikely to be consequential. Anticipating this, A is then less likely to choose In when p is high.

One way of capturing this idea is to assume that B's utility of acting prosocially has a "warm glow" component (Andreoni 1989, 1990): besides his material payoffs, B cares about *appearing* altruistic, whether his decision matters ex post or not (impure altruism). Formally, assume that B's utility function is given by $u_B(\sigma) = \mathbb{E}_B[m_B(\sigma)] + \Phi(\sigma_B)$ where $\Phi(1) = \phi > 0$ and $\Phi(0) = 0$, meaning that B gets warm glow utility ϕ from choosing $Meet$ and 0 otherwise. Under the strategy method, B must select an action without knowing whether In was realized and form expectations about his material payoffs. Letting $q(p, \sigma_A^*) := p\sigma_A^* + \frac{1}{2}(1-p)$ denote B's belief that In was realized, B will therefore choose $Meet$ over $Take$ provided that

$$5(1 - q(p, \sigma_A^*)) + 10q(p, \sigma_A^*) + \phi \;\geq\; 5(1 - q(p, \sigma_A^*)) + 14q(p, \sigma_A^*)$$

$$\Leftrightarrow \;\; q(p, \sigma_A^*) \;\leq\; \frac{\phi}{4}$$

[19] In order to classify subjects, I discard A's choice for $p = 0$, as decisions were inconsequential in this case. For instance, a subject who played In when $p = 0$ and Out afterwards is categorized as somebody who always played Out. Accounting for decisions at $p = 0$ would only slightly change the classification of A players (+1 for monotone increasing, +1 for monotone decreasing, -1 for always Out, -1 for always In).

Under some restrictions on ϕ, it can be shown that $\sigma_B(p)$ is decreasing in p and therefore A's best response $\sigma_A(p)$ is also decreasing in p; see Appendix for more details.

4.2.2 Individual heterogeneity in behavior and beliefs

Table 1: Distribution of individual strategies

Behavioral pattern for $j \in \{A, B\}$	Parameter conditions (model of Section 2.2)	% of A players (freq)	% of B players (freq)	Total % (freq)
$\sigma_j = 0$ for all p	$\alpha + \theta < \frac{1}{2}$ (selfish case)	26.3 (10/38)	39.5 (15/38)	32.9 (25/76)
$\sigma_j = 1$ for all p	$\alpha \geq \frac{1}{2}$ (pure altruism)	23.7 (9/38)	18.4 (7/38)	21.1 (16/76)
$\frac{\partial \sigma_j}{\partial p} \geq 0$ (> 0 for some p)	$\alpha < \frac{1}{2}$ and $\alpha + \theta \geq \frac{1}{2}$ (intentions matter)	21.1 (8/38)	23.7 (9/38)	22.4 (17/76)
$\frac{\partial \sigma_j}{\partial p} \leq 0$ (< 0 for some p)	$\varnothing$ [impure altruism]	21.1 (8/38)	10.5 (4/38)	15.8 (12/76)
other (non monotone)	$\varnothing$ [random play]	7.9 (3/38)	7.9 (3/38)	7.9 (6/76)

Table 1 presents the distribution of behavioral types described in the previous section. There are 4 main findings. First, over 75% of the observed behavioral patterns can be rationalized by the theory discussed in Section 2.2.3; in particular, very few subjects switch their action more than once.[20] Second, the behavior of a large fraction of subjects is irresponsive to changes in the environment. As much as 50% of A players and 58% of B players select the same action for all values of p. While an even fraction of A players select the prosocial and the selfish actions, the proportion of selfish B players is twice as large as the proportion of pure altruists. Third, among subjects who react to the environment, nearly half (17/35) behave in a way consistent with intention-based reciprocity. For the B players, those subjects represent the second largest category after selfish types (23.7% of the sample); their proportion is twice as large as the proportion of subjects who switch from the prosocial to the selfish action at higher values of p. However, for the A players, the proportion of subjects who play a monotone increasing strategy (consistent with intentions mattering) is counterbalanced by an equal proportion of subjects who play a monotone decreasing strategy. Therefore, the flat response observed at the aggregate level for the A players hides substantial heterogeneity at the

[20] Among the 6 subjects who do so, 5 exhibit a fairly erratic behavior, consistent with random play. The last subject exhibits a behavior close to monotone increasing (case where intentions matter), which might have been what this subject intended to play. See OA-A.2 for more details.

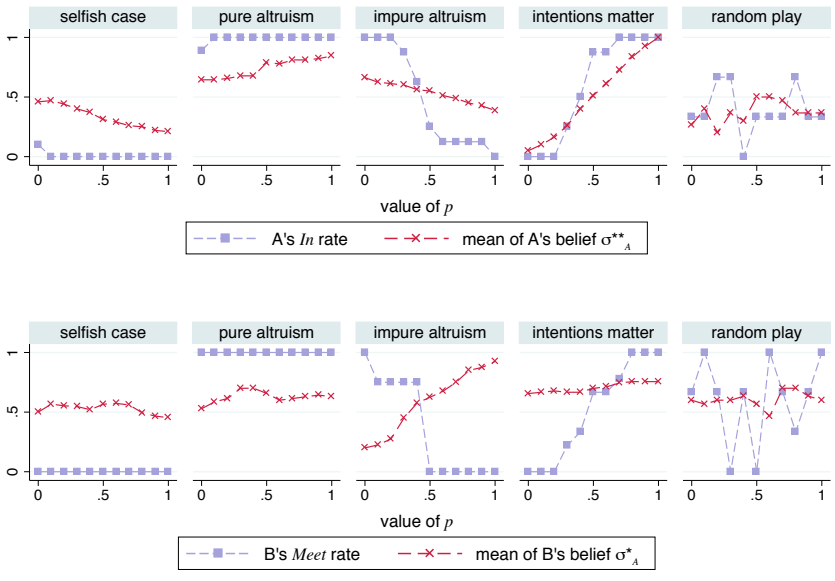
individual level in the way subjects react to changes in the value of p.

A similar level of individual heterogeneity can be observed for belief patterns (see OA Table 1). Relative to behavior, beliefs appear to be more responsive to changes in the value of p: while over 50% of subjects pick the same action irrespective of p, only 26% report the same beliefs regardless of the amount of noise. The largest category corresponds to subjects whose beliefs (σ_A^* for B and σ_A^{**} for A) are monotone increasing in p, consistent with the interpretation that trust intentions matter. This is the modal category for both A and B (34% of the sample for both roles). As previously observed for behavior, the fraction of B players with monotone increasing beliefs is about twice as large as those with monotone decreasing beliefs (13/38 vs 7/38); on the other hand, there is an almost equal proportion of A players in each category (13/38 vs 12/38). Subjects with non monotone beliefs represent about 15% of the sample.

4.2.3 Relationship between individual behavior and beliefs

A final step to understand subjects' response to the noise is to bring together behavior and beliefs. For each of the 5 behavioral types of Table 1, Figure 4 presents the mean belief and rate of prosociality at each value of p (top panel for A and bottom panel for B). A full breakdown for each subject can be found in OA-A.2.

Figure 4: Rates of prosociality and mean beliefs by behavioral type



Beliefs vary greatly across behavioral types. For the A players, beliefs follow behavior fairly closely. Among subjects whose propensity to choose In is monotone increasing in p (intentions matter case), the mean belief increases sharply and is remarkably close to the actual In rate for most values of p. For subjects whose propensity to be prosocial is decreasing in p (impure altruism case), the mean belief also follows a monotone decreasing pattern, although beliefs decrease less sharply than the In rate. Since the A players were asked for their second-order beliefs, the observed similarity between behavior and beliefs could partly reflect subjects' general tendency to believe that others can predict their behavior well (Vanberg 2008, Ellingsen et al. 2010, Butler et al. 2015).

The beliefs of the B players convey additional information about the nature of their preferences. First, among those whose behavior is insensitive to p, beliefs appear to be orthogonal to behavior; in particular, both groups have very similar beliefs on average and those beliefs are far away from their respective rates of prosociality. This finding is consistent with the interpretation of observed behavior as reflecting selfish play and pure altruism, since beliefs about A's prosociality should not influence the propensity to be prosocial of those two types. Consistent with a warm glow interpretation, subjects with a monotone decreasing behavior are less likely to choose $Meet$ as their confidence that A chose In increases, that is, as their decision becomes more likely to count. Finally, the mean belief of subjects whose behavior is consistent with intention-based reciprocity is above 0.5 for all values of p but follows only a modest increase (from 0.66 at $p = 0$ to 0.76 at $p = 1$). Beliefs are ascending for only about half of these subjects (4/9), and constant or descending for the other half (see OA for the breakdown). Coming back to Assumptions $A1$-$A3$ in Subsection 2.2.2, these findings suggest that the effect of p on B's reciprocity cannot be solely explained by more optimistic beliefs about A's propensity to go In (indirect channel), but also by the amount of risk contained in the decision to go In (direct channel).

Looking more closely at subjects with monotone increasing strategies (intentions matter case), one can ask how much noise is perceived as too much noise for prosocial behavior to occur. First, both A and B stop acting prosocially ($\sigma_A(p) = \sigma_B(p) = 0$) once $p < 0.3$; on the other hand, the rate of prosociality reaches 1 once $p \geq 0.7$ for A and once $p \geq 0.8$ for B. In between, most of the decline in prosocial behavior occurs over the interval [0.2,0.5], where a decrease in p by 0.1 leads to a drop in the In (resp. $Meet$) rate by an average of 29.1 (22.2) percentage points. A similar decrease in p by 0.1 over (0.5,0.7] (resp. (0.5, 0.8]) has a more modest effect on rates of prosociality, with a decrease by an average of 6.25 percentage points for A (resp. 11.1 percentage points for B). In other words, prosociality tends to prevail as long as A's action is most likely to be implemented ($p > 0.5$), but becomes seriously threatened when A stops to be the main first-stage actor.

Conclusion 2*: At the individual level, there is substantial heterogeneity in how subjects respond to noisy environments. While intentions seem to matter for 20-25% of subjects, behavior is irresponsive to p for about half of the sample. In addition, social image concerns appear almost as equally important as trust intentions to explain aggregate behavior. Among those for whom intentions seem to matter, prosociality falls sharply once A's decision is implemented with a 50% chance or less.*

6. Conclusion

Many experimental studies of social dilemma games find that perceived intentions affect one's propensity to be prosocial, in particular when it comes to trust relationships. If intentions matter, then the ability to convey trust intentions in a credible manner should influence the onset of trust relationships. The present paper tests this conjecture in a noisy binary trust game where actions may only imperfectly reflect trust intentions. I manipulate the strength of the trust signal by varying the probability with which the Sender's action is implemented. Unlike most of the previous literature, the experimental design exploits within-subject variation to identify the effect of intentions on individual behavior and classify subjects into types depending on how they respond to the noise. When actions are noisy signals of trust intentions, how much noise becomes too much noise for the trust-reciprocity outcome to emerge?

I find that prosocial behavior and beliefs are affected by the credibility with which Senders can signal their intentions to trust. The trust-reciprocity outcome is less likely to emerge as the value of p decreases i.e. when the Sender has less control over the outcome; furthermore, beliefs become increasingly more pessimistic. This paper therefore provides evidence that prosociality is sensitive to the credibility with which Senders can convey trust intentions through their actions. One implication of this finding is that trust relationships could be efficiently promoted by policies designed to increase transparency of the decision-making environment. For instance, in collaborations involving multiple actors and decision-making stages, procedures that make explicit the actual contribution of each actor could foster trust by boosting one's confidence that individual efforts will be acknowledged and reciprocated.

Although intentions matter, they do not matter for everyone. The within-subject analysis of this paper reveals that the effect of intentions on individual strategies is highly heterogeneous. The behavior of half of the Senders and more than half of the Receivers is insensitive to whether Senders have control over the final outcome. Furthermore, if perceived intentions seem to be a key determinant of prosociality, the individual analysis suggests that social image concerns might be almost as equally important to explain the behavioral patterns observed in the aggregate data.

Because a large heterogeneity in individual responses can limit the efficacy of a given policy, more research is needed to understand the distribution of social preferences in the population.

References

[1] Andreoni, J. (1989), "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy*, vol. 97, 1447-1458.

[2] Andreoni, J. (1990), "Impure altruism and donations to public goods: A theory of warm-glow giving", *Economic Journal*, vol. 100, 464-477.

[3] Arrow, K. (1972), "Gifts and Exchanges", *Philosophy and Public Affairs*, I, 343-362.

[4] Attanasi, G., P. Battigalli, and E. Manzoni (2016), "Incomplete Information Models of Guilt Aversion in the Trust Game", *Management Science*, vol. 62 (3), 648-667.

[5] Attanasi, G., P. Battigalli, and R. Nagel (2015), "Disclosure of Belief-Dependent Preferences in a Trust Game", working paper.

[6] Battigalli, P. and M. Dufwenberg (2009), "Dynamic Psychological Games", *Journal of Economic Theory*, vol. 144, 1-35.

[7] Berg, J., J. Dickhaut and K.A. McCabe (1995), "Trust, Reciprocity, and Social History", *Games and Economic Behavior*, vol. 10, 290-307.

[8] Blount, S. (1995), "When social outcomes aren't fair: the effect of causal attributions on preferences", *Organizational Behavior and Human Decision Processes*, vol. 63, 131-144.

[9] Brandts, J. and G. Charness (2011), "The strategy versus the direct-response method: a first survey of experimental comparisons", *Experimental Economics*, vol. 14, 375-398.

[10] Brandts, J. and C. Solà (2000), "Reference points and negative reciprocity in simple sequential games", *Games and Economic Behavior*, vol. 2, 227-238.

[11] Butler, J., P. Giuliano and L. Guiso (2015), "Trust, Values and False Consensus", *International Economic Review*, vol. 56 (3), 889-915.

[12] Camerer, C.F. (2003), *Behavioral Game Theory*, Princeton University Press.

[13] Casari, M. and T. Cason (2009), "The Strategy Method Lowers Measured Trustworthy Behavior", *Economics Letters,* vol. 103, 157–159

[14] Charness, G. (2004), "Attribution and reciprocity in a simulated labor market: an experimental investigation.", *Journal of Labour Economics*, vol. 22, 665-688.

[15] Charness, G. and M. Dufwenberg (2006), "Promises and Partnership", *Econometrica*, vol. 74 (6), 1579-1601.

[16] Charness, G. and D. Levine (2007), "Intention and Stochastic Outcomes", *The Economic Journal*, vol. 117, 1051-1072.

[17] Charness, G. and M. Rabin (2002), "Understanding Social Preferences with Simple Tests", *Quarterly Journal of Economics*, vol. 117 (3), 817-869.

[18] Charness, G. and M. Rabin (2005), "Expressed preferences and behavior in experimental games", *Games and Economic Behavior*, vol. 53, 151–69.

[19] Cox, J. C. (2004), "How to identify trust and reciprocity", *Games and Economic Behavior*, vol. 46, 260-281.

[20] Cox, J. C. and C. A. Deck (2006), "Assigning Intentions When Actions Are Unobservable: The Impact of Trembling in the Trust Game", *Southern Economic Journal*, vol. 73 (2), 307-314.

[21] Cox, J. C., K. Sadiraj, and V. Sadiraj (2008), "Implications of trust, fear, and reciprocity for modelling economic behavior", *Experimental Economics*, vol. 11, 1-24.

[22] Dufwenberg, M. (2002), "Marital Investment, Time Consistency and Emotions", *Journal of Economic Behavior and Organization*, vol. 48, 57-69.

[23] Dufwenberg, M. and G. Kirchsteiger (2004), "A theory of sequential reciprocity", *Games and Economic Behavior*, vol. 47, 268-298.

[24] Ellingsen, T., M. Johannesson, G. Torsvik, and S. Tjøtta (2010), "Testing Guilt Aversion", *Games and Economic Behavior*, vol. 68 (1), 95-107.

[25] Falk, A. and U. Fischbacher (2006), "A theory of reciprocity", *Games and Economic Behavior*, vol. 54, 293–315.

[26] Falk, A., E. Fehr, and U. Fischbacher (2003), "On the Nature of Fair Behavior", *Economic Inquiry*, vol. 41 (1), 20-26.

[27] Falk, A., E. Fehr, and U. Fischbacher (2008), "Testing theories of fairness - Intentions matter", *Games and Economic Behavior*, vol. 62 (1) 287-303.

[28] Geanakoplos, J., D. Pearce, and E. Stacchetti (1989), "Psychological Games and Sequential Rationality", *Games and Economic Behavior*, vol. 1, 60-79.

[29] McCabe, K., S. J. Rassenti, and V. L. Smith (1998), "Reciprocity, trust, and payoff privacy in extensive form bargaining", *Games and Economic Behavior*, vol. 24, 10-24.

[30] McCabe, K., M. Rigdon, and V. L. Smith (2003), "Positive Reciprocity and Intentions in Trust Games", *Journal of Economic Behavior and Organization*, vol. 52, 267-275.

[31] Offerman, T. (2002), "Hurting Hurts More Than Helping Helps", *European Economic Review*, vol. 46, 1423-37.

[32] Putnam, R. (1995), "The Case of the Missing Social Capital", mimeographed.

[33] Rabin, M. (1993), "Incorporating fairness into game theory and economics", *American Economic Review*, vol. 83, 1281–1302.

[34] Rand, D. G., D. Fudenberg, and A. Dreber (2015), "It's the Thought That Counts: The Role of Intentions in Noisy Repeated Games", *Journal of Economic Behavior and Organization*, vol. 116, 481-499.

[35] Schlag, K. and J. van der Weele (2013), "Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality", *Theoretical Economics Letters*, 3:1, 38-42.

[36] Tadelis, S. (2011), "The Power of Shame and the Rationality of Trust", working paper.

[37] Vanberg, C. (2008), "Why do people keep their promises? An experimental test of two explanations", *Econometrica*, vol. 76, 1467-1480.

Appendix

Comparative statics for the impure altruism case

Following the discussion of Section 4.2.1, assume that B maximizes $u_B(\sigma) = \mathbb{E}_B[m_B(\sigma)] + \Phi(\sigma_B)$ where $\Phi(1) = \phi > 0$ and $\Phi(0) = 0$. Letting $q(p, \sigma_A^*) := p\sigma_A^* + \frac{1}{2}(1-p)$ denote B's belief that In is realized, B chooses $Meet$ over $Take$ provided that

$$5(1 - q(p, \sigma_A^*)) + 10q(p, \sigma_A^*) + \phi \;\geq\; 5(1 - q(p, \sigma_A^*)) + 14q(p, \sigma_A^*)$$

$$\Leftrightarrow \quad q(p, \sigma_A^*) \;\leq\; \frac{\phi}{4} \quad (**)$$

There are 3 cases to consider, depending on the value of ϕ:

Case 1: $\phi \geq 4$. In this case, $(**)$ is always satisfied, regardless of the value of p and σ_A^*. Thus $\sigma_B(p) = 1$ for all p, that is, B chooses $Meet$ irrespective of the amount of noise. As a response, A always chooses In, i.e. $\sigma_A(p) = 1$ for all p.

Case 2: $\phi \in (2, 4)$. Here, B chooses $Meet$ if and only if

$$q(p, \sigma_A^*) \leq \frac{\phi}{4} \;\Leftrightarrow\; \sigma_A^* \leq \bar{\sigma}(p, \phi) \;:=\; \frac{1}{2} + \frac{\phi - 2}{4p}$$

In this case, $\bar{\sigma}(p, \phi) \geq 0$ and $\frac{\partial \bar{\sigma}}{\partial p} \leq 0$; furthermore, $\bar{\sigma}(p, \phi) > 1 \Leftrightarrow p < \frac{\phi}{2} - 1$. Thus, for:

(i) $p \in [0, \frac{\phi}{2} - 1]$, B chooses $Meet$ for all σ_A^*. In turn, A chooses In regardless of σ_A^{**}.
(ii) $p \in (\frac{\phi}{2} - 1, 1]$, B chooses $Meet$ iff $\sigma_A^* \leq \bar{\sigma}(p, \phi)$ and thus A chooses In iff $\sigma_A^{**} \leq \bar{\sigma}(p, \phi)$.

In other words, when $\phi \in (2, 4)$, both $\sigma_A(p)$ and $\sigma_B(p)$ are (weakly) decreasing in p.

Case 3: $\phi \in [0, 2]$. In this case, it is easily verified that for:

(i) $p \in [0, 1 - \frac{\phi}{2})$, B chooses $Take$ for all σ_A^*. In turn, A chooses Out regardless of σ_A^{**}.
(ii) $p \in [1 - \frac{\phi}{2}, 1]$, B chooses $Meet$ iff $\sigma_A^* \leq \bar{\sigma}(p, \phi)$ and thus A chooses In iff $\sigma_A^{**} \leq \bar{\sigma}(p, \phi)$.

Since $\bar{\sigma}(p, \phi)$ is now (weakly) increasing in p, both $\sigma_A(p)$ and $\sigma_B(p)$ are weakly increasing in p. However, since $\bar{\sigma}(p, \phi)$ is bounded above by $\frac{1}{2}$ when $\phi \leq 2$, B will never choose $Meet$ for any $\sigma_A^* > \frac{1}{2}$. This prediction is largely inconsistent with the data presented in Section 4.2. In particular, 7 of the 9 subjects in role B whose behavior is consistent with intention-based reciprocity chose $Meet$ at values of p for which $\sigma_A^* > 0.5$, suggesting that the warm glow case $\phi \leq 2$ cannot explain their behavior.

Table A1: Effect of p on prosocial behavior

	Panel A: Individual behavior				**Panel B: Action profiles**			
	A's *In* Rate		B's *Meet* Rate		(*In*, *Meet*)		(*Out*, *Take*)	
Sample	Full	$p \neq 0$	Full	$p \neq 0$	Full	$p \neq 0$	Full	$p \neq 0$
(I) Test of linear trend								
Estimated effect of p	0.012	0.003	0.177	0.203	0.184*	0.187*	-0.005	-0.019
	[0.138]	[0.142]	[0.115]	[0.121]	[0.094]	[0.095]	[0.120]	[0.128]
Predicted effect of p	+		+		+		−	
(II) Rates of prosociality by noise category								
$p_L \leq 0.3$	0.493	0.5	0.329	0.325	0.118	0.123	0.296	0.298
$p_M \in [0.4, 0.7]$	0.487	0.487	0.395	0.395	0.224	0.224	0.342	0.342
$p_H \geq 0.8$	0.5	0.5	0.474	0.474	0.254	0.254	0.281	0.281
Predictions*			*t*-Stat values					
$\sigma_{i,p_L} \leq \sigma_{i,p_M}$	0.10	0.17	1.04	1.13	1.82**	1.80**	0.70	0.67
$\sigma_{i,p_L} \leq \sigma_{i,p_H}$	0.00	0.00	1.64*	1.71**	1.99**	1.99**	0.17	0.2
$\sigma_{i,p_M} \leq \sigma_{i,p_H}$	0.32	0.3	1.76**	1.76**	0.91	0.91	1.23	1.23
$^*\sigma_i \in \{\sigma_A, \sigma_B\}$								
Observations	418	380	418	380	418	380	418	380

Notes: Linear probability models with the dependent variable equal to 1 if A (B) chose In ($Meet$) for Panel A, and if the profile $(In, Meet)$ (resp. $(Out, Take)$) was realized for Panel B, regressed on p treated as a continuous variable in (I) and as a categorical variable in (II). Full ($p \neq 0$) sample includes (excludes) observations for $p = 0$. Standard errors in square brackets clustered at the subject level for Panel A and at the level of a match for Panel B. Significance in (II) assessed with one-sided t-tests; * and ** indicate p-value < 0.1 and < 0.05.

Table A2: Effect of p on beliefs

	Panel A: Mean belief						**Panel B: Median belief**					
	B's belief σ_A^*		A's belief σ_A^{**}		Combined		B's belief σ_A^*		A's belief σ_A^{**}		Combined	
Sample	Full	$p \neq 0$	Full	$p \neq 0$	Full	$p \neq 0$	Full	$p \neq 0$	Full	$p \neq 0$	Full	$p \neq 0$
(I) Test of linear trend												
Estimated effect of p	0.097	0.075	0.146	0.150	0.122*	0.112	0.250	0.337*	0.429***	0.667***	0.333**	0.429***
	[0.089]	[0.089]	[0.116]	[0.120]	[0.073]	[0.074]	[0.180]	[0.174]	[0.167]	[0.171]	[0.128]	[0.137]
Predicted effect of p	+		+		+		+		+		+	
(II) Rates of prosociality by noise category												
$p_L \leq 0.3$	0.561	0.575	0.458	0.462	0.509	0.518	0.5	0.5	0.4	0.4	0.5	0.5
$p_M \in [0.4, 0.7]$	0.620	0.620	0.526	0.526	0.573	0.573	0.6	0.6	0.5	0.5	0.6	0.6
$p_H \geq 0.8$	0.625	0.625	0.567	0.567	0.596	0.596	0.8	0.8	0.8	0.8	0.8	0.8
Predictions*	t-Stat values											
$\tilde{\sigma}_{A,p_L} \leq \tilde{\sigma}_{A,p_M}$	1.43*	1.24	1.39*	1.39*	2.0**	1.88**	1.69**	1.71**	0.8	1.25	2.65***	2.65***
$\tilde{\sigma}_{A,p_L} \leq \tilde{\sigma}_{A,p_H}$	0.98	0.81	1.23	1.23	1.58*	1.48*	4.19***	4.14***	2.2**	3.36***	6.18***	6.07***
$\tilde{\sigma}_{A,p_M} \leq \tilde{\sigma}_{A,p_H}$	0.14	0.14	0.98	0.98	0.88	0.88	6.17***	6.17***	3.68***	5.34***	8.49***	8.22***
*$\tilde{\sigma}_A \in \{\sigma_A^*, \sigma_A^{**}\}$												
Observations	418	380	418	380	836	760	418	380	418	380	836	760

Notes: Linear regressions where the dependent variable is B's elicited belief $\sigma_A^* \in \{0, 0.1, ..., 1\}$ that A chose In, A's elicited belief $\sigma_A^{**} \in \{0, 0.1, ..., 1\}$ about B's belief, or both (Combined), regressed on p treated as a continuous variable in (I) and as a categorical variable in (II). Panel A (B) shows OLS (quantile) regressions. Full ($p \neq 0$) sample includes (excludes) observations for $p = 0$. Standard errors in square brackets clustered at the subject level. Significance in (II) assessed with one-sided t-tests; *, ** and *** indicate p-value < 0.1, < 0.05 and < 0.01.