Intention-Based Reciprocity and Signalling of Intentions*

Séverine Toussaert†

February 19, 2016

Intentions count in your actions. Abu Bakr

Abstract

Many experiments find that trust intentions are a key determinant of prosociality, but ignore the uncertainty pertaining to informal trust agreements. If intentions matter, then trust should depend on whether intentions can be transparently conveyed. This conjecture is formalized and tested in a noisy trust game where I vary the extent to which trust can be credibly signalled. I find that (i) prosociality decreases when intentions become noisier; (ii) Subjects are willing to pay to signal their trust. Therefore, not only do intentions count, but players internalize this fact. However, the effect of intentions on individual behavior is highly heterogeneous.

JEL classification: C91, D63, D64

Keywords: Trust; Intentions; Reciprocity; Noise; Signalling; 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 Yuksel and Elliot Lipnowski for numerous discussions, as well as Andrew Schotter, Tore Ellingsen, Magnus Johannesson, Sebastian Fehrler, Gary Charness, Pierpaolo Battigalli, three 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.

†New York University, Department of Economics. Email: st1445@nyu.edu

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. One consequence of their implicit nature is that 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, trust relationships are characterized by the following three conditions: (i) the relationship is mutually beneficial, (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 of the mechanisms proposed to explain these findings is that Receivers like to reciprocate the risk taken by their Sender for it signals their 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 randomly determined 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. This concern, if verified, would make results from the lab quite uninformative about the world. It is therefore important to assess the extent to which

[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. (2013) 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).

the onset of trust might depend on the transparency with which intentions can be conveyed. If trust relationships are fragile in the presence of noise, then potential trustors might benefit from strengthening their signal of trust, even if this comes at a cost.

In an environment where actions may only imperfectly reflect trust intentions, this paper investigates how changes in the ability to signal trust may affect trust relationships. To study this question, I consider a noisy binary trust game where perceived intentions are manipulated through the interaction of two variables: (i) the probability p with which the Sender's action is implemented; (ii) the information transmitted to the Receiver about the Sender's action. In the baseline game, the Receiver faces the Sender's decision with some common knowledge probability p; otherwise, 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 a given amount of noise, I then compare prosocial behavior in two information treatments: one treatment where the Receiver is exogenously informed of his partner's decision; the other treatment where he possesses no information. Finally, I study whether Senders internalize the importance of signalling their good intentions by exploring a third treatment where they can communicate their action to the Receiver at some cost. By a revealed preference argument, if Senders pay to communicate their action, they must believe that stronger signals of trust will foster the reciprocity of their partner.

If intentions matter, then prosocial behavior should be affected by the interaction of (i) and (ii), that is, by how the amount of noise interacts with the Receiver's knowledge of the Sender's action. Intuitively, when p is low, whether the Receiver knows what action the Sender took is irrelevant: since the Sender's decision is unlikely to matter, the Receiver cannot attribute trusting intentions to the Sender. Hence no credible signal of trust can be sent. On the other hand, when p is high, the Receiver should be more likely to reciprocate if he can be ensured that the Sender chose the trusting action. 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 and her choice is known to the Receiver. As a result, costly communication can be worthwhile for the Sender provided she has enough control over the outcome. I formalize this intuition and test the model's predictions.

The present paper speaks to a burgeoning literature connecting prosociality to intentions in environments of imperfect information. Most closely connected to this study, two papers consider noisy binary games where the intended action is implemented with some probability p, and reversed otherwise.[3] In a binary trust game, Cox et al. (2006) find that the Receiver's degree of prosociality

[3] 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

does not significantly decrease when the Sender's action is only implemented with probability $p = \frac{3}{4}$ 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. (2013) 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 expands on these papers in several directions. First, both papers restrict attention to small departures from the perfect information case and do not analyze how the amount of noise $(1-p)$ may interact with the Receiver's information. Furthermore, these papers only exploit between-subject variation to identify the effect of noise on prosocial behavior; as a result, they cannot assess whether individual heterogeneity exists in how subjects respond to the noise. Finally, the present paper is, to my knowledge, the first one to study the impact of costly communication of intentions on prosocial behavior.[4]

In the rich information environment considered in this paper, I find that not only is prosocial behavior sensitive to whether trust intentions can be clearly conveyed, but Senders understand the importance of signalling credible intentions. First, players tend to move away from relationships of mistrust when the amount of noise is sufficiently small and the Sender's action is common knowledge. This result holds whether the Receiver's information is exogenous or determined by the Sender. Secondly, Senders make a strategic use of costly communication by choosing to pay only if it can credibly signal their trust. These findings therefore bring new evidence that not only do intentions count in one's actions, but players internalize this fact. At the same time, the analysis of individual strategies reveals that the effect of perceived intentions on prosocial behavior is highly heterogeneous in the population, with a large fraction of players being irresponsive to the environment. Furthermore, prosociality in this experiment also appears to be guided by social image concerns made salient through the use of the strategy method.

The remainder of the paper is organized as follows. Section 2 introduces the noisy trust game on which the experiment is based and analyzes the role of perceived intentions in a simple model. Section 3 describes the experimental design and derives comparative statics predictions. Section 4 presents the main results, while Section 5 discusses observed deviations from the theory. Section 6 concludes. Instructions and supplemental material can be found in an Online Appendix (henceforth OA) available on the author's website.

follows Nash equilibrium predictions for rational and selfish agents.

[4] Although some studies show that pre-play communication may foster trust relationships, the evidence has been so far limited to cheap talk communication from Receivers to Senders (see Charness and Dufwenberg 2006). Here I focus on the effect of truthful and costly communication from Senders to Receivers.

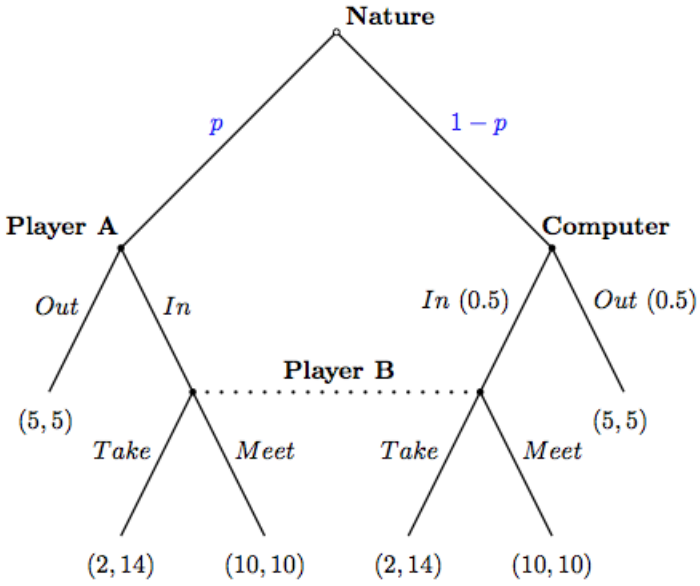
2. Intentions in a Noisy Trust Game: Theory and Predictions

In this section, I introduce the noisy trust game on which the experiment is based (Section 2.1), present a simple intention-based model to analyze the game and show how intentions to trust may depend on the environment (Section 2.2).

2.1 The Noisy Binary Trust Game

Consider the noisy trust game represented in Figure 1 by game tree $\Gamma_1(p, \Omega_A)$, where p parametrizes the amount of noise in the game and Ω_A refers to the information that player B (the Receiver) possesses about the action of player A (the Sender). A formal definition of Ω_A is introduced in Section 2.2. Payoffs are in dollars.

Figure 1: The game form $\Gamma_1(p, \Omega_A)$



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: player A can choose to invest in a relationship with player B or stay *Out* of the relationship; if A chooses *In*, B can choose to meet A's investment by sharing gains or decide to take most of the benefits for himself. At the other extreme where $p = 0$, 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 her 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? To wit, 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, her 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 feel more confident about choosing In as p increases. This intuition is formalized in the next section.

2.2 A Simple Intention-Based Model

The previous discussion suggested that B may care not only about realized outcomes but also about their origin: if he could be convinced that In originated from A's genuine intention to trust, B would be more likely to share. In this section, I produce a formal analysis of the noisy trust game, which captures this intuition. After introducing the key ingredients of the model, I derive comparative statics with respect to the amount of noise and analyze the set of equilibria.

2.2.1 Strategies, Information and Beliefs

Strategies

Let $\sigma_A \in \triangle(S_A)$ be player A's propensity to choose In in $\Gamma_1(p, \Omega_A)$ where $S_A := \{In, Out\}$.

[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.

Let $\sigma_B \in \triangle(S_B)$ be player B's propensity to choose $Meet$ in $\Gamma_1(p, \Omega_A)$ where $S_B := \{Meet, Take\}$. 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, 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 $\sigma_i \in (0, 1)$ will be interpreted as coming from a statistical distribution of pure strategies played by individuals in role i who were drawn at random.[7]

Two key elements of the model are the information and the beliefs held by B about A's action:

Information

An *information structure* is a map Ω_A summarizing the information that B possesses about A's action.[8] In the following, assume Ω_A is common knowledge. Two cases will be contrasted in the experiment: (1) *No information case*: $\Omega_A(s_A) = S_A$ for all $s_A \in S_A$; (2) *Perfect information case*: $\Omega_A(s_A) = \{s_A\}$ for all $s_A \in S_A$.

Beliefs

Let $\sigma_A^* := \mathbb{E}_B\left[\sigma_A \mid \Gamma_1(p, \Omega_A)\right]$ denote player B's belief about σ_A, that is, B's *first-order belief*: σ_A^* captures B's *confidence* in A's propensity to choose In.

Let $\sigma_A^{**} := \mathbb{E}_A\left[\sigma_A^* \mid \Gamma_1(p, \Omega_A)\right]$ denote player A's belief about σ_A^*, that is, A's *second-order belief*: σ_A^{**} measures A's *confidence perception*.[9]

When the environment is noisy (i.e. when $p \in (0, 1)$) and In is realized, one important belief denoted by $\mu_B(\sigma_A^*)$ measures B's confidence that In was A's intention:

$$\mu_B(\sigma_A^*) := \frac{p\sigma_A^*}{p\sigma_A^* + \frac{1}{2}(1-p)}$$

and $\mu_B(\sigma_A^{**})$ is defined similarly.

[7] Thus, $\sigma_i(s_i) \in (0, 1)$ will refer 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. 2014).

[8] Formally, an information structure is a map $\Omega_A : S_A \to 2^{S_A} \setminus \varnothing$ with the following two properties: (i) $s_A \in \Omega_A(s_A)$ and (ii) $s'_A \in \Omega_A(s_A) \Leftrightarrow s_A \in \Omega_A(s'_A)$.

[9] To be accurate, second-order beliefs in epistemic game theory are defined as probability measures over the product space of strategies and first-order beliefs. In this sense, σ_A^{**} is only a feature of A's second-order belief.

2.2.2 Preferences

As is fairly common in the literature on belief-dependent motivations, I assume for simplicity that only one player exhibits social preferences.[10] In particular, player A is assumed to be 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. In the Online Appendix (OA-A.1), I relax this assumption to allow player A to also exhibit social concerns. I show that although the set of equilibria is enlarged, there is a range of the parameter space in which the main intuition of the model remains intact.[11] The assumption of selfish preferences for A was therefore retained for its simplicity and to minimize departures from the standard theory.

To study the role of intentionality, I assume that the preferences of B take the following form:

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

Here B may care about A's material payoffs for two reasons. First, B may care about A for purely altruistic reasons, as captured by the altruistic parameter $\alpha \geq 0$. Secondly, B may care about A's payoffs to the extent that he believes In was A's intention to trust, where the sensitivity of B to A's intention is captured by the intention parameter $\theta \geq 0$. The pair of parameters (α, θ) is assumed to be common knowledge.[12]

It is worth relating the above specification of B's preferences to other specifications proposed in the literature to model intention-based reciprocity. As in Charness and Rabin (2002), this specification captures intention-based reciprocity in a reduced-form manner, that is, without explicitly introducing the formal apparatus developed in psychological game theories of intention-based reciprocity (Rabin 1993, Dufwenberg and Kirchsteiger 2004, Falk and Fischbacher 2006).[13] In these models, B's prosociality is affected by how kind he perceives A's action, which depends on what A expects

[10] See Geanakoplos et al. (1989), Dufwenberg (2002), Charness and Dufwenberg (2006), Battigalli and Dufwenberg (2009) or Tadelis (2011) for examples of games where only one player is assumed to exhibit social concerns.

[11] On the other hand, this extended model generates some unappealing equilibria where A chooses In even if she expects B to $Take$.

[12] Results could be generalized by assuming that B players have heterogeneous preferences and that only the distribution of (α, θ) is common knowledge; A's choice would then depend on the distribution of (α, θ).

[13] Here B is not formally a psychological type because beliefs do not enter B's utility function directly, only his expected utility. Indeed, notice that B's expected utility from choosing $Take$ after In given belief σ_A^* is given by

$$\begin{aligned}\mathbb{E}\left[u_B(Take)|In,\, \sigma_A^*\right] &= \mu_B(\sigma_A^*)\left[m_B(In,Take) + (\alpha+\theta)m_A(In,Take)\right] + (1-\mu_B(\sigma_A^*))\left[m_B(In,Take) + \alpha m_A(In,Take)\right] \\ &= m_B(In,Take) + \left[\alpha + \theta\mu_B(\sigma_A^*)\right] m_A(In,Take)\end{aligned}$$

I thank an anonymous referee for this careful remark.

B's decision to be. A key feature of these models is therefore B's second-order belief σ_B^{**} (i.e. B's belief about A's expectation of B's choice). In the Online Appendix (OA-A.3.1), I show that the model I propose produces similar comparative statics as the model of Dufwenberg and Kirchsteiger (2004), without requiring the elicitation of an additional set of beliefs, and if anything, can account for a larger portion of the data. This model was therefore chosen for its tractability.

2.2.3. Comparative Statics and Equilibria

In this subsection, I analyze the game for some arbitrary information structure Ω_A with associated game form $\Gamma_1(p, \Omega_A)$; comparative statics with respect to the information structure are derived in Section 3. I first analyze the game without assuming equilibrium and then proceed to the equilibrium analysis. However, similarly to Charness and Dufwenberg (2006), the main analysis will be performed without making equilibrium assumptions, which seem quite restrictive in the environment under study.[14]

Assume In is realized. Then B will choose $Meet$ if and only if

$$10 + 10[\alpha + \theta\mu_B(\sigma_A^*)] \geq 14 + 2[\alpha + \theta\mu_B(\sigma_A^*)]$$

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

There are three cases to consider: (1) When $\alpha \geq \frac{1}{2}$, B exhibits strong altruistic concerns and therefore chooses $Meet$ irrespective of his beliefs; (2) When $\theta + \alpha < \frac{1}{2}$, B mostly cares about his own material payoffs and thus chooses $Take$ irrespective of his beliefs; (3) When $\alpha < \frac{1}{2}$ and $\theta + \alpha \geq \frac{1}{2}$, A's intentions matter for B. In particular when $\theta + \alpha > \frac{1}{2}$, B chooses $Meet$ if and only if

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

where $\bar{\sigma}(p, \alpha, \theta) := \frac{(\frac{1}{2}-\alpha)(1-p)}{2p(\theta+\alpha-\frac{1}{2})}$ is decreasing and convex in p, as well as decreasing in θ and α.[15]

Since $\bar{\sigma}(p, \alpha, \theta) \leq 1$ if and only if $p \geq \Lambda(\alpha, \theta) := \frac{1-2\alpha}{4\theta-(1-2\alpha)}$, one finds the following:

[14] The equilibrium assumption seems unsuitable to study one-shot interactions, where there is no learning. Furthermore, the restriction that beliefs be correct in equilibrium is particularly strong in a setting where players have belief-dependent motivations. I will therefore refrain from making this assumption when deriving the main comparative statics of Section 3. We note however that the key model predictions hold in equilibrium.

[15] In the knife-edge case where $\theta + \alpha = \frac{1}{2}$, B chooses $Meet$ if and only if $\mu_B(\sigma_A^*) = 1$ i.e. B must be fully convinced that In was A's decision; this can only occur when $p = 1$. To see how $\bar{\sigma}(p, \alpha, \theta)$ changes with p for different values of (α, θ), see OA-B.1.

- When $p \in [0,\ \Lambda(\alpha,\theta))$, B chooses $Take$ irrespective of his beliefs. As a result, A chooses Out.
- When $p \in [\Lambda(\alpha,\theta),1]$, B chooses $Meet$ iff $\sigma_A^* \geq \bar{\sigma}(p,\alpha,\theta)$. In turn, A chooses In iff $\sigma_A^{**} \geq \bar{\sigma}(p,\alpha,\theta)$.

Fixing α and θ, B is therefore more likely to choose $Meet$ as the amount of noise decreases (i.e. as p increases). Since A understands B's reciprocity concerns, A is in turn more likely to choose In as the environment becomes more transparent.

I now analyze the set of equilibria of this game for the non-trivial cases where intentions matter, that is, $\alpha < \frac{1}{2}$ and $\theta + \alpha > \frac{1}{2}$. In order to do so, consider the auxiliary game $\Gamma_2(p,\Omega_A;\sigma_A^*)$ obtained from $\Gamma_1(p,\Omega_A)$ by substituting utilities for material payoffs at the end nodes. Following Dufwenberg (2002), I adopt the following notion of equilibrium:

Definition: A strategy profile (σ_A, σ_B) of the noisy trust game is a *psychological sequential equilibrium* if: (1) (σ_A, σ_B) is a SPE in the standard game $\Gamma_2(p,\Omega_A;\sigma_A^*)$; (2) Beliefs are correct so that $\sigma_A = \sigma_A^* = \sigma_A^{**}$.

Solving $\Gamma_2(p,\Omega_A;\sigma_A^*)$ by backward induction, one finds a multiplicity of equilibria supported by different beliefs when the value of p is high but not equal to 1. This is the object of the following proposition:[16]

Proposition: *Suppose that* $\alpha < \frac{1}{2}$ *and* $\theta + \alpha > \frac{1}{2}$.

1. *When* $p \in [0,\ \Lambda(\alpha,\theta))$, *there is a unique equilibrium: B chooses* $Take$ *and A chooses* Out.

2. *When* $p \in [\Lambda(\alpha,\theta),1)$, *there is a multiplicity of equilibria:*

   (a) $\sigma_B = 0$ *and* $\sigma_A = \sigma_A^* = \sigma_A^{**} = 0$ *is an equilibrium of self-fulfilling mistrust. This is the only equilibrium with* $\sigma_A^* < \bar{\sigma}(p,\alpha,\theta)$.

   (b) $\sigma_B = 1$ *and* $\sigma_A = \sigma_A^* = \sigma_A^{**} = 1$ *is an equilibrium of self-fulfilling trust. This is the only equilibrium with* $\sigma_A^* > \bar{\sigma}(p,\alpha,\theta)$.

   (c) *In addition, there is a mixed strategy equilibrium such that* $\sigma_A = \sigma_A^* = \sigma_A^{**} = \bar{\sigma}(p,\alpha,\theta)$ *and* $\sigma_B = \frac{3}{8}$.

3. When $p = 1$, *there is a unique equilibrium: B chooses* $Meet$ *and A chooses* In.

Again, which strategy profile can be supported as an equilibrium outcome of the game depends on the amount of noise $(1-p)$. The cooperative outcome $(In, Meet)$ can only be supported as an equilibrium for sufficiently high values of p (that is, for values of p above $\Lambda(\alpha,\theta)$).

[16] See Appendix A at the end of this paper for a proof of this proposition.

3. Experimental Design and Hypotheses

I now introduce the experimental design, analyze the different treatments in light of the theory developed in the previous section and finally derive comparative statics predictions.

3.1 Experimental Design and Procedures

3.1.1 Treatments

The experimental design was built around two main treatment variables, which characterize the environment: (1) the probability p that A's decision is implemented; (2) the information structure Ω_A faced by B. The value of p varied within subject; more precisely, subjects played 11 noisy trust games, one for each value of p in the set $\{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. A large range of values was selected in order to analyze the sensitivity of individual strategies to the amount of noise. To facilitate comparisons with existing work, the boundary cases of $p = 0$ (dictator game) and $p = 1$ (standard trust game) were retained. On the other hand, the information structure Ω_A varied between subjects; in particular, subjects played the noisy trust game in one of the following three treatment conditions:

No Information (NI): B chooses between $Meet$ and $Take$ without knowing A's decision.

Exogenous Information (EI): B is exogenously informed of A's decision before making a choice.

Costly Communication (COM): Before B makes a choice, A can inform B of her true decision at a cost of 1 dollar and this is common knowledge.

The design of all treatments followed the same basic structure. At the start of each session, subjects were randomly assigned a role, either A or B, and matched with one single person in the other role for the entire session. All interactions took place through computer terminals.[17] A neutral frame was adopted in the instructions; in particular, B made a choice between "*Up*" (for $Take$) or "*Down*" (for $Meet$). The experiment was divided in two parts corresponding to the elicitation of strategies (Part 1) and the elicitation of beliefs about their matched partner (Part 2). While subjects knew at the beginning of the session that the experiment would have two parts, they were not told the purpose of the second part. Furthermore, subjects received no feedback about choices or earnings between the two parts. Each session ended with the administration of a short questionnaire to assess subjects' understanding of the experiment. I now describe Part 1 (elicitation of strategies) and Part 2 (belief elicitation) in detail.

[17] The experiment was programmed and conducted with the software z-Tree (Fischbacher 2007).

3.1.2 Elicitation of Strategies

In the first part, strategies were elicited using the *strategy method* (Selten 1967).[18] More precisely, A (B) made a choice between In or Out ($Meet$ or $Take$) for each possible value of p and was told that one value of p would be randomly selected for payment. Furthermore, in all treatments, B was asked to make a choice without knowing whether In was realized and to behave as if this was the case. Finally, in the last two treatments, B made a choice for each possible choice of A. That is, in treatment EI, B made a choice for each of the two cases, "A chose In" (case EI-In) and "A chose Out" (case EI-Out). In treatment COM, B made a choice for each of the following three cases: "A paid to inform you that she chose In" (case COM-In), "A paid to inform you that she chose Out" (case COM-Out) and "A didn't pay to inform you" (No-COM).[19] Therefore, A made 11 choices in each treatment, one for each value of p, while B made 11 choices in treatment NI, 22 choices in treatment EI and 33 in COM.[20] In COM, A also made a choice between informing and not informing B of her action for each value of p. A could only communicate her true action and the cost of \$1 was incurred only if B's decision affected the outcome (i.e. if In was realized).

3.1.3 Elicitation of Beliefs

Data on B's first-order beliefs and A's second-order beliefs was collected in the second part of the experiment. For each value of p, B was first asked to assess the likelihood that A chose In, which corresponds to B's first-order beliefs σ_A^*. In turn, A was asked to guess B's answer, which corresponds to A's second-order beliefs σ_A^{**}. In COM and for each value of p, B was also asked to assess the likelihood that A paid to inform B for each of the two cases (1) A chose In; (2) A chose Out. Again, A was asked to guess B's answer to these questions. Therefore, subjects answered 11 questions in total in the first two treatments and 33 questions in the last treatment. Subjects were paid according to their guess for the randomly selected value of p. In COM, one of the three blocks of questions was randomly selected for payment.[21]

B's first-order beliefs were incentivized using a method referred to as the Lottery Rule which,

[18] 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 should expect a downward bias on the effect of perceived intentions in the noisy trust game.

[19] While both theory and introspection suggest that the A players who chose Out would never choose to inform B of their choice, the case COM-Out was added to address the legitimate concern that A's choice to inform B may come from experimenter demand effects rather than from A's real understanding of the game.

[20] To keep the environment symmetric between players and across treatments, A and B were asked all 11 questions. However, choices for some questions were inconsequential (i.e. when $p = 0$ for A and when $p = 1$ in case EI-Out and COM-Out for B) and will be dropped from the econometric analysis of the aggregate results.

[21] More precisely, subjects were paid (with equal chance) either for their answer to the first block of questions or for their answer to the second/third block, depending on A's actual choice between In or Out. See instructions in OA-C.3 for more details.

unlike the widely used Quadratic Scoring Rule, provides a dominant strategy to reveal correct beliefs independently of the subject's risk attitudes. This rule is similar in spirit to the Becker-Degroot-Marschak (1964) mechanism.[22] More precisely, player B was asked to assess the percentage chance that A chose In (or paid to inform him) 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). 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 a correct guess (out of the 11 possible options) and 0 otherwise.

3.1.4 Summary of the Dataset

The dataset comprises 14 sessions (4 to 5 per treatment) which were all conducted at the Center for Experimental Social Science (CESS) of New York University. Overall, 226 subjects participated in the experiment with an average of 16 subjects per session. Each session lasted between 45 and 60 minutes and subjects earned on average between 10 and 15 dollars.[23]

Table 1: Summary Statistics

Treatment	# of sessions	# of pairs	Total # of obs.
No Information (NI)	5	38	A: 418 B: 418
Exogenous Information (EI)	4	36	A: 396 B: 792 ($= 396 \times 2$)
Costly Commmunication (COM)	5	39	A: 429 B: 1,287 ($= 429 \times 3$)
Total	14	113	A: 1,243 B: 2,497

Notes: Choices between In or Out ($Meet$ or $Take$) form the unit of observation. Number of observations for B in EI (COM) corresponds to choices for the two (three) cases EI-In and EI-Out (COM-In, COM-Out and No-COM).

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

[23] Because the sessions conducted for treatment EI were a bit larger on average than for the other two treatments, I conducted one session less for this treatment to obtain a similar number of subjects per treatment. Sessions were longer for COM than for NI and EI; as a result, subjects in COM received a \$7 show-up fee instead of \$5 in the other two treatments. It is not believed that this slight difference in show-up fee affected decisions in the game.

3.2 Predictions

The three treatments described above correspond to the following information structures:

$$NI:\ \Omega_A^1[S_A] = S_A \qquad EI:\ \Omega_A^2[S_A] = \{\{In\}, \{Out\}\}$$

$$COM:\ \Omega_A^3[S_A] = \Omega_A^1[S_A] \sqcup \Omega_A^2[S_A]$$

In the No Information treatment (NI), B has no information about A's choice; in this environment, A and B may hold any beliefs σ_A^*, $\sigma_A^{**} \in [0,1]$. By contrast, in the Exogenous Information treatment (EI), B is fully informed about A's choice before making a decision, implying σ_A^*, $\sigma_A^{**} \in \{0,1\}$. In both cases, the information structure is exogenously given to A and B. On the other hand, in the Costly Communication treatment (COM), the information structure is *endogenously* determined by A: it is given by $\Omega_A^2[S_A]$ if A chose "Communication" (C) (meaning σ_A^*, $\sigma_A^{**} \in \{0,1\}$); otherwise, it is given by $\Omega_A^1[S_A]$. In the latter case, let $\sigma_{A|NC}^* \in [0,1]$ denote B's posterior belief that A chose In given that she chose "No Communication" (NC), and define $\sigma_{A|NC}^{**} \in [0,1]$ as A's belief about $\sigma_{A|NC}^*$.[24] Due to the endogeneity of Ω_A^3, behavior in treatment COM is not directly comparable to behavior in the other two treatments, NI and EI. Instead, from the comparison of behavior in treatments NI and EI, I will draw conclusions for behavior in treatment COM.

To derive the main predictions, notice first that when mostly altruistic ($\alpha \geq \frac{1}{2}$) or mostly selfish ($\alpha + \theta < \frac{1}{2}$), B's behavior is insensitive to changes in the environment parametrized by (p, Ω_A): B chooses $Meet$ when $\alpha \geq \frac{1}{2}$ and $Take$ when $\alpha + \theta < \frac{1}{2}$, regardless of the environment. To derive comparative statics with respect to the environment, I therefore focus on the non-trivial case where intentions matter ($\alpha < \frac{1}{2}$ and $\theta + \alpha \geq \frac{1}{2}$). Results from Section 2 showed that for low values of p (i.e. when $p \in [0, \Lambda(\alpha, \theta))$), B chooses $Take$ and A chooses Out, regardless of the information structure. On the other hand, for high values of p (i.e. when $p \in [\Lambda(\alpha, \theta), 1]$), behavior depends on whether players' beliefs, σ_A^* and σ_A^{**}, are above the threshold $\bar{\sigma}(p, \alpha, \theta) = \frac{(\frac{1}{2} - \alpha)(1-p)}{2p(\theta + \alpha - \frac{1}{2})}$. Since this threshold is decreasing in p, one obtains the first prediction:

Prediction 1: Suppose σ_A^*, $\sigma_A^{**} > 0$. Then A's propensity to go In and B's propensity to $Meet$ are weakly monotone increasing in p.

Note that when $\sigma_A^* = 0$ (as in case *EI-Out* and *COM-Out*), the behavior of B should be insensitive to p: no matter the value of p, B is certain to be facing the computer's choice upon observing In and therefore chooses $Take$.

[24] That is, $\sigma_{A|NC}^* := \mathbb{E}_B[\sigma_A \mid \text{A chose "No Communication"}]$ and $\sigma_{A|NC}^{**} := \mathbb{E}_A[\sigma_{A|NC}^*]$.

I now compare behavior in NI and EI across all values of p. For low values of p, behavior should be identical in both treatments: B chooses $Take$ and A chooses Out. For high values of p, behavior will depend on beliefs. Under NI, the information structure puts no constraint on beliefs. In this context, whether subjects' beliefs are above the threshold $\bar{\sigma}(p, \alpha, \theta)$ will likely depend on individual characteristics. On the other hand, under EI, the information structure fully determines players' beliefs:

- If A chooses In, then $\sigma_A^* = \sigma_A^{**} = 1$. Thus, B chooses $Meet$ for all $p \in [\Lambda(\alpha, \theta), 1]$.
- If A chooses Out, then $\sigma_A^* = \sigma_A^{**} = 0$. Thus, B chooses $Take$ for all $p \in [\Lambda(\alpha, \theta), 1]$.
- As a result, A chooses In for all $p \in [\Lambda(\alpha, \theta), 1]$.

Finally, notice that in the limit as p tends to 1, $\bar{\sigma}(p, \alpha, \theta)$ tends to 0. In this case, behavior in NI is identical to EI: B chooses $Meet$ and A chooses In. Intuitively, when $p = 1$, B cannot be convinced that A chose Out conditional on observing In (i.e. $\mu_B(\sigma_A^*) = 1$, irrespective of σ_A^*). Thus, one obtains the second prediction:

Prediction 2:

(2a) For any value of p, B's propensity to $Meet$ in NI should be weakly lower compared to EI-In and weakly higher compared to EI-Out.

(2b) In turn, A's propensity to go In should be weakly higher in EI compared to NI.

In particular, subjects in role A who understand B's reciprocity concerns should feel more confident about choosing In in a more transparent environment, that is, under perfect information and for high values of p. Furthermore, the benefits of perfect information should vanish as p tends to 1.

This observation has consequences for COM: if offered the option, subjects in role A may choose to pay for a perfect information environment provided p is sufficiently high but strictly below 1. To understand the trade-off faced by A in COM, first notice that if A chooses to inform B of her action, B will behave as in EI: B will only choose $Meet$ if $p \in [\Lambda(\alpha, \theta), 1]$ and A chose In (implying $\sigma_A^* = 1$); otherwise, he will choose $Take$ (i.e. if $p \in [0, \Lambda(\alpha, \theta))$ and/or A chose Out). Since communication is costly, A will therefore never choose this option if p is low or she chose Out. On the other hand, if A chooses not to inform B, the latter must form a guess $\sigma_{A|NC}^*$ about A's action as in NI: here, B chooses $Meet$ if and only if $\sigma_{A|NC}^* \geq \bar{\sigma}(p, \alpha, \theta)$ and $p \in [\Lambda(\alpha, \theta), 1]$; otherwise B chooses $Take$. As a result:

- When $p \in [0, \Lambda(\alpha, \theta))$, A chooses (Out, NC).
- When $p \in [\Lambda(\alpha, \theta), 1]$, A chooses (In, NC) iff $\sigma_{A|NC}^{**} \geq \bar{\sigma}(p, \alpha, \theta)$. Otherwise, A chooses (In, C).

Notice that the relationship between p and A's propensity to choose (In, C) will be flat provided $\sigma^{**}_{A|NC} \geq \bar{\sigma}(p, \alpha, \theta)$ for all $p \geq \Lambda(\alpha, \theta)$, implying no signalling, and otherwise will be inverted U-shaped. For low values of p, no A should choose (In, C); for intermediate values of p (i.e. above $\Lambda(\alpha, \theta)$ but strictly below 1), the pessimistic A players will choose (In, C); however, since $\bar{\sigma}(p, \alpha, \theta)$ is decreasing and convex in p, A's propensity to switch to (In, NC) should be weakly increasing in p over $[\Lambda(\alpha, \theta), 1]$. In the limit, as $p = 1$ (implying $\bar{\sigma}(p, \alpha, \theta) = 0$), there is no benefit to communication and therefore all the A players should choose (In, NC).

Prediction 3:

(3a) For any value of p, B's propensity to *Meet* in *No-COM* should be weakly lower compared to *COM-In* and weakly higher compared to *COM-Out*.

(3b) A's propensity to communicate *In* as a function of p should be either flat or inverted U-shaped.

All predictions are summarized in Table A of the Appendix at the end of this paper.

4. Results

This section tests Predictions 1-2-3. In 4.1, I test whether the prosociality of A and B increases when A's decision is implemented with higher probability (Prediction 1). In 4.2, I contrast treatments *NI* and *EI* to study how prosociality depends on what B knows about A's action (Prediction 2). Finally, in 4.3, I focus on treatment *COM* to test whether the availability of a costly signalling technology fosters trust outcomes by inducing A to reveal her intentions to B (Prediction 3).

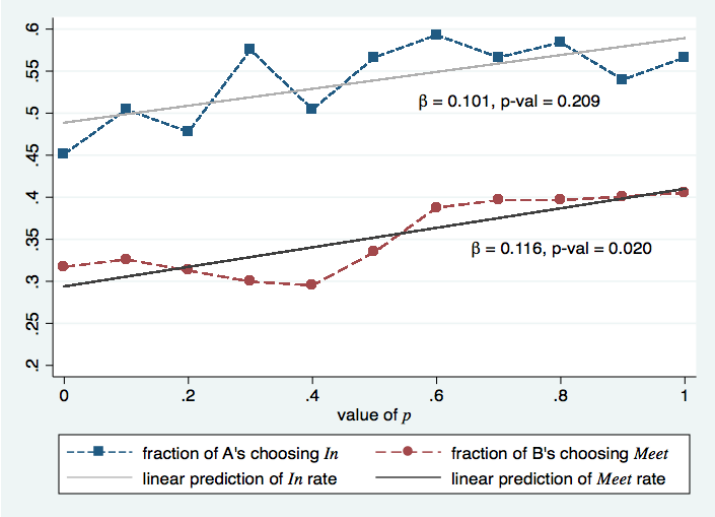
All baseline regressions mentioned in the analysis are *OLS* regressions where the unit of observation is either a binary choice or a guess in $\{0, 0.1, ..., 1\}$ made by a given subject for a specific value of p and treatment (case). Standard errors are clustered at the subject level (or at the level of a match when appropriate) to account for within-subject correlation across decisions. Significance is assessed with one-sided t-tests when there is a directional hypothesis and two-sided t-tests otherwise. Choices which were inconsequential (when $p = 0$ for A and when $p = 1$ in cases *EI-Out* and *COM-Out* for B) are excluded from the regression analysis; results are qualitatively similar in the full sample (see econometric analysis in OA-B).

4.1 Is prosociality sensitive to the value of p? (Prediction 1)

For each value of p, Figure 2 shows the fraction of A (resp. B) players who chose *In* (resp. *Meet*), pooling observations across all treatments. Prosocial behavior is overall slightly monotone increasing in p for both A and B, although the trend is only significant for B and is essentially driven by behavior

in *COM-In* and, to a lesser extent, by behavior in *COM-Out*, while the theory predicted no effect in the latter case.[25] I return to this finding in Section 5 when analyzing individual behavior.

Figure 2: Proportion of *In* and *Take* choices by value of p



Notes: Averages presented for the full sample to preserve symmetry between players; in particular for $p = 1$, B's decisions in *EI-Out* and *COM-Out* are included in the computation of the average. Reported β coefficient is from a linear regression of an indicator $In = 1$ ($Meet = 1$) on the value of p for the full sample; corresponding coefficient on restricted sample is $\beta = 0.076$ (p-value $= 0.366$) for A and $\beta = 0.127$ (p-value $= 0.016$). See OA-B.2.1 for details.

Although there is no clear linear trend, I show in the Online Appendix (Table 6 & 7, OA-B.2.2) that prosociality is sensitive to whether p takes a low, medium or high value, where $p_L \in \{0, 0.1, 0.2\}$, $p_M \in \{0.3, 0.4, 0.5, 0.6\}$ and $p_H \in \{0.7, 0.8, 0.9, 1\}$.[26] First, A's propensity to go *In* is significantly lower when p is low compared to medium or high, except for treatment *NI* where behavior is irresponsive to the noise. Secondly, B's propensity to *Meet* is significantly higher when p is high rather than medium or low in *NI*, *COM-In*, *COM-Out*. Again, this latter finding is at odds with the theory, as is the lack of effect of p in *EI-In* and *No-COM*, results which are discussed in Section 5.1. But interestingly, changes in prosociality seem to occur at different noise thresholds for A and B: while A's tendency to go *In* significantly increases from low to medium values of p and remains relatively constant afterwards, B's propensity to *Meet* significantly increases only once p reaches high values. Corroborating this finding, the mean threshold value $\bar{p}^*$ at which monotonic

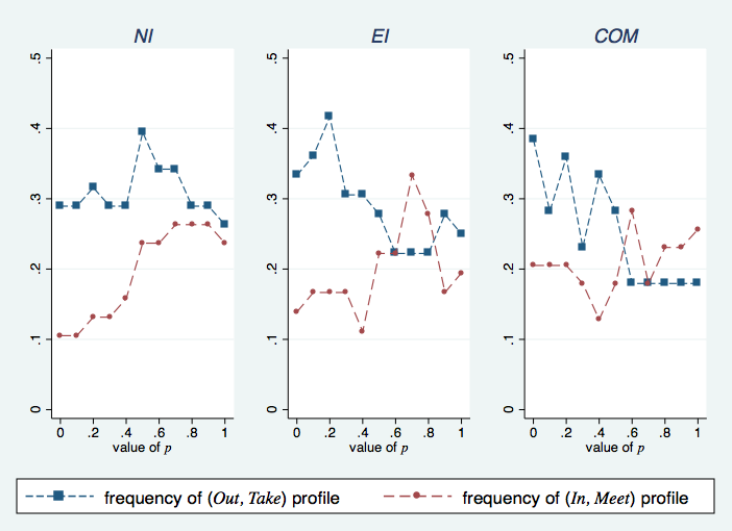
[25] Regression results show a positive and significant effect of p in *COM-In* ($\beta = 0.280$, p-value $= 0.02$) and *COM-Out* ($\beta = 0.216$, p-value $= 0.06$). The effect is also positive in *NI* but misses marginal significance ($\beta = 0.177$, p-value $= 0.13$), and is close to zero in the other cases. See OA-B.2.1 for more details.

[26] These findings are only partially robust to alternative splits of the data at the cutoff values 0.3 and 0.7. Although the qualitative results remain unchanged, observed differences lose significance for A with alternative splits, except for treatment *CI* when comparing behavior at p_L and p_M. Results are robust for B in all treatment cases and, if anything, monotonicity is stronger with alternative splits. See OA-B.2.2, Table 8 & 9, for more details.

subjects switch to prosociality is significantly higher for B than for A (0.54 vs 0.40, p-value = 0.007 on a two-sided t-test) and the distribution of individual thresholds p^* for B is located to the right of A (see Figure 1 Panel (b) in OA-B.1). Hence, B's reciprocity appears to be less robust to the introduction of noise than A's trust.[27]

Combining the behavior of A and B, another way to see the positive impact of reduced noise on prosocial behavior is through the distribution of realized action profiles. If intentions matter, the Trust action profile $(In, Meet)$ should be more likely to emerge as the value of p increases, and conversely for the No Trust profile $(Out, Take)$. Remember that each subject was paired with a single partner and made decisions for each possible contingency (that is, for each value of p and for B, for each possible choice of A). Pooling observations across all pairs, Figure 3 shows the frequency with which the profiles $(In, Meet)$ and $(Out, Take)$ were realized at each value of p.

Figure 3: Frequency of $(In, Meet)$ and $(Out, Take)$ profiles by value of p



Overall, more (resp. fewer) realizations of the $(In, Meet)$ (resp. $(Out, Take)$) profile occur as p increases; furthermore, the frequency of the Trust profile catches up with No Trust once p exceeds 0.5, except in NI. Table 2 confirms these findings: the effect of p goes in the right direction in all cases and is statistically significant in half of the cases: in NI, the fraction of $(In, Meet)$ significantly increases with p, while in EI and COM, the fraction of $(Out, Take)$ significantly drops.

[27] It is also worth noting that despite being embedded in a more complex setting, decisions at the extreme values of p (0 and 1) are very much in line with previous findings in the literature, with similar rates of trust and reciprocity, and a higher tendency towards prosociality when $p = 1$ compared to when $p = 0$. See OA-B.2.3 for a comparison of these findings with the literature.

Table 2: % (freq.) of the Trust and No Trust profiles by treatment and value of p

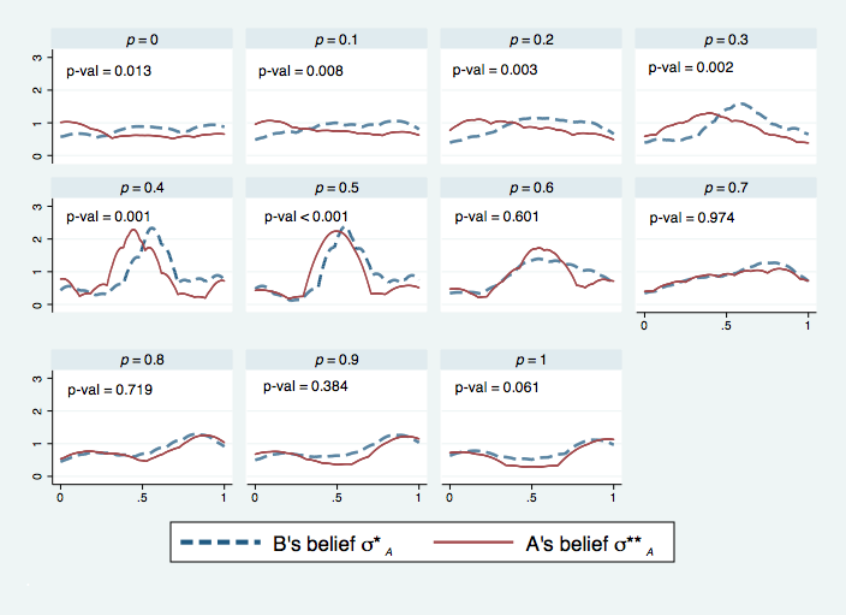
Action profile	(In, $Meet$)				(Out, $Take$)			
Treatment	*All*	*NI*	*EI*	*COM*	*All*	*NI*	*EI*	*COM*
$p \leq 0.5$	**16.64**	**15.26**	16.67	17.95	**31.50**	31.58	**33.33**	**29.74**
	(94/565)	(29/190)	(30/180)	(35/195)	(178/565)	(60/190)	(60/180)	(58/195)
	$\wedge^{***}$	$\wedge^{**}$	$\wedge$	$\wedge$	$\vee^{**}$	$\vee$	$\vee^{**}$	$\vee^{**}$
$p > 0.5$	**24.86**	**25.26**	25.90	23.59	**23.05**	30.53	**20.48**	**17.95**
	(137/551)	(48/190)	(43/166)	(46/195)	(127/551)	(58/190)	(34/166)	(35/195)

Notes: Results from a linear probability model with the dependent variable equal to 1 if an individual pair played the profile (In, $Meet$) (respectively, (Out, $Take$)) and 0 otherwise, regressed on dummies for $p \leq 0.5$ and $p > 0.5$; results qualitatively similar for cutoff values $p < 0.5$ and $p \geq 0.5$ (see OA-B.2.5). Realized action profiles for $p = 0$ dropped from all treatments; action profiles for $p = 1$ dropped in EI and COM when A chose Out. Standard errors clustered at the level of each pair. Significance assessed with one-sided t-tests; ** indicates p-value < 0.05.

A final step to understand the effect of p on individual decisions is to look at subjects' belief patterns. Remember that for each treatment and value of p, B assessed the likelihood that A chose In by picking a number between 0 and 100% (in increments of 10), thus eliciting σ_A^*; then A guessed B's answer, corresponding to σ_A^{**}. If intentions matter and players form consistent beliefs, then σ_A^* and σ_A^{**} should be increasing in p. Figure 4 shows the distribution of A's and B's beliefs for each value of p. The first thing to notice is that the distribution of beliefs is very sensitive to the amount of noise, with the mode of the distribution progressively shifting towards 1 for both A and B. In other words, as p increases, B becomes progressively more convinced that A will choose In and A rightly guesses so. In line with players' behavior, major changes in the shape of the distribution of beliefs occur at low, medium and high values of p. When $p < 0.3$, there is a large variance in subjects' beliefs, with A making more pessimistic predictions than B. When $p \in [0.3, 0.6]$, beliefs progressively become more centered around 0.5. Finally, when $p > 0.6$, the mass of both distributions moves towards 1 (with a second mode at 0) and players' beliefs become more closely aligned. Therefore, perceptions of prosociality are also greatly shaped by the amount of noise.

***Conclusion 1**: At the aggregate level, the introduction of noise appears to be detrimental to prosocial behavior and expectations thereof. In particular, B's reciprocity appears to be a less robust phenomenon than A's trust as A's action becomes less likely to be implemented. One consequence is that lower values of p are associated with fewer realizations of the Trust profile and/or more realizations of the No Trust profile.*

Figure 4: Distribution of beliefs of A and B by value of p



Notes: Distributions are kernel density estimates; p-values correspond to Kolmogorov-Smirnov tests of the hypothesis that the two distributions are identical. Observations pooled across all treatments ($N = 113$ for both A and B).

4.2 Are trust relations more likely when B knows A's action? (Prediction 2)

I now test the second prediction by contrasting prosocial behavior in treatments NI and EI, where the information structure is exogenously given to A and B. If intentions matter, B's propensity to *Meet* should be highest knowing that A chose In (case EI-In), followed by not knowing A's action (case NI), and finally, lowest knowing that A chose Out (case EI-Out). In turn, A's propensity to go In should be higher in EI compared to NI. Table 3 shows the percentage of In and *Meet* choices for each treatment (case), pooling observations across all subjects and values of p. Figure 5 presents the corresponding breakdown by value of p.

Starting with B, Table 3 shows that *Meet* is chosen more frequently when A chose In rather than Out (column 4 vs 5). This behavioral difference is observed for all values of p at the exception of $p = 1$.[28] Also consistent with the theory, the fraction of *Meet* choices is significantly higher when A's action is unknown compared to when A chose Out (column 3 vs 5); furthermore, differences between NI and EI-Out tend to increase as p increases, reaching significance for all $p > 0.3$ (except $p = 0.5$). Interestingly, there is no real distinction for B between knowing that A chose In and not knowing A's action (column 3 vs 4).[29] This finding is consistent with Cox and Deck (2006), who

[28] One-sided t-tests show that differences are significant for $p \in \{0.1, 0.2, 0.3, 0.6, 0.7, 0.8\}$.

[29] Despite a slightly higher fraction of *Meet* choices in NI compared to EI-In, differences are not significant when averaging over p. Breaking down by value of p, differences are marginally significant for $p \in \{0.4, 0.9, 1\}$.

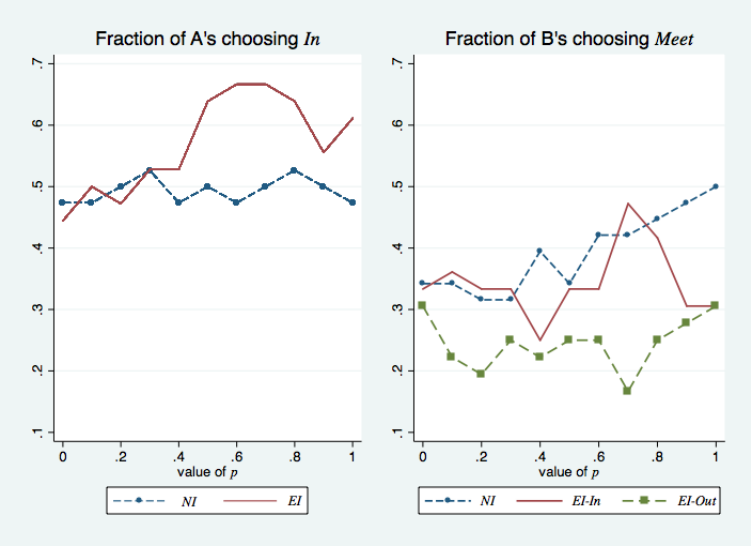
Table 3: Percentage of In and $Meet$ choices by information condition

	***In* choices of A players**		***Meet* choices of B players**		
Column	(1)	(2)	(3)	(4)	(5)
Treatment - Case	NI	EI	NI	EI-In	EI-Out
% of pro-social choices	49.47	58.06	39.23	34.34	23.89
t-Stat	-	1.00	0.56	2.86***	1.76**
Comparison		(1) vs (2)	(3) vs (4)	(4) vs (5)	(3) vs (5)
Predicted sign	-	(1) $\leq$ (2)	(3) $\leq$ (4)	(4) $\geq$ (5)	(3) $\geq$ (5)
Observations	380	360	418	396	360

Notes: Results from a linear probability model with the dependent variable equal to 1 if A (B) chose In ($Meet$) and 0 otherwise, regressed on dummies for the treatment (case). Observations for $p = 0$ excluded in the analysis for A; observations for $p = 1$ excluded in EI-Out. Standard errors clustered at the subject level. Significance is assessed with one-sided t-tests. ** and *** indicate p-value < 0.05 and 0.01 respectively.

find that B is not significantly less likely to cooperate in a noisy trust game where A's action is reversed with probability 0.25 compared to a standard trust game. Their interpretation is that B is willing to give A the benefit of the doubt. Altogether, these results suggest that B is more likely to punish a lack of trust than to reward a trusting act, an interpretation in line with previous studies documenting a higher propensity to punish harmful behavior than to reward friendly behavior (Offerman 2002, Charness 2004, Charness and Rabin 2002, 2005).

Figure 5: Proportion of In and $Meet$ choices in NI and EI by value of p



Does A foresee B's responsiveness to the information structure, trusting more frequently when her action is revealed to B? In line with the theory, the A players are more likely to choose In when B can condition his decision on A's action. When averaging over all values of p, the 8.6 percentage point difference between the two treatments is not statistically significant at conventional levels (column 1 vs 2 of Table 3). However, the theory only predicts behavioral differences when p is sufficiently high but not equal to 1 (that is, for $p \in [\Lambda(\alpha, \theta), 1)$): on the one hand, choosing In cannot be perceived by B as a signal of trust when A has little agency over the outcome; on the other hand, B cannot be convinced that A chose Out in NI if p is equal to 1 and In is the realized outcome. Figure 5 shows that differences between A's propensity to choose In in NI and EI depend on the value of p in the expected direction. For values of p strictly below 0.5, A's behavior is almost identical in the two treatments. However, the percentage of A players who choose In is systematically higher in EI for $p \geq 0.5$ and differences are statistically significant for $p = 0.6$ and $p = 0.7$.[30] Thus, the behavior of the A players is also sensitive to the information structure, provided A's decision is implemented with high enough probability.

Given that the behavior of both A and B is responsive to the information structure, a natural question to ask is whether the perfect information environment facilitates coordination on the Trust profile (In, $Meet$) and/or helps players to move away from the No Trust profile ($Out, Take$). Looking back at Table 2, one can see that although the relative frequency of Trust and No Trust is almost identical between the two treatments when $p \leq 0.5$ (i.e. 15.3% and 31.6% for Trust and No Trust in NI; 16.7% and 33.3% for Trust and No Trust in EI), differences emerge for $p > 0.5$ in the form of a lower frequency of (Out, $Take$) realizations (30.5% in NI versus 20.5% in EI). At the same time, the proportion of (In, $Meet$) realizations is still nearly identical across the two treatments for high values of p, implying a higher level of miscoordination in EI compared to NI. This higher miscoordination in EI for $p > 0.5$ is 78% of the time due to A unduly trusting B, compared to 55% of the time in NI.

Conclusion 2*: When exogenously given to A and B, the information structure appears to affect the behavior of both players. First, B is more likely to sanction A for choosing Out in EI but gives A the benefit of the doubt in NI. In turn, A is more likely to choose In if her action is revealed to B, provided the noise is low but not too close to zero. Under these latter conditions, players tend to move away from the No Trust profile, but A chooses to trust B more often than what is optimal.*

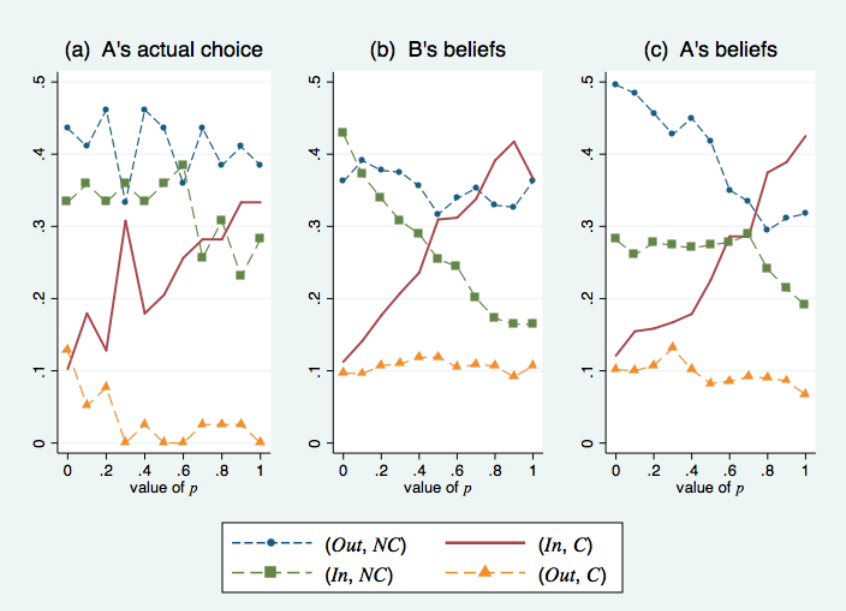
[30] The p-value on a one-sided t-test is 0.048 for $p = 0.6$ and 0.075 for $p = 0.7$.

4.3 Does costly signalling promote trust relations? (Prediction 3)

The previous section showed that B is responsive to the information structure, and in particular, as to whether A chose Out. As a result, A may have an incentive to communicate her intention to trust in COM. Remember that for each value of p, subjects in role A made a choice of action and information structure in $\{In, Out\} \times \{C, NC\}$, where C (resp. NC) refers to "Communication" ("No Communication"). Pooling observations across all subjects and values of p, A chose (In, C) about 25% of the time (97/390); the corresponding choice frequencies for (In, NC), (Out, NC) and (Out, C) were respectively, 32% (125/390), 41% (159/390) and 2% (9/390).[31] These numbers indicate that subjects in role A did use the communication technology and seem to have understood its strategic nature. At the individual level, over 56% of the A players selected (In, C) at least once and decisions to communicate In represent 92% of the cases where communication was chosen.

The theory also predicts that the distribution of choices should depend on the value of p. For low values of p, communicating In cannot trigger a positive response of B since In is a weak signal of trust. For high values of p, communication may induce B to switch from $Take$ to $Meet$ by clearing his doubts about A's intentions. However, in the limit when $p = 1$, there is no value to communication since B cannot be convinced that A chose Out if In was realized. Figure 6 Panel (a) presents the distribution of A's choices in COM broken down by value of p.

Figure 6: Distribution of A's actual choices and beliefs thereof by value of p



[31] As in the previous sections, observations for $p = 0$ are excluded from the analysis since A's choice of action is inconsequential in this case. As will be seen below, most of the incoherent choices (Out, C) were made when $p = 0$.

The fraction of A players who chose (Out, NC) is fairly stable across values of p. In contrast, the proportion of A players who communicated In is globally increasing in p: while about 10% of the A players chose (In, C) when their choice was inconsequential (i.e. for $p = 0$), more than 33% of the A players did so when their choice was implemented for sure (for $p = 1$), a difference which is highly statistically significant (p-value = 0.013 on a two-tailed t-test). For high values of p, the increase in the fraction of (In, C) choices appears to have been partially compensated by a decrease in the fraction of (In, NC) choices, although the overall decrease is small.[32] These observations are only partially consistent with the theory. On the one hand, A is more likely to choose (In, C) for high values of p, that is, when there is hope to shift B's action from $Take$ to $Meet$. On the other hand, there is no drop in the proportion of A players who use the costly communication technology for $p = 1$; as much as 54.2% of the A players who chose In when $p = 1$ do pay to communicate their action. I return to this departure from the theory in the discussion section.

Are beliefs about A's decisions consistent with the behavioral patterns observed? Remember that for each value of p in COM, B was not only asked to form a guess σ_A^* of how likely A chose In, but also to assess the likelihood that A chose to inform B in case (1) A chose In, and (2) A chose Out. Denote these subjective likelihoods respectively by s_1^* and s_2^*. In turn, A was asked to guess B's answers to these 3 types of questions, denoted by σ_A^{**}, s_1^{**} and s_2^{**}. From these answers, one can compute B's joint belief that A made a given choice of action and information structure, as well as the corresponding second-order belief of A.[33] Panel (b) of Figure 6 shows B's beliefs about the distribution of choices made by A for each value of p, while Panel (c) presents the corresponding second-order beliefs of A.

The first thing to notice is that the B players correctly anticipate the monotonic pattern in A's propensity to choose (In, C). Despite slightly overestimating this propensity for high values of p, differences between B's beliefs and A's actual choice of (In, C) are not significant for any value of p, with an average percentage point deviation of 6.3. The A players are remarkably accurate in guessing the answers of the B players, with guesses that on average deviate from B's beliefs by 3.7 percentage points. The B (resp. A) players also predict a decrease in A's propensity to choose (In, NC) (resp. in B's beliefs), although the fall predicted by B is much sharper than the actual decrease. Somewhat curiously, B overestimates A's tendency to communicate Out and this is almost perfectly predicted by A. Overall, beliefs are however fairly consistent with each other and with actual behavior.

[32] The effect of p on the probability to use the communication technology among those who chose In is significant at the 10% level in a linear regression; this finding is robust to excluding observations for $p = 0$.

[33] B's belief that A chose (Out, NC) is computed as $(1-\sigma_A^*)(1-s_2^*)$; the corresponding beliefs for (In, NC), (In, C) and (Out, C) are respectively $\sigma_A^*(1 - s_1^*)$, $\sigma_A^* s_1^*$ and $(1 - \sigma_A^*)s_2^*$. The beliefs of A are computed in the same manner.

I now examine whether the A players who signal In benefit from a signalling premium in the form of an increase in B's propensity to choose $Meet$. If B is responsive to A's signalling of trust intentions, $Meet$ should be chosen more frequently in $COM\text{-}In$ compared to $No\text{-}COM$ and in $No\text{-}COM$ compared to $COM\text{-}Out$. Table 4 shows the percentage of $Meet$ choices for each treatment case and separating high and low values of p.

Table 4: Percentage of $Meet$ choices in COM by case and value of p

	$p \leq 0.5$			$p > 0.5$		
Column	(1)	(2)	(3)	(4)	(5)	(6)
Treatment Case	*COM-In*	*COM-Out*	*No-COM*	*COM-In*	*COM-Out*	*No-COM*
% of *Meet*	29.49	28.21	39.74	49.74	40.38	40.51
t-Stat	0.22	1.92**	1.35*	1.34*	0.00	1.71**
Comparison	(1) vs (2)	(2) vs (3)	(1) vs (3)	(4) vs (5)	(5) vs (6)	(4) vs (6)
Predicted sign	(1) $\geq$ (2)	(2) $\leq$ (3)	(1) $\geq$ (3)	(4) $\geq$ (5)	(5) $\leq$ (6)	(4) $\geq$ (6)
Observations	234	234	234	195	156	195

Notes: Results from a linear probability model with the dependent variable equal to 1 if B chose $Meet$ and 0 otherwise, regressed on dummies for the treatment case, dividing the sample into $p \leq 0.5$ and $p > 0.5$. Observations for $p = 1$ excluded in $COM\text{-}Out$; ; results essentially unchanged when including all observations. Standard errors clustered at the subject level. Significance is assessed with one-sided t-tests. * and ** indicate p-value < 0.1 and < 0.05.

Consistent with the theory, the B players do reward the signalling of trust intentions for high values of p by choosing $Meet$ more often in $COM\text{-}In$ relative to $COM\text{-}Out$ and $No\text{-}COM$. However, the computation of A's optimal strategy given B's actual play (see OA-B.3) shows that the signalling premium is too low to justify the signalling cost: (In, NC) systematically yields higher expected payoffs than (In, C) for high values of p, at the exception of $p = 0.6$ where (In, C) is optimal. We further note that for low values of p, no signalling triggers a significantly higher cooperation rate of B than communicating Out, but also than communicating In. This latter finding is at odds with the theory, although it is consistent with B's disbelief that A chose (In, C) when p was low.

***Conclusion 3**: When the information structure is endogenously determined by A, the A players use the costly communication technology in a strategic manner: communication is used almost exclusively to signal In and more often when A's decision is implemented with high probability. While B tends to reward A's signalling of trust intentions, the signalling premium nevertheless appears to be too low to justify the communication cost.*

5. Discussion

The aggregate analysis of Section 4 showed that intentions are a key determinant of B's prosociality and this fact is internalized by A. First, players' choices and beliefs appear to be responsive to the probability p that A's decision is implemented, with higher values of p being associated with more frequent realizations of the Trust profile relative to No Trust. Secondly, the behavior of both players is affected by the information structure: common knowledge of A's action leads players to move away from No Trust relationships as B tends to sanction A's lack of trust, which makes A trust more frequently. This pattern holds both when the information structure is exogenously given to the players and when it is endogenously determined by A. In particular, A uses costly communication as a way to credibly signal intentions to trust.

While these results provide support to the theoretical framework developed in this paper, two major findings run against the theory: (1) the absence of a clear monotone relationship between prosocial behavior and the amount of noise $(1-p)$, mediated by intentionality; (2) the high frequency of communication decisions in the absence of noise. These findings are discussed in turn below.

5.1 Understanding the weak effect of p on aggregate behavior: individual analysis

Results from Section 4 showed only moderate support for Prediction 1 regarding the existence of a monotone increasing relationship between prosocial behavior and the value of p. At the aggregate level, there is no clear trend in A's propensity to go In as p increases, in particular in NI where behavior is completely irresponsive to the noise. For B, the positive and significant trend is partially driven by behavior in $COM\text{-}Out$, where the theory predicted no effect, while aggregate behavior in $EI\text{-}In$ and $No\text{-}COM$ fails to be monotone in p.

To interpret these findings, the within-subject design adopted in this experiment can provide further insights through the study of individual strategies. Based on an inspection of individual choices, subjects were classified into five categories according to whether and how they switched their action as p increases. The first two categories correspond to subjects who never switch their action i.e. who always choose In ($Meet$) or always choose Out ($Take$). The next two categories refer to subjects who switch their action exactly once. Player A (B) is classified as *monotone increasing* if he switches only once from Out to In ($Take$ to $Meet$) as the value of p increases. Similarly, player A (B) is *monotone decreasing* if he switches only once from In to Out ($Meet$ to $Take$) as p increases. The last category refers to the non monotone subjects who switch multiple times.[34]

[34]To classify subjects, I discarded A's choice when $p = 0$ and B's choice when $p = 1$ for the cases $EI\text{-}Out$ and $COM\text{-}Out$. For instance, a subject who played In when $p = 0$ and Out afterwards was classified as "*always Out*". Accounting for all choices, the classification would be left unchanged for the B players, while it would be altered

As summarized in Table A (see Appendix), the theory presented in this paper can only rationalize the first three types of behavior. For the A players, there is a one-to-one relationship in all treatments between A's strategy and B's preferences: *always In* (resp. *Out*) is optimal if and only if $\alpha \geq \frac{1}{2}$ (resp. $\alpha + \theta < \frac{1}{2}$), that is, under the assumption that B is altruistic (selfish); on the other hand, A will adopt a *monotone increasing* strategy if B is assumed to care about A's intentions ($\alpha < \frac{1}{2}$ and $\theta + \alpha \geq \frac{1}{2}$). For the B players, the relationship between B's strategy and preferences depends on the treatment case. If A chose In or A's action is unknown, B chooses *always Meet* when altruistic, *always Take* when selfish and adopts a *monotone increasing* strategy if intentions matter. However, when A chose *Out*, *always Take* is optimal both if B is selfish and if B cares about A's intentions, so a monotone increasing behavior cannot be rationalized in this context. Provided that intentions matter for B, one should therefore expect:

1. a higher fraction of B players who *always Take* in *EI-Out* and *COM-Out*.
2. no *monotone increasing* B players in *EI-Out* and *COM-Out*; a positive fraction otherwise.
3. a positive fraction of *monotone increasing* A players in all treatments.
4. no *monotone decreasing* or *non monotone* behavior of A and B in any treatment.

Table 5 (resp. 6) presents the distribution of strategies for player A (resp. B).

Overall, 23% of the A players exhibit a monotone increasing behavior, consistent with the belief that intentions matter for B. This is the second largest category behind those who always choose In. With the 20% of subjects who always choose *Out*, the theory presented in this paper can rationalize close to 70% of the choices made by the A players. Interestingly, the fact that A's behavior is only weakly monotone increasing in p in the aggregate partly comes from a sizable fraction of A players (about 18% overall) who exhibit a monotone decreasing behavior. This fraction is highest in NI (21.1%), where it equals the fraction of monotone increasing types.[35] An examination of their answers to the exit questionnaire sheds some light on their motivations. Some subjects predicted that B would $Take$ if his decision were to count with a high enough chance (that is, if In occurred with a high probability). On the other hand, they expected B to choose $Meet$ for low values of p since *Out* occurs with a probability of at least $\frac{1}{2}(1-p)$.[36] In this respect, the adopted monotone decreasing strategy appears to be a best response to their belief about B's strategy.

for 9 subjects in role A with only a minor change in the overall distribution of behaviors (+3 monotone increasing subjects, +2 monotone decreasing, +2 non monotone).

[35] On the other hand, the absence of a clear monotonicity in COM seems mostly due to the presence of non monotonic subjects (25.6% compared to about 8% in both NI and EI). One reason might be the increased environment complexity in COM since A had to choose both an action (In or *Out*) and a communication pattern (NC or C).

[36] Comments are compiled in Section E of the Online Appendix; here is one example taken from (Session 3, NI):
- "[...] When p was equal to 50 or higher, I chose *Out* in order to increase my chances of earning \$5 over \$2 (assuming that every individual is looking out for her/himself). As p lowered, I gave my partner the benefit of the doubt, hoping that he or she would choose *Down*."

Table 5: Pattern of choice for A across all values of p

A's choice pattern	% of A's (freq.) in treatment: *All*	*NI*	*EI*	*COM*
always In	24.8 (28/113)	23.7 (9/38)	27.8 (10/36)	23.1 (9/39)
always Out	20.4 (23/113)	26.3 (10/38)	19.4 (7/36)	15.4 (6/39)
monotone increasing	***23.0*** (26/113)	21.1 (8/38)	27.8 (10/36)	20.5 (8/39)
monotone decreasing	17.7 (20/113)	**21.1** (8/38)	**16.7** (6/36)	15.4 (6/39)
non monotone	14.2 (16/113)	7.9 (3/38)	8.3 (3/36)	**25.6** (10/39)
Total % explained (freq.)	68.1 (77/113)	71.1 (27/38)	75.0 (27/36)	59.0 (23/39)

▭ : key patterns inconsistent with the theory

Turning to B, *always Take* is the largest category in all treatments, despite noticeable variation across cases. As predicted by the theory, this category is significantly larger in *EI-Out* and *COM-Out*.[37] There is also a large variance in the proportion of monotone increasing subjects across cases. This fraction is highest in *COM-In*, where the theory predicts that behavior should be highly sensitive to p, and lowest in *EI-Out*, where behavior should be insensitive to p. These results suggest that intentions are indeed a determinant of prosocial behavior.[38] With those who *always Meet*, the model can rationalize between 62% and 85% of the choices made by the B players.

However, two major findings run against the theory presented in this paper. First, over 10% of subjects adopted a monotone decreasing strategy in *NI*, *EI-In* and *No-COM*. In combination with the noise generated by non monotone play, the presence of monotone decreasing subjects could explain the null effect of p in these cases. Secondly, the fraction of monotone increasing B players

[37] The p-values from one-sided t-tests are 0.012 (0.085) for *EI-Out* against *EI-In* (*NI*) and 0.048 (0.092) for *COM-Out* against *COM-In* (*No-COM*).

[38] Answers provided in the questionnaire seem to give substance to this interpretation. Here are two examples:

"If A's choice is implemented and A chose *In*, I would appreciate A's decision and choose *Down*. If p is higher than 80[%], which means A's choice is more likely to be implemented, I choose *Down*." (Session 1, *NI*)

" [...] if role A chose to be *In* and her decision would be implemented with a chance of 100%, it seemed to me that I should choose *Down* as they were cooperative, and thus in my opinion, more deserving of their share. If I thought they would choose *Out* then I would pick *Up*, in order to sway the profit my way." (Session 3, *EI*).

Table 6: Pattern of choice for B across all values of p

B's choice pattern	% of B's (freq. in brackets) in treatment case: *NI*	*EI-In*	*EI-Out*	*COM-In*	*COM-Out*	*No-COM*
always Meet	18.4 (7/38)	16.7 (6/36)	13.9 (5/36)	15.4 (6/39)	12.8 (5/39)	23.1 (9/39)
always Take	39.5 (15/38)	41.7 (15/36)	***55.6*** (20/36)	35.9 (14/39)	***48.7*** (19/39)	41.0 (16/39)
monotone increasing	23.7 (9/38)	13.9 (5/36)	8.3 (3/36)	***33.3*** (13/39)	**23.1** (9/39)	15.4 (6/39)
monotone decreasing	**10.5** (4/38)	**11.1** (4/36)	5.6 (2/36)	7.7 (3/39)	5.1 (2/39)	**10.3** (4/39)
non monotone	7.9 (3/38)	**16.7** (6/36)	**16.7** (6/36)	7.7 (3/39)	10.3 (4/39)	**10.3** (4/39)
Total % explained (freq.)	81.6 (31/38)	72.2 (26/36)	69.4 (25/36)	84.6 (33/39)	61.5 (24/39)	79.5 (31/39)

: key patterns inconsistent with the theory

in *COM-Out* is not only positive but it is relatively large (23.1%). Although these two findings might appear to be disconnected at first glance, subjects' answers to the exit questionnaire point towards a common mechanism: in line with the expectations of the monotone decreasing A players, B's prosociality in both cases appears to be inversely related to his chances of determining the final outcome, which happens only if In is realized.[39] Otherwise stated, B is more likely to be nice when the cost of appearing nice is low, that is, when B's decision is likely to be inconsequential. One way to capture this idea is to suppose that B's decision to choose $Meet$ has a "warm glow" component (Andreoni 1989, 1990): B cares about *appearing* altruistic, whether his decision matters ex post or not. In OA-A.2, I present an extension of the model, which allows for a warm glow effect. To gather the main intuition, consider the case where $\alpha = \theta = 0$ and assume that B gets warm glow utility $\phi > 0$ from choosing $Meet$ in addition to his expected material payoffs. Define

[39] Here are two examples illustrating this point:
- "The higher the probability, the more frequently I decided to go with Up because it would benefit me more than choosing $Down$ would. When p was very low, I put $Down$ and as p got larger, I chose Up instead." (Session 3, NI)
- "My decisions depended on the value of p in that if A chose Out and his decision definitely was going to be implemented, I was more likely to choose $Down$ because my decision would not have mattered. The higher the probability that In would be the result, the more certain I was to choose Up. " (Session 1, EI)

$q(\sigma_A^*) := p\sigma_A^* + \frac{1}{2}(1-p)$ as B's subjective belief in the event "In was realized". Then B chooses $Meet$ over $Take$ if and only if

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

$$\text{i.e.} \quad \phi \geq 4q(\sigma_A^*)$$

Notice that the RHS is weakly increasing in p provided $\sigma_A^* \geq \frac{1}{2}$ and decreasing otherwise. As a result, the warm glow effect can generate both the monotone increasing responses observed in *EI-Out* and *COM-Out* (where $\sigma_A^* = 0$), as well as the monotone decreasing behavior observed in *EI-In* and *COM-In* (where $\sigma_A^* = 1$). The warm glow hypothesis is also supported by the beliefs of the monotone decreasing players in NI. For these B players, σ_A^* and $q(\sigma_A^*)$ are indeed monotone *increasing* in p; otherwise stated, the more they expect A to choose In (and therefore, In to be realized), the more likely they are to $Take$.[40]

5.2 Why does A pay for signalling when p=1?

The second main departure from the theory concerns A's propensity to pay to signal her action even when B is already guaranteed to be facing A's decision i.e. when $p = 1$. In this limit case, not only is there no drop in A's propensity to choose (In, C), but it also reaches a maximum. Furthermore, the B players form expectations in line with A's behavior and continue to give a signalling premium when $p = 1$ by choosing $Meet$ more often in *COM-In* than *No-COM* (51% vs 41%).

These findings suggest that signalling may carry additional value beyond the one of simply revealing A's action. In particular, by paying to inform B of her decision to go In, A also signals that she expects B to be more likely to $Meet$ if informed. If B cares about meeting A's expectations (for instance because he is guilt averse), then costly communication from A will indeed induce B to choose $Meet$. The explanations given by some of the A players in the exit questionnaire converge towards this interpretation of the data.[41] This mechanism could also help to explain why B's overall propensity to $Meet$ in *COM-Out* is almost 10 percentage points higher compared to *EI-Out* and only slightly lower compared to *COM-In*: by paying a cost to reveal her decision to go Out, A

[40] Among the monotone decreasing B players of treatment NI, a standard linear regression shows that the effect of p on σ_A^* (resp. $q(\sigma_A^*)$) is positive and highly significant (p-value = 0.01 for σ_A^* and = 0.002 for $q(\sigma_A^*)$). Similar regressions for the *No-COM* treatment case indicate that no such a relationship exists between p and $\sigma_{A|NC}^*$ or $q(\sigma_{A|NC}^*)$; if anything, the effect of p is negative but insignificant. The relationship however becomes positive (but still insignificant) when replacing the posterior $\sigma_{A|NC}^*$ with the prior belief σ_A^*.

[41] Here are two examples:

- "I did [inform B] to hope to add guilt into the mix." (Session 4, *COM*)

- "I informed my partner of my decision for those cases that I chose In. I felt this would encourage my partner to show generosity rather than act selfishly by showing them that I had appealed to their good nature instead of making the more cynical decision. [...]" (Session 2, *COM*)

signals her expectation that B will respond more favorably than under No Communication, which in turn should motivate B to choose *Meet*. Appendix B formalizes this intuition by assuming that B's preferences exhibit a form of guilt aversion.

5.3 Can other models better rationalize the data?

Given the above discussion, the reader may wonder whether other theories cannot provide a better fit of the data. In the Online Appendix (OA-A.3), I analyze comparative statics in the noisy trust game $\Gamma_1(p, \Omega_A)$ for three of the most popular models in the literature on social preferences: the model of sequential reciprocity of Dufwenberg and Kirchsteiger (2004), the model of simple guilt of Battigalli and Dufwenberg (2007) and the model of inequity aversion of Fehr and Schmidt (1999) (henceforth DK, BD and FS). In order to make meaningful comparisons, I assess the relative explanatory power of these models under the assumption of equilibrium behavior, which allows to make testable predictions without relying on beliefs that were not elicited in the experiment.[42]

I show that none of these models performs uniformly better than the model proposed in this paper and, if anything, they tend to perform worse. First, the DK model of sequential reciprocity overall produces similar comparative statics as the present model, including the inverted U-shaped relationship between costly signalling and the value of p not supported by the data; however, it cannot explain why over 20% of subjects exhibit prosocial behavior regardless of the value of p (*always In/Meet* types). Secondly, although the BD model of guilt aversion can explain why A may still choose costly signalling when $p = 1$, it does not predict that A's propensity to do so should increase with p. Furthermore, this model lacks predictive power to explain B's response when knowing that A chose *Out* and cannot account for the 23% of subjects in NI who exhibit a monotone increasing response to p. Finally, according to outcome-based models such as the FS model of inequity aversion, behavior should be insensitive to p and the information structure. This is true even if one assumes that players care about expected rather than realized payoffs, because in the present case A and B receive the same outside payoffs. To generate behavior that is responsive to p and/or the information structure in the FS model, one would need to consider a payoff structure in which B is behind when *Out* is realized.

[42] A key ingredient of DK (2004) and BD (2007) is indeed B's second-order belief σ_B^{**} i.e. (i.e. B's belief about A's expectation of B's choice). Since the present experiment did not elicit this type of beliefs, it remains an open question whether these models would perform better once the equilibrium assumption is relaxed.

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 transparent 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. To manipulate perceived intentions, I consider a general environment where I vary both the probability with which the Sender's action is implemented and the Receiver's knowledge about the Sender's action. Unlike most of the previous literature, the experimental design exploits both between- and within-subject variation to identify the effect of intentions, and uses the strategy method to elicit subjects' strategies. Moreover, I study the role of intentions from a new angle by allowing Senders to communicate their action at a cost. When actions are noisy signals of trust intentions, how much noise becomes too much noise for the trust-reciprocal outcome to emerge?

I find that whether the trust-reciprocal outcome emerges is indeed affected by the credibility with which Senders can signal their intentions to trust; furthermore, Senders understand the signalling content of their actions. First, prosociality is positively affected by how much agency Senders have over the outcome and by whether their action is common knowledge. As a result, players tend to move away from relationships of mistrust when the Sender's actions become more transparent signals of trust. Secondly, Senders use costly communication in a strategic manner, that is, as a way to signal their intention to trust. This paper therefore not only provides new evidence that trust intentions matter but also that potential trustors internalize this fact. One policy implication 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. For instance, in the baseline treatment (NI), the behavior of nearly 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. In particular, the prosociality of some players is inversely related to their chances of affecting the final outcome: in some sense, these players care more about *appearing* altruistic than being altruistic per se. One conjecture is that the direct-

response method would eliminate this concern by requiring subjects to make decisions only after having been informed of the outcome. If this conjecture were to be verified, this would suggest that the strategy method might not be as conservative as what was previously thought.[43] In any event, more research is needed regarding differences between the strategy and direct-response method.

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 (2014), "Incomplete Information Models of Guilt Aversion in the Trust Game", forthcoming in *Management Science*.

[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 (2007), "Guilt in Games", *American Economic Review, Papers and Proceedings*, vol. 97, 170-176.

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

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

[9] 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.

[10] 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.

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

[43] For instance, Brandts and Charness (2011) emphasize in their comparative survey of the two elicitation methods that there is no evidence of a treatment effect found using the strategy method that was not found using the direct-response method.

[12] Butler, J., P. Giuliano and L. Guiso (2013), "Trust, Values and False Consensus", forthcoming in the *International Economic Review*.

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

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

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

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

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

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

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

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

[21] 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.

[22] 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.

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

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

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

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

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

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

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

[30] 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.

[31] 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.

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

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

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

[35] Rand, D. G., D. Fudenberg, and A. Dreber (2013), "It's the thought that counts: the role of intentions in reciprocal altruism", working paper.

[36] 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.

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

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

Appendix

A Proof of Proposition (Page 10)

I only present the proof of Part 2 since Part 1 follows immediately from the computations in Page 9. So consider the case where $p \in [\Lambda(\alpha,\theta),1]$. If $\sigma_A^* < \bar{\sigma}(p,\alpha,\theta)$, then B chooses $Take$ ($\sigma_B = 0$) and, by backward induction, A chooses Out ($\sigma_A = 0$). Since beliefs must be correct in equilibrium, one obtains $\sigma_A^* = \sigma_A^{**} = \sigma_A = 0$. On the other hand, if $\sigma_A^* > \bar{\sigma}(p,\alpha,\theta)$, then B chooses $Meet$ ($\sigma_B = 1$) and, by backward induction, A chooses In ($\sigma_A = 1$). Again, since beliefs must be correct, one obtains $\sigma_A^* = \sigma_A^{**} = \sigma_A = 1$. Finally suppose that $p \in (\Lambda(\alpha,\theta),1)$ so that $\bar{\sigma}(p,\alpha,\theta) \in (0,1)$. In this case, there is an equilibrium in which players are mixing. Indeed, if $\sigma_A^* = \bar{\sigma}(p,\alpha,\theta)$, then B is indifferent between $Meet$ and $Take$. Since beliefs must be correct in equilibrium, we have $\sigma_A = \bar{\sigma}(p,\alpha,\theta)$ and $\sigma_A^{**} = \bar{\sigma}(p,\alpha,\theta)$. Finally, for A to mix, she must be indifferent between In and Out. This happens if $5 = 10\sigma_B + 2(1-\sigma_B) \Leftrightarrow \sigma_B = \frac{3}{8}$.

B Costly communication and guilt aversion

The discussion in Section 5.2 suggested that B might be responsive to the cost paid by A even when $p = 1$, for it signals A's optimistic expectations about B. To formalize this idea, let $\sigma_{B|h}^*$ denote A's confidence that B will $Meet$ after history h (where $h \in \mathcal{H} := \{In, Out, NC\}$ refers to the information that B possesses about A's action) and define $\sigma_{B|h}^{**}$ as B's beliefs about A's confidence, that is, $\sigma_{B|h}^{**} := \mathbb{E}_B[\sigma_{B|h}^*]$. Assume that the preferences of B are of the form:

$$u_B(\sigma, \sigma_B^{**}) = m_B(\sigma) + \gamma \sigma_B^{**} m_A(\sigma)$$

where $\gamma > 0$ measures B's "trust responsiveness", that is, the extent to which B cares about meeting A's expectations. One can think of trust responsiveness as a moral obligation to fulfill trust or as an aversion to letting down an A player who trusted. Continue to assume that A only cares about maximizing her material payoffs i.e. $u_A(\sigma_A, \sigma_B^*) = \mathbb{E}_A[m_A(\sigma)]$.

To see how A's decision to signal may encourage B to $Meet$, notice that if A chooses to communicate her decision to go In and B believes in A's rationality, then by forward induction reasoning, B will infer that

$$u_A((In,C), \sigma_{B|In}^{**}) \geq u_A((In,NC), \sigma_{B|NC}^{**})$$

$$\Leftrightarrow \; [p+\frac{1}{2}(1-p)][9\sigma_{B|In}^{**} + (1-\sigma_{B|In}^{**})] + \frac{1}{2}(1-p)\cdot 5 \geq [p+\frac{1}{2}(1-p)][10\sigma_{B|NC}^{**} + 2(1-\sigma_{B|NC}^{**})] + \frac{1}{2}(1-p)\cdot 5$$

$$\Leftrightarrow \; \sigma_{B|In}^{**} \geq \sigma_{B|NC}^{**} + \frac{1}{8} \;\; (*)$$

Thus, B should feel more trusted if A chose to pay for signalling.[44]

Furthermore, after any history h (whether A chose to communicate or not her action i.e. $1_{\{C\}} = 1$ or 0), B will choose $Meet$ if and only if

$$10 + (10 - 1_{\{C\}})\gamma\sigma^{**}_{B|h} \geq 14 + (2 - 1_{\{C\}})\gamma\sigma^{**}_{B|h}$$

$$\Leftrightarrow \ \sigma^{**}_{B|h} \geq \frac{1}{2\gamma} \ \ (**)$$

By condition $(*)$, $\sigma^{**}_{B|In} \geq \frac{1}{2\gamma}$ is satisfied whenever $\sigma^{**}_{B|NC} \geq \frac{1}{2\gamma}$ is satisfied, so B should $Meet$ more often when A chose to communicate her action.

This psychological game has a multiplicity of equilibria, depending on players' beliefs. For instance, $((Out, NC), Take)$ is an equilibrium of the game under the beliefs $\sigma^{*}_{B|h} = \sigma^{**}_{B|h} = 0$ for all $h \in \{In, Out, NC\}$. However, under forward induction reasoning and provided some restriction on the preference parameter γ, only $((In, C), Meet)$ and $((In, NC), Meet)$ can be equilibria of the game. To see this, suppose that $\gamma > 4$. Then if A chooses (In, C), B must at least believe that $\sigma^{**}_{B|In} \geq \frac{1}{8}$ by $(*)$. Since $\frac{1}{8} > \frac{1}{2\gamma}$, condition $(**)$ is satisfied so B will choose to $Meet$. Since beliefs must be correct in equilibrium, $\sigma_{B|In} = \sigma^{*}_{B|In} = \sigma^{**}_{B|In} = 1$. As a result, choosing Out cannot be optimal for A and by $(*)$, (In, C) will be optimal provided $\sigma^{**}_{B|NC} \in [0, \frac{7}{8}]$. Otherwise, (In, NC) will be optimal.

[44] Similarly, if A chose to communicate Out (i.e. $u_A((Out, C), \sigma^{*}_{B|Out}) \geq u_A((Out, NC), \sigma^{*}_{B|NC})$), B must believe:

$$\frac{1}{2}(1-p)[9\sigma^{**}_{B|Out} + (1 - \sigma^{**}_{B|Out})] + [p + \frac{1}{2}(1-p)] \cdot 5 \geq \frac{1}{2}(1-p)[10\sigma^{**}_{B|NC} + 2(1 - \sigma^{**}_{B|NC})] + [p + \frac{1}{2}(1-p)] \cdot 5$$

$$\Leftrightarrow \ \sigma^{**}_{B|Out} \geq \sigma^{**}_{B|NC} + \frac{1}{8}$$

Table A: Optimal strategies for A and B

Case 1: B is mostly altruistic ($\alpha \geq \frac{1}{2}$)

For all treatments (NI, EI and COM) and all values of $p \in [0,1]$:

A: In
B: $Meet$

Case 2: B is mostly selfish ($\alpha + \theta < \frac{1}{2}$)

For all treatments (NI, EI and COM) and all values of $p \in [0,1]$:

A: Out
B: $Take$

Case 3: Intentions matter for B ($\alpha < \frac{1}{2}$ and $\alpha + \theta \geq \frac{1}{2}$)

Treatment	$p \in [0, \Lambda(\alpha,\theta))$	$p \in [\Lambda(\alpha,\theta), 1]$
NI	A: Out B: $Take$	A: In iff $\sigma_A^{**} \geq \bar{\sigma}(p,\alpha,\theta)$ B: $Meet$ iff $\sigma_A^{*} \geq \bar{\sigma}(p,\alpha,\theta)$
EI	A: Out B: $Take$	A: In B: $Meet$ in EI-In $Take$ in EI-Out
COM	A: (Out, NC) B: $Take$	A: (In, NC) if $\sigma_{A\|NC}^{**} \geq \bar{\sigma}(p,\alpha,\theta)$ (In, C) otherwise B: $Meet$ in COM-In $Take$ in COM-Out $Meet$ in No-COM iff $\sigma_{A\|NC}^{*} \geq \bar{\sigma}(p,\alpha,\theta)$

Notes: This table summarizes the optimal strategies for player A and B as a function of the treatment and the amount of noise parametrized by p. Predictions do not assume equilibrium behavior.