CEIS TOR VERGATA Research Papers

Università di Roma Tor Vergata

ISSN 2610-931X

CEIS Tor Vergata

RESEARCH PAPER SERIES

Vol. 23, Issue 1, No. 593 – February 2025

Rage against the Machine or Humans?

Luca Delle Foglie, Stefano Papa and Giancarlo Spagnolo

Rage against the machine or humans?*

Luca Delle Foglie,† Stefano Papa,‡ Giancarlo Spagnolo,§

Abstract

We examine how betrayal aversion and ambiguity attitudes influence trust. To disentangle these effects, we use a Trust game and manipulate trustors' perception of being the intentional recipients of trustees' betrayal by varying the nature of the latter: a human or a machine that replicates human choices in probability. After confirming that this manipulation does not affect ambiguity attitudes or beliefs about others' behavior, we find that both factors significantly influence trust. Nonetheless, even when controlling for these attitudes and beliefs, participants exhibit lower trust in humans than in machine. Furthermore, using Noldus' FaceReader technology to measure emotions during trustors' decision-making process, we find that participants express greater anger toward human trustees. Our results indicate that both betrayal aversion and ambiguity attitudes play important roles in shaping trust decisions.

Keywords: Ambiguity attitudes · Anger · Betrayal cost · Emotions · FaceReader · Trust game

JEL Classification: A13 · C91 · D03 · D64 · D90

*We thank Pierpaolo Battigalli, Giorgio Coricelli, Giovanni Di Bartolomeo, Martin Dufwenberg, Charles Noussair, Maria Porter, James Tremewan and all the participants of "Developments in Economic Thinking: A Workshop in Memory of Gary Charness", "Workshop on Understanding Decision-making under Uncertainty through Financial Literacy and Emotions", "IBE Colloquium: Behavioral Economics, Then & Now" for their valuable comments.

†Department of Economics and Finance, University of Rome Tor Vergata; CIMEO.

‡Department of Economics and Finance, University of Rome Tor Vergata; CIMEO.

§University of Rome Tor Vergata, EIEF, SITE & CEPR.

1 Introduction

Trust is an important social and economic lubricant that can facilitate trade and growth (Arrow, 1972; Knack and Keefer, 1997; Algan and Cahuc, 2010), financial participation (Guiso et al., 2004, 2008; Giannetti and Wang, 2016), firm productivity (La Porta et al., 1996; Bloom et al., 2012), vaccine acceptance (Alsharawy et al., 2022) and democratic stability (Algan et al., 2017), among other things. Understanding the determinants of trust in society is therefore of first order importance.

Koehler and Gershoff (2003), Bohnet and Zeckhauser (2004) and Bohnet et al. (2008) proposed that the aversion to be taken advantage of, "betrayal aversion", is an important independent determinant of people's decision to trust others (in addition to beliefs and material incentives). Bohnet and Zeckhauser (2004) defined this component, and verified it's existence, in the form of a risk premium required by people to trust. A number of other studies followed up, confirming the importance of betrayal aversion in a variety of contexts.[1] However, these studies do not control for subjects' ambiguity attitudes.

Baillon et al. (2018) recently developed a way to measure ambiguity attitudes at individual level, and Li et al. (2019) show that these attitudes and ambiguity-neutral beliefs both matter for trust decisions. Li et al. (2020) suggest that the betrayal aversion measured in previous papers may be the result of not appropriately controlling for the source of ambiguity and for the complexity of the situation.

This paper reports results from an experiment designed to shed more light on the existence of betrayal aversion and its relationship with ambiguity attitudes. We use a Trust game where subjects' ambiguity attitudes are measured and controlled for following Li et al. (2019) methodology, and where the amount and source of ambiguity (other individuals' decisions) is kept constant. In this environment we manipulate the trustor's perception of being

[1] See for example Bohnet et al. (2008), Fehr (2009), Aimone and Houser, (2012), Aimone et al. (2015), Cubitt et al. (2017), Butler et al. (2016), Butler and Miller (2018) and Benndorf et al. (2024), among others. Some other studies, for example Fetchenhauer and Dunning (2012), Evans and Krueger (2017) and Humphrey and Mondorf (2021), failed to find betrayal aversion.

the direct recipient of the trustee's betrayal intention by introducing a "Machine" that randomly selects the trustee decision from a distribution of choices made by subjects in the same population to which human trustees belong. In addition, we measure the emotional reaction of individuals that decide whether to trust, which should be different in presence of betrayal aversion rather than only ambiguity aversion.

In more detail, Bohnet and Zeckhauser (2004) designed an innovative method based on minimum acceptable probabilities (MAPs) to measure possible betrayal costs in trust games while accounting for risk attitudes.[2] These authors compare the MAPs stated in the Trust game to those from a similar Risky Dictator game (where the resolution of the uncertain prospect is determined by nature and not by humans), finding that subjects stated higher MAPs in the Trust than in the Risky Dictator game. Assuming that subjects adhere to the Substitution Axiom of von Neumann–Morgenstern utility, they exclude that MAPs could have been affected by ambiguity attitudes, and conclude that betrayal aversion drove their results.

However, Li et al. (2019) show that ambiguity attitudes are significant determinants of the trustor decision, and that it is important to correct for ambiguity attitudes when measuring the expectations on the trustee decision. Moreover, Li et al. (2020) identifies a non-strategic element of ambiguity, named social ambiguity, implying that people tend to perceive actions taken by humans outside of any strategic interactions differently than they perceive natural events that lack human agency and free will. The authors argue that social ambiguity attitudes, combined with reasonable belief assumptions and accounting for the complexity of the game, can explain Bohnet and Zeckhauser's (2004) and Bohnet et al.'s (2008) findings without betrayal aversion. They then conduct an experiment, using a slightly modified version of the methodology developed by Baillon et al. (2018), to measure ambiguity attitudes in a treatment characterized by social ambiguity without strategic interaction and

[2]MAP is the minimum probability of receiving the good outcome for which the subject accepts to expose herself to an uncertain prospect instead of opting for a particular payoff. MAPs are also used in several of the follow-up studies on betrayal aversion mentioned in footnote 1.

moral implications, as well as in a Trust game (betrayal ambiguity treatment). They test for the presence of betrayal aversion in the form of a higher ambiguity aversion observed in the Trust game. They find the same level of ambiguity aversion in the two situations, concluding that there may be no room for betrayal aversion. This conclusion, however, appears based on the implicit assumption that betrayal aversion is solely measurable by a difference in ambiguity aversion between social and betrayal ambiguity treatments.

We believe that ambiguity attitudes are a necessary control in the analysis of the trustor decision, but that the results of Li et al. (2020) are not definitive evidence of the non-existence of betrayal aversion. Recent research shows that complex belief elicitation methodologies increase the salience of monetary incentives and reduce that of psychological factors relevant to the topic on which the belief is elicited (Gangadharan et al., 2024). Given the complexity of the methodology implemented to elicit ambiguity attitudes, it is possible that this makes the psychological cost deriving from the possible betrayal, during the performance of the task, less salient. Therefore, it may happen that subjects show the same level of ambiguity aversion but externalize their betrayal aversion showing both a lower propensity to trust and a more negative emotional state when they might be betrayed. We believe that one should verify the presence of betrayal aversion by looking at both the propensity to trust and emotions during the trust choice in two environments (where the experimenter exogenously vary the perception of the association between a direct betrayal intention and a bad economic outcome), controlling for ambiguity attitudes and ambiguity-neutral beliefs on the probabilities of possible economic outcomes.

To shed light on these issues, we study a Trust game augmenting it with a treatment aimed at reducing a trustor's perception of being the direct recipient of the human intention behind the trustee decision. This treatment consists of changing the nature of the trustee decision maker, while keeping constant the same human-driven ambiguity source: the trustee decision maker is either a participant as usual (H treatment), or a Machine (M treatment) that exactly replicates human's trustee decisions in probability based on a pool of choices made

by participants drawn from the same population. We then verify whether this treatment affects the propensity to choose trust while measuring and controlling at the individual level, with the method developed in Li et al. (2019), the ambiguity attitudes and the ambiguity-neutral expectations on the choice of the trustee decision maker.

Besides confirming the results by Li et al. (2019) on the importance of ambiguity attitudes in the decision to trust, we find that, ceteris paribus, subjects show a significantly lower propensity to trust in the H treatment.

Furthermore, we measure the emotions expressed at the facial level by the subjects during the trustor decision using Noldus' FaceReader.[3] We find that subjects manifested significantly more anger - an emotion identified in the literature as related to the discovery or anticipation of facing a betrayal[4] - during the trustor decision in the H treatment.

We believe that the lower propensity to trust and the higher level of anger expressed during the trustor decision in the H treatment, controlling for ambiguity attitudes and ambiguity-neutral expectations, are evidence consistent with the existence of betrayal aversion, together with ambiguity aversion.

Finally, we present additional findings that are consistent with the literatures on gender differences in the Trust game –we find significant gender differences in both the propensity to trust and trustworthiness, consistent with Croson and Gneezy (2009) and Buchan et al. (2008) and partially with Van Den Akker et al. (2020)– and on motivated reasoning, consistent with the literature surveyed by Gino et al. (2016).

Besides contributing to the literature on betrayal aversion and ambiguity attitudes already discussed above,[5] our work also contributes to the more general literature investigating the determinants of trust using the Trust game introduced by Berg et al. (1995). In addition

[3]To the best of our knowledge, this is the first empirical work in which subjects' emotions are measured during the trustor decision. Kugler et al. (2020), using Noldus' FaceReader, looked for correlations between trustor decision and emotions felt in the 10 seconds before the trustor decision.

[4]See e.g. Baron (1992), Koehler and Gershoff (2003), Gershoff and Koehler (2011), Aimone and Houser (2012) and Schniter et al. (2020).

[5]We contribute to the literature on ambiguity attitudes by confirming the results of Li et al. (2019) about those being important determinants of trust.

to providing evidence supporting the importance of expectations about others' trustworthiness in determining the propensity to trust, we also find evidence of self-similar reasoning consistent with Li et al. (2019).

The paper also contributes to the literature on trust in interactions between humans and machines or computers. Houser et al. (2010) and Schniter et al. (2020) do not find significant differences in average trusting behavior when humans interact with computers programmed to replicate human behavior, as in our paper, rather than with other humans. However, these studies do not control for ambiguity attitudes. We find that, after controlling for ambiguity attitudes, humans tend to trust machines more than individuals.

Diecidue et al. (2024) find that humans trust more the machine learning analyst because they attribute to it a higher accuracy rate. In our paper, instead, there is no difference in expected payoffs, which – together with the analysis of emotions – leads us to attribute the higher trust in the machine to betrayal aversion.

Following up on Abdellaoui et al. (2011) finding that subjects are sensitive to the source of uncertainty (e.g. weather vs stock market), Farjam (2019) finds that people (prefer computer generated to human-driven risk but) are indifferent between computer generated and human-driven ambiguity. In our study subjects made choices under ambiguity rather than risk, and we held the human-driven nature of ambiguity constant across treatments. However, the two treatments may have induced subjects to manifest different ambiguity attitudes and beliefs about the trustee decision. Therefore, we hypothesize and empirically verify that there are no differences in the above variables between the treatments.

Our study also contributes to the literature on emotions in the Trust game, including Capra (2004), Aimone et al. (2014), Engelmann et al. (2019), Kugler et al. (2020) and Schniter et al. (2020), among many others (see Farolfi et al. et al., 2021 for an extensive survey). We measure emotions during the trustor decision using Noldus' FaceReader and identify an increase in anger in the H treatment (relative to the M one). This could be linked to the anticipation of the possibility of being betrayed after choosing to trust, in line with

the literature on betrayal aversion and anger (see footnotes 1 and 4).

The rest of the paper unfolds as follows. Section 2 presents the experimental design, Section 3 some definitions and our hypotheses, and Section 4 some details on the analyzable sample. Sections 5 reports the results, Section 6 shows some additional findings, and Section 7 concludes.

2 Experimental Design

We implemented a variant of the experimental design developed by Li et al. (2019) modified in several aspects (see Fig. 1). The motivations for these modifications are discussed in detail in subsection 2.2. Here, we proceed to illustrate the experimental design.

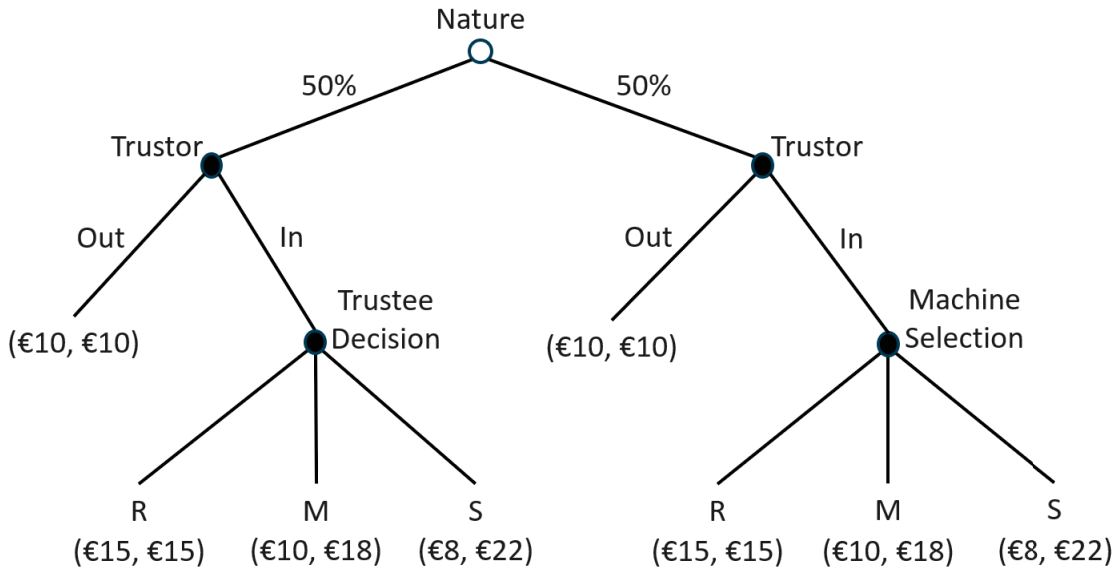

Fig. 1 Trust game

The subjects participate in a Trust game in which each begins by making the trustee decision, then faces 24 decision-making situations where they choose between an ambiguous prospect and a risky one, and finally make the trustor decision. After the experimenter reads the Trust game instructions aloud, participants first make the trustee decision without knowing whether they will be assigned to the Human (H) or Machine (M) trustee decision-maker

treatment. However, they are informed that half of the pairs will be randomly assigned to the H treatment and the other half to the M treatment with equal probability. Additionally, participants know that the roles of trustor and trustee/passive receiver will be assigned with equal probability in each randomly matched pair,[6] and they are informed of the selection rule the Machine will apply to select one of the available options in the trustee decision.[7] Regarding the trustee decision, participants must choose from the following three options:[8]

- Option A: Pay €15 to each of you
- Option B: Pay you €18, pay the partner €10
- Option C: Pay you €22, pay the partner €8

After all the subjects have made the trustee decision, each participant is privately informed via a message on the screen whether the effective trustee decision-maker for their randomly matched pair will be one of the two participants (with equal probability) or the Machine. Regarding the Machine, the participants were informed that it selects one of the available options for the trustee decision according to an unknown probability distribution, which is based on the choices made by other participants from the same University (and thus the same population) in previous similar experiments.[9]

After the trustee decision, subjects face a series of 24 decision-making situations involving ambiguous and risky prospects, designed to elicit ambiguity attitudes and ambiguity-neutral first-order beliefs about the choice made by the actual trustee decision maker (the partner in the H treatment, the Machine in the M treatment) between Option A (Reciprocate), Option B (Middle), and Option C (Selfish).[10] We implement the methodology developed by Baillon

[6]During the experiment, the words trustor and trustee were never used; the two roles were referred to as Role X and Role Y respectively.

[7]Subjects were also told that the option selected by Machine is implemented in the same way as if it had been chosen by the subject assigned the trustee role.

[8]The terms Reciprocate, Middle and Selfish were never used during the experiment; these choices were called Option A, Option B and Option C respectively.

[9]The trustee decisions used to build this probability distribution are those made by the participants who took part in the experiment conducted by Delle Foglie and Papa (2024).

[10]Before the 24 decision-making situations, an experimenter read aloud the instructions for this part of the experiment, and subjects answered control questions.

et al. (2018) and Li et al. (2019), with slight modifications outlined in section 2.2. In each of the 24 situations, subjects choose between an ambiguous payoff scenario, where the payoff depends on the trustee decision (or Machine selection), and a risky payoff scenario, where the payoff depends on a lottery with a known winning probability. An example of such a decision-making situation is shown below:

Please choose one of the following two Alternatives:

- Alternative 1: Pay you €3 if your partner chose Option A,[11] pay €0 otherwise
- Alternative 2: Pay you €3 with 50% chance, pay €0 otherwise

The above describes the first decision situation in a block of four, where the event on which the win depends in the ambiguous prospect (i.e., Alternative 1) is that the actual trustee decision maker (either the partner or the Machine) has chosen Option A in the trustee decision. Applying the bisection method,[12] it is possible to estimate the matching probability $m_A \in [5, 95]$ that makes the subject indifferent between a risky prospect of earning €3 with a probability of m_A% and a prospect of earning the same amount if the relevant counterpart (the partner in the H treatment, the Machine in the M treatment) chooses Option A.

This experimental phase consists of 24 decision-making situations, as it is necessary to measure both components of ambiguity attitudes—namely, the ambiguity aversion index and the ambiguity-generated insensitivity index as defined by Li et al. (2019)—and the ambiguity-neutral first-order beliefs. To this end, we need to elicit the matching probabilities for each single event (i.e., that the relevant counterpart chose Option A, B, or C) and composite event (i.e., that the relevant counterpart chose Option A or B, A or C, or B or C).

[11] The wording "your partner" is replaced with "the Machine" in the M treatment.

[12] The bisection method involves adjusting the winning probability in Alternative 2 as follows: In the first step, the winning probability is set to 50%. If the subject chooses Alternative 1 (or 2) in the first step, the winning probability is then increased (or decreased) by 24 percentage points in the second step. The third step follows the same procedure, but the change in the winning probability is 12 percentage points. The fourth and final step also follows the same pattern, with a change of 6 percentage points. Finally, in the last step, the matching probability for the specific event is elicited by adjusting the final winning probability by ±3 percentage points.

For each single and composite event, subjects must make four choices between ambiguous and risky prospects, resulting in a total of 24 choices. The order in which the subjects faced the 6 blocks was randomized at the individual level.[13]

Once all subjects have completed the 24 decision-making situations involving ambiguous and risky prospects, they proceed to a 20-second waiting screen. This waiting period allows for the simultaneous activation of the video cameras on all computers using μCap software (Doyle and Schindler, 2019). The first 10 seconds of video recording are used to measure the baseline emotional state of each subject immediately before making the trustor decision. Following this, the subjects make the trustor decision, which involves choosing between the following two options:

- Option 1: Follow your partner's ("the Machine" in the M treatment) instruction for payment
- Option 2: Pay €10 to each of you

During the trustor decision phase, videos are recorded at each computer station to measure the emotions experienced by the subjects throughout the decision-making process. Once all subjects have made their choices, two screens are presented sequentially: the first screen informs the subject of the role—either trustor or trustee—she has been randomly assigned within the pair, and the second screen reveals what the other party had chosen in the role assigned to the subject. Both screens are displayed for a predetermined duration of 20 seconds, with the first 10 seconds used to assess the emotional reaction to the specific stimulus received.[14] After the second screen, the video recording is stopped, and subjects are shown summary screens displaying their earnings from Part 1 (Trust game), Part 2 (24 decision-making situations between ambiguous and risky prospects) and their total payoff,

[13] A block refers to a sequence of 4 choices between ambiguous and risky prospects, where the event in Alternative 1 remains constant, but the winning probability changes based on the subject's previous choice between Alternative 1 and 2.

[14] The decision to use only the first 10 seconds to measure the emotional reaction to the received stimulus is based on prior relevant scientific literature (Breaban and Noussair, 2018).

including the show-up fee of €2, thus concluding the experimental session. Regarding the earnings for Part 2, each subject is compensated for only one (randomly selected) decision-making situation out of the 24, based on their choice between Alternative 1 and Alternative 2 for that particular situation.

In summary, after the subjects had read and signed the release form and after starting z-Tree, z-Leaf and μCap, each experimental session is composed of the following sequential steps summarized in Fig. 2:

1. Reading the instructions for Part 1 (Trust game) aloud and provided on the screen to each subject.

2. *Trustee decision.* Each subject chooses between Option A (Reciprocate), B (Middle) and C (Selfish) for the Role Y (trustee) decision knowing that, after having made it, each couple will be randomly assigned with equal probability to the H or M treatment.

3. *Treatment assignment.* Participants in each session are randomly paired, with half of the pairs randomly assigned to the H treatment and half to the M treatment. Each subject is notified on the screen, after making the trustee decision, which treatment her pair has been assigned.

4. Reading the instructions for Part 2 (24 decision-making situations) aloud and provided on the screen to each subject.

5. Completing control questions to check understanding of instructions.

6. *24 decision-making situations.* Subjects face, in individually randomized order, 6 blocks of 4 decision-making situations (24 situations in total) between ambiguous and risky prospects.

7. *Starting video recording.* After all subjects completed the previous step, the webcams at each station are started at the same time, i.e., 20 seconds before entering the screen

for the trustor decision. The first 10 seconds of video recording are used to measure pre-stimulus emotions.

8. *Trustor decision.* Each subject chooses between Option 1 (trust) and Option 2 (distrust) for the Role X (trustor) decision.[15]

9. *Role revelation.* After all subjects have completed the previous step, each subject is told what role, X (trustor) or Y (trustee), she has been assigned in the randomly matched pair.

10. *Trust game feedback.* Each subject is informed of what the counterpart, in the role assigned, has chosen, i.e., each trustor (trustee) is informed of the counterpart's trustee (trustor) decision. In particular, regarding the feedback given to the trustors, the trustee decision made by the actual counterpart is communicated: the choice made by the partner if their pair is assigned to the H treatment; the option selected by the Machine if their pair is assigned to the M treatment.

11. *End of video recording.* After all subjects completed the previous step, the webcams at each station are stopped simultaneously.

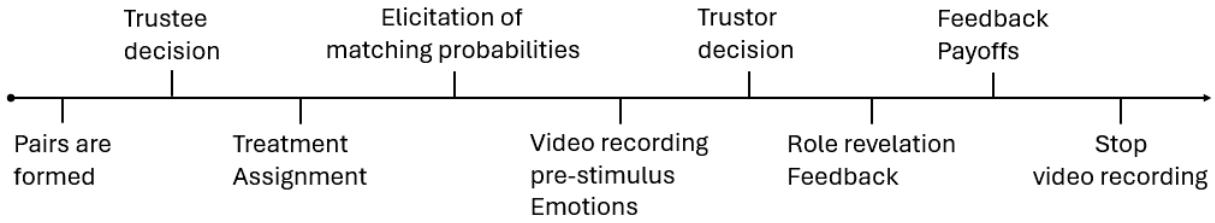

Fig. 2 Timing

[15]Thanks to the use of the μCap software, for each subject the section of the video (the duration of which was determined by the subject herself) in which she made the trustor decision was precisely isolated.

2.1 Procedures

The research project was pre-registered in the Open Science Framework Registries prior to data collection (`https://osf.io/6bvsd`). The experiment was conducted at the University of Rome Tor Vergata, using a between-subjects design with 216 students (across 7 sessions, of which 5 sessions included 32 subjects and 2 sessions included 28 subjects, with each session consisting of one round). Each participant took part in a single experimental session. Upon arrival, each subject was assigned to an isolated computer station equipped with a webcam and provided with a release form, which they were required to read and sign before the start of the experiment, granting their informed consent to be filmed for scientific purposes. The experiment was programmed using z-Tree software Fischbacher (2007) and conducted with the support of z-Tree, z-Leaf, and μCap software (Doyle and Schindler, 2019), ensuring precise temporal synchronization between the experiment and video recordings through input generated by programming code written in z-Tree. The experimental sessions were held between March and April 2024.[16]

2.2 Variations from Li et al. (2019)

In their experimental design, subjects first make the choice as trustor, then face 24 decision-making situations to elicit, at the individual level, the matching probabilities, the ambiguity aversion index, and the ambiguity-insensitivity index, before finally making the choice as trustee. In our experiment, the order of these choices is reversed: subjects first make the choice as trustee, then face the 24 decision-making situations, and finally make the choice as trustor. This change was implemented to make the two treatments identical, except for the nature of the trustee decision maker. If we had followed the experimental design of Li et al. (2019), subjects assigned to the M treatment would not have been able to make the trustee decision, which would have created an undesired difference between the two treatments.

[16]Specifically, the experimental sessions took place on the following dates: one session on March 25, one on March 26, two on March 27, and three on April 15, for a total of 7 sessions.

Our approach was to have all subjects make the trustee decision at the start of the experiment, even before treatment assignment (which was revealed to all subjects immediately after the trustee decision was made). At this point, subjects knew that the actual trustee decision maker would either be an experimental subject from the randomly matched pair (with roles of trustor and trustee/passive receiver assigned with equal probability) or the Machine with equal probability. This variation could have induced the experimental subjects to apply backward induction, in particular to find in their own trustee decision a basis for the elaboration of their expectations on the trustee decision of the counterpart. This is true in both treatments and it is possible to control this aspect by taking into account the appropriate control variables in the regression models, i.e., including the dummy variables *TrusteeR* (1 if the subject chose Reciprocate as trustee, 0 otherwise) and *TrusteeS* (1 if the subject chose Selfish as trustee, 0 otherwise) in the regression models.

Secondly, a key feature of Li et al. (2019)'s experimental design is that subjects are included in the analyzable sample only if their choices meet minimum requirements of logical consistency.[17] During the 24 decision-making situations involving ambiguous and risky prospects, we displayed a reminder on the screen that included the alternatives from the last decision-making situation in each block of four previously encountered situations, along with the subject's own choices in these specific situations. This approach was intended to ensure that any exclusion of subjects from the analyzable sample was due solely to a lack of discriminatory power between different levels of likelihood, rather than issues related to individual mnemonic abilities. The latter can play a significant role in shaping individual choices, as discussed by Battigalli and Generoso (2024) and demonstrated by Delle Foglie and Papa (2024).

Finally, in the experimental design of Li et al. (2019), subjects were paid for one of the 24 decision-making situations with a 96% chance (4% chance for each decision-making

[17]Specifically, subjects must not overestimate the probability of single events relative to that of event compositions. If a subject commits this logical fallacy repeatedly, it is not possible to estimate their ambiguity attitudes and ambiguity-neutral beliefs (Li et al., 2019).

situation), or for the trust game with a 4% chance of being selected as either trustor or trustee (2% chance for each role). In contrast, our experimental design paid subjects for both the Trust game (using the same payment structure as in Li et al., 2019) and for the elicitation of expectations.[18] The payment for expectations was set at 3 euros, ensuring that the gain from the elicitation of expectations accounted for only a small portion of the subject's total earnings from the experiment.

3 Definitions and Hypotheses

The main objective of this research is to determine whether the nature of the trustee decision maker—whether another person or a machine acting based on an algorithm—affects a subject's propensity to trust that counterparty, in a context where the ambiguity regarding the probability of the counterparty choosing/selecting a particular option from a set of available choices remains the same.[19]

As for ambiguity aversion and the lack of discriminatory power regarding different levels of likelyhoods (a-insensitivity), we use the indices developed by Baillon et al. (2018), while for the measurement of ambiguity-neutral first order beliefs we apply the formula developed by Li et al. (2019).

Definition 1 Ambiguity aversion index:

$$b = 1 - \overline{m}_S - \overline{m}_C \tag{1}$$

b is defined in $[-1, 1]$ where $b = 0$ indicates ambiguity neutrality, $b = -1$ means that the subject is strongly ambiguity seeker, and $b = 1$ means that she is strongly ambiguity

[18] Each subject was paid for one randomly selected decision-making situation out of the 24 they completed.

[19] While the selection of a particular option by a person among a set of available choices can be defined as a "choice," motivated by reasoning and intention, the same cannot be said for the selection made by a machine based on a probability distribution, as there is no reasoning or intention involved in the machine's action. Therefore, we refer to the decision made by an experimental subject as a "choice," while the decision made by the machine is referred to as a "selection."

averse. $\overline{m}_S = (m_R + m_M + m_S)/3$ where m_R, m_M, m_S are the matching probabilities for the single events that the counterpart choose *Reciprocate-Middle-Selfish* respectively, and $\overline{m}_C = (m_{RM} + m_{RS} + m_{MS})/3$ where m_{RM}, m_{RS}, m_{MS} are the matching probabilities for the composite events that the counterpart choose *Reciprocate or Middle - Reciprocate or Selfish - Middle or Selfish* respectively.[20]

Definition 2 a(mbiguity-generated)-insensitivity index:

$$a = 3\left[\frac{1}{3} - (\overline{m}_C - \overline{m}_S)\right] \tag{2}$$

a, defined in $[-1, 1]$, is a measure of the lack of discriminatory power of the subject regarding different levels of likelihood, where $a = 0$ indicates the maximum discriminatory power, while $a = 1$ indicates the total lack of discriminatory power.[21]

Definition 3 Ambiguity-neutral first order beliefs:

$$p_i = \frac{3(\overline{m}_C - \overline{m}_S) + 3m_i - 3m_{jk} + 2(1-a)}{6(1-a)} \tag{3}$$

where $\{i, j, k\} = \{Reciprocate, Middle, Selfish\}$. $p_{R,M,S}$ is defined in $[0, 1]$ and $p_R + p_M + p_S = 1$. Quoting Li et al. (2019): "These can be interpreted as the beliefs of an ambiguity neutral twin of the decision maker, who is exactly the same as the decision maker except that she is ambiguity neutral. That is, a-neutral probabilities are additive subjective probabilities that result after correcting for ambiguity attitudes.". Consistent with what Li et al. (2019) did, in our analyses we will use $p_R - p_S$ as a variable that measures the positivity (negativity) of the subjects' beliefs regarding the other trustee's decision.

To induce the same ambiguity attitudes regarding the choice/selection of the trustee decision maker in both treatments, we set the Machine's selection rule in the M treatment as

[20]The matching probability m (m_i or m_{ij}) of an event E (E_i or E_{ij}) is the probability such that the decision maker is indifferent between the ambiguous prospect of earning X if the event E happens, and the risky prospect of earning X with known probability $m\%$. For further details on this, see Li et al. (2019).

[21]See Baillon et al. (2018) for further details on this index.

a random draw from a set of trustee decisions made by a sample of subjects from the same population as those who participated in this experiment (all of which was made perfectly clear to the experimental subjects). This approach should also induce equal expectations regarding the probability distribution underlying the trustee's decision, since the probability that a subject will choose Reciprocate-Middle-Selfish in the H treatment is identical to the probability that another subject from the same population would have chosen Reciprocate-Middle-Selfish in a previous experiment (M treatment). The only distinction between a subject's choice and the Machine's selection lies in the presence of intentionality: the former reflects a deliberate intention by an individual to make their matched participant experience the consequences of their Reciprocate-Middle-Selfish choice, while the latter involves a random assignment of previously made choices, without any intention directed toward the person affected by the outcome. Given the selection rule applied by the Machine, we should assume that the assignment to the M treatment has no effect on ambiguity attitudes or ambiguity-neutral first-order beliefs. However, one might suspect that the experimental subjects could have found the Machine's selection rule difficult to understand. If this were the case, it could potentially impact ambiguity aversion, the a-insensitivity index, and ambiguity-neutral first-order beliefs. Therefore, we test the null hypothesis that the treatment assignment had no effect on these variables.

H2-3-4: The assignment to H or M treatment has no effect on the ambiguity aversion index, the a-insensitivity index, and $p_R - p_S$.

After confirming that the treatment has no effect on ambiguity attitudes and ambiguity-neutral beliefs, we proceed with the central research question of this paper. We hypothesize that any discrepancy in the propensity to choose trust between the two treatments, controlling for all other relevant factors, can be attributed to the experimental subject's perception, during the trustor decision, of the presence (or absence) of an intention to betray the trustor as a motivator for the selfish trustee decision.[22]

[22]One might argue that the trustor decision is also influenced by the fact that in the M treatment, there

H1: subjects assigned to the H treatment show, ceteris paribus, a lower propensity to trust than subjects assigned to the M treatment, consistent with betrayal aversion theory.

Finding evidence supporting H1-2-3-4 would lead us to conclude that subjects, ceteris paribus, anticipate greater disutility from trusting in the H treatment than in the M treatment. This would confirm the existence of a non-monetary loss associated with the possibility of experiencing the Selfish choice from a human partner, consistent with the concept of betrayal aversion as defined by Bohnet and Zeckhauser (2004) and Bohnet et al. (2008).

After the experimental sessions, Noldus' FaceReader software (`https://www.noldus.com/facereader`) was used to generate data on the emotions of subjects at the individual level. Specifically, during the experiment, subjects were filmed individually by a webcam (one per computer station) with their prior consent. The perfect synchronization of the video recording with the timing of the experiment was achieved using μCap software (Doyle and Schindler, 2019). After the experimental sessions, the videos were analyzed using FaceReader software available in the CIMEO laboratory at Sapienza University of Rome. The output produced by FaceReader consists of observations made every tenth of a second at the individual level for several variables: (I) the intensity of seven emotions (happiness, sadness, anger, surprise, scare, disgust, contempt), with each emotion measured independently on a scale from 0 to 1; (II) a synthetic indicator of the subject's emotional state called Valence; (III) an indicator of the degree of activity of the subject called Arousal.[23] In line with findings from Baron (1992), Schlösser et al. (2013), and Aimone et al. (2014), regarding

is a subject in the role of passive receiver, who is subject to both the trustor's choice and the Machine's selection of the trustee decision, without the right of reply. However, this is unlikely, as Houser et al. (2010) and Schniter et al. (2020) both found in similar experiments that the presence or absence of a passive human receiver has no influence on the behavior of a sender in an investment game. In this regard, we provide a check in Appendix B, suggesting that the difference in the trustor decision between the two treatments was not driven by prosocial motivations.

[23]Valence is calculated at each moment as the intensity of happiness (the only positive emotion) minus the intensity of the most intense negative emotion (sadness, anger, scare, or disgust). Arousal, based on the activation of 20 Action Units from the Facial Action Coding System (Ekman and Friesen, 1978), measures whether the subject is active or not. For more details on these variables, see *FaceReader Methodology Note* (`https://info.noldus.com/hubfs/resources/noldus-white-paper-facereader-methodology.pdf?utm_campaig`).

the importance of anticipated and action-related emotions in economic decision-making, we hypothesize that emotional changes occur in the mood of experimental subjects during the trustor decision in the H treatment, but not (or to a lesser extent) in the M treatment.[24] In particular, we hypothesize that in the H treatment, negative emotions arise from the prospect of trusting a human counterpart, whereas this does not happen (or happens less) in the M treatment, since the selection made by the Machine is not associated with a human intention explicitly directed toward the trustor.[25]

H5a: During the trustor decision, ceteris paribus, subjects assigned to the H treatment experience lower Valence than subjects assigned to the M treatment.

Valence captures only the most intense negative emotion experienced at a given moment, thereby overlooking the informational content provided by changes in other negative emotions. As outlined in the exploratory analysis section of the pre-registration, we also examine the presence of significant associations between the nature of the trustee decision maker and the individual emotions that contribute to the determination of Valence, namely happiness, sadness, anger, and scare.[26] We specifically focus on anger, as previous research by Baron (1992), Koehler and Gershoff (2003), Gershoff and Koehler (2011), and Aimone and Houser (2012) has found evidence suggesting a significant correlation between betrayal aversion and anger.

H5b: During the trustor decision, ceteris paribus, subjects assigned to the H treatment experience a higher (lower) level of anger, sadness, scare (happiness) than subjects assigned to the M treatment.

[24]Baron (1992) found that most people would feel angrier when blinded by a vaccine for a disease compared to being blinded by an untreated disease. Schlösser et al. (2013) demonstrated the predictive power of anticipated and action-related emotions in risky choice contexts. Aimone et al. (2014) found that betrayal aversion stems from a desire to avoid negative emotions resulting from the realization that one's trust was betrayed.

[25]In the M treatment, there may still be a high level of negative emotions related to the uncertainty faced by the subjects.

[26]We exclude disgust from the analysis because we do not believe it is an anticipated emotion related to betrayal aversion. However, we also tested this and found no association between the M treatment and the disgust experienced by subjects during the trustor decision.

Another piece of evidence for the existence of a betrayal cost can be found in the reactions of subjects who chose to trust when they receive feedback about the trustee's decision. If a betrayal cost exists, subjects who, after trusting the counterpart, learn that the trustee has chosen the Selfish option should experience more intense negative emotions when the trustee is human, compared to when the trustee is a Machine. We hypothesize that these emotions are likely to be anger and/or sadness, in line with the findings of Schniter et al. (2020).

H6-7: Subjects who are assigned the role of trustor, who choose to trust the counterpart, and who discover that the trustee decision maker has chosen (selected) the Selfish option, experience a higher (lower) level of sadness and/or anger if they are assigned to the H (M) treatment.

4 Analyzable sample

Out of the 216 subjects who participated in the experiment, not all could be included in the analyses. Specifically, subjects whose choices in the task for eliciting ambiguity attitudes and a-neutral beliefs about the trustee's decision were logically inconsistent could not be included, as it is not mathematically possible to estimate the required variables in such cases. In their study, Li et al. (2019) excluded from the analyzable sample all subjects who failed at least two monotonicity tests or who indirectly declared matching probabilities leading to $\overline{m}_S = \overline{m}_C$.[27] However, we find that for our data, this selection criterion is neither sufficient nor necessary. On one hand, there are subjects who (I) failed only one monotonicity test and (II) exhibit $\overline{m}_C > \overline{m}_S$ and $a < 1$, but for whom it is still not possible to estimate consistent expectations (6 subjects in total).[28] On the other hand, there are subjects who (I) failed

[27]The monotonicity test verifies the condition $m_{ij} \geq m_i$, meaning that the matching probability of a composite event E_{ij} (i.e., the counterpart choosing i or j) should not be less than the matching probability of a single event E_i. There are three single events, (E_R, E_M, E_S), and three composite events, (E_{RM}, E_{RS}, E_{MS}), so six monotonicity tests are performed for each subject.

[28]For example, one subject from this group has $m_R = 0.35$, $m_M = 0.41$, $m_S = 0.65$, $m_{RM} = 0.41$, $m_{RS} = 0.47$, and $m_{MS} = 0.77$, resulting in $\overline{m}_S = 0.47 < 0.55 = \overline{m}_C$ and $a = 0.76$. For this subject, $p_R = -0.375$, $p_M = 0.375$, and $p_S = 1$.

two monotonicity tests and (II) exhibit $\overline{m}_C > \overline{m}_S$ and $a < 1$, but for whom it is possible to estimate consistent expectations (5 subjects in total).[29]In the online Appendix C of Li et al. (2019), regarding the derivation of equation (3), one can find that the original monotonicity assumption consisted of ($m_{ij} \geq m_i$ and $\overline{m}_C > \overline{m}_S$). As we have demonstrated, the first condition is sufficient but not necessary, meaning that adhering to both conditions would exclude from the analyzable sample subjects who have perfectly estimated and analyzable values of a, p_R, p_M, and p_S.[30] In light of the above, we included in our analyzable sample 180 subjects who met the following two conditions: (1) $\overline{m}_C > \overline{m}_S$, ensuring $a < 1$, and (2) $m_{ij} + m_{ik} \geq m_j + m_k$ for $i, j, k = \{Reciprocate, Middle, Selfish\}$, ensuring that p_R, p_M, and p_S are non-negative and sum to 1. Further details on the second necessary and sufficient condition are provided in Appendix A. Tables 1 and 2 present summary statistics for the analyzable sample in the H and M treatments, respectively.

[29]For example, one subject from this group has $m_R = 0.29$, $m_M = 0.41$, $m_S = 0.35$, $m_{RM} = 0.59$, $m_{RS} = 0.23$, and $m_{MS} = 0.71$, leading to $\overline{m}_S = 0.35 < 0.51 = \overline{m}_C$ and $a = 0.52$. For this subject, $p_R = 0.0625$, $p_M = 0.6875$, and $p_S = 0.25$.

[30]The fact that this condition is not strictly necessary is implicitly acknowledged by Li et al. (2019), since they also include subjects who fail one monotonicity test in their analyzable sample.

	Summary statistics - H treatment				
	Mean	Median	SD	Min	Max
Trustor	0.52	1	0.50	0	1
Trustee	1.80	2	0.86	1	3
Ambiguity aversion	-0.07	-0.04	0.18	-0.56	0.40
a-insensitivity	0.25	0.22	0.27	-0.56	0.88
p_R	0.33	0.33	0.18	0	1
p_M	0.29	0.33	0.14	0	0.69
p_S	0.38	0.33	0.19	0	1
Male	0.43	0	0.50	0	1

Table 1: 87 observations from H treatment, consisting of 77+7+3 subjects who failed 0, 1, 2 monotonicity tests respectively. Trustor = 1 if the subject chooses to trust, 0 otherwise. Trustee = 1, 2 or 3 if the subject chooses option Reciprocate, Middle or Selfish respectively. Ambiguity aversion and a-insensitivity are indexes measuring motivational and cognitive component of ambiguity attitudes respectively. p_R, p_M, p_S are the ambiguity-neutral expectations on the probability that the trustee decision maker chooses Reciprocate, Middle or Selfish respectively. Male = 1 if the subject is male.

	Summary statistics - M treatment				
	Mean	Median	SD	Min	Max
Trustor	0.67	1	0.47	0	1
Trustee	1.91	2	0.93	1	3
Ambiguity aversion	-0.05	0	0.19	-0.66	0.86
a-insensitivity	0.23	0.10	0.27	-0.74	0.94
p_R	0.33	0.33	0.14	0	0.77
p_M	0.31	0.33	0.11	0	0.75
p_S	0.36	0.33	0.12	0.08	0.67
Male	0.46	0	0.50	0	1

Table 2: 93 observations from M treatment, consisting of 84+7+2 subjects who failed 0, 1, 2 monotonicity tests respectively. Trustor = 1 if the subject chooses to trust, 0 otherwise. Trustee = 1, 2 or 3 if the subject chooses option Reciprocate, Middle or Selfish respectively. Ambiguity aversion and a-insensitivity are indexes measuring motivational and cognitive component of ambiguity attitudes respectively. p_R, p_M, p_S are the ambiguity-neutral expectations on the probability that the trustee decision maker chooses Reciprocate, Middle or Selfish respectively. Male = 1 if the subject is male.

The analyzable sample for testing **H5a** and **H5b** consists of 169 observations. The sample size was reduced from 180 to 169 due to two technical issues: (1) the FaceReader software is unable to measure facially expressed emotions when participants partially or fully cover their faces, and (2) in one experimental session, technical malfunctions prevented recording of videos that could be perfectly synchronized with the experiment's timing. Summary statistics for this subsample are presented in Table 3.

	Summary statistics				
	Mean	Median	SD	Min	Max
$Trustor$	0.51	1	0.50	0	1
$Trustee$	1.86	2	0.90	1	3
$Ambiguity\ aversion$	-0.06	0	0.19	-0.66	0.86
$a-insensitivity$	0.23	0.16	0.27	-0.74	0.94
p_R	0.33	0.33	0.16	0	1
p_M	0.30	0.33	0.13	0	0.75
p_S	0.37	0.33	0.16	0	1
$Male$	0.44	0	0.50	0	1
$H\ treatment$	0.49	0	0.50	0	1
$PS\ Valence$	-0.19	-0.16	0.17	-0.68	0.47
$PS\ Happiness$	0.02	0.002	0.06	0.00	0.49
$PS\ Sadness$	0.18	0.14	0.16	0.002	0.68
$PS\ Anger$	0.05	0.02	0.09	0.00	0.63
$PS\ Scare$	0.01	0.003	0.02	0.00	0.20
$PS\ Arousal$	0.39	0.38	0.16	0.04	0.09
$TD\ Valence$	-0.17	-0.15	0.15	-0.68	0.35
$TD\ Happiness$	0.02	0.004	0.05	0.00	0.36
$TD\ Sadness$	0.15	0.10	0.14	0.003	0.68
$TD\ Anger$	0.07	0.03	0.10	0.00	0.51
$TD\ Scare$	0.01	0.004	0.02	0.00	0.15
$TD\ Arousal$	0.37	0.38	0.10	0.08	0.60

Table 3: 169 observations, consisting of 151+13+5 subjects who failed 0, 1, 2 monotonicity tests respectively. "PS" means pre-stimulus, indicating that the value of that specific emotion is equal to the average of the values manifested by the subject in the 10 seconds before starting the trustor decision (one observation every tenth of a second). "TD" means trustor decision, indicating that the value of that variable is equal to the average of the values manifested by the subject during the completion of the trustor decision (one observation every tenth of a second).

5 Results

In this section we present our findings with respect to the hypotheses presented in Section 3.[31]

5.1 Hypotheses 2-3-4

We first provide evidence in support of **H2-3-4**, which allows us to assert that any differences observed between the two treatments regarding the propensity to trust are ceteris paribus, i.e., holding ambiguity attitudes and ambiguity-neutral first-order beliefs constant. Using the Wilcoxon rank-sum test, we find that the treatment assignment has no significant effect on the ambiguity aversion index (-0.07 vs -0.05, W = 3646.5, p = 0.2453). This is also true for the a-insensitivity index (0.25 vs 0.23, W = 4304, p = 0.4539) and $p_R - p_S$ (-0.05 vs -0.03, W = 4078, p = 0.9247). Thus, the data robustly support the hypotheses **H2-3-4**. As a robustness check, we verify these findings using regression models with standard errors clustered by experimental sessions. Since the ambiguity aversion index, a-insensitivity index, and $p_R - p_S$ are continuous variables within the range (-1, 1), we employ Beta regression models with a logit link function. In these models, the dependent variables are transformed as $y' = \frac{y-a}{b-a}$, where $a = -1$ and $b = 1$ (Cribari-Neto and Zeileis, 2010). Furthermore, since $p_R - p_S$ takes values at the extreme points -1 and 1, we apply the further transformation $y'' = \frac{y'(n-1)+0.5}{n}$, where n is the sample size (180) (Smithson and Verkuilen, 2006).

[31] Data are available upon request.

	Beta Regression model		
	Model 1	Model 2	Model 3
Ambiguity aversion	Dependent variable	-0.29 (0.42)	0.64 (0.42)
a-insensitivity	-0.14 (0.16)	Dependent Variable	-0.08 (0.08)
$p_R - p_S$	0.22* (0.13)	-0.28** (0.11)	Dependent Variable
$Male$	0.03 (0.04)	-0.09 (0.12)	-0.00 (0.9)
H treatment	-0.02 (0.06)	-0.01 (0.10)	-0.09 (0.11)
$TrusteeR$	-0.05 (0.08)	0.08 (0.14)	0.30*** (0.09)
$TrusteeS$	0.02 (0.05)	-0.14 (0.11)	-0.36** (0.17)
const	-0.06 (0.06)	0.52**** (0.15)	0.00 (0.12)

Table 4: 180 observations, Beta Regression models with standard errors clustered by experimental session. Since the variables Ambiguity aversion and a-insensitivity are defined in the interval (-1, 1), those were transformed into $y' = (y - a)/(b - a)$ where $a = -1$ and $b = 1$. Since $p_R - p_S$ takes values in [-1, 1], this has been transformed into $y'' = (y'(n - 1) + 0.5)/n$ where $n = 180$ and $y' = (y - a)/(b - a)$ with $a = -1$ and $b = 1$. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$; $^{****}p < 0.001$

The results in Table 4 further confirm the hypotheses **H2-3-4**. Figures 3, 4, and 5 display the boxplots for the ambiguity aversion index, a-insensitivity index, and $p_R - p_S$, respectively, for the H and M treatments.

Result 0: We find no evidence to reject the null hypotheses **H2-3-4** that ambiguity aversion, a-insensitivity and $p_R - p_S$ are independent of the treatment.

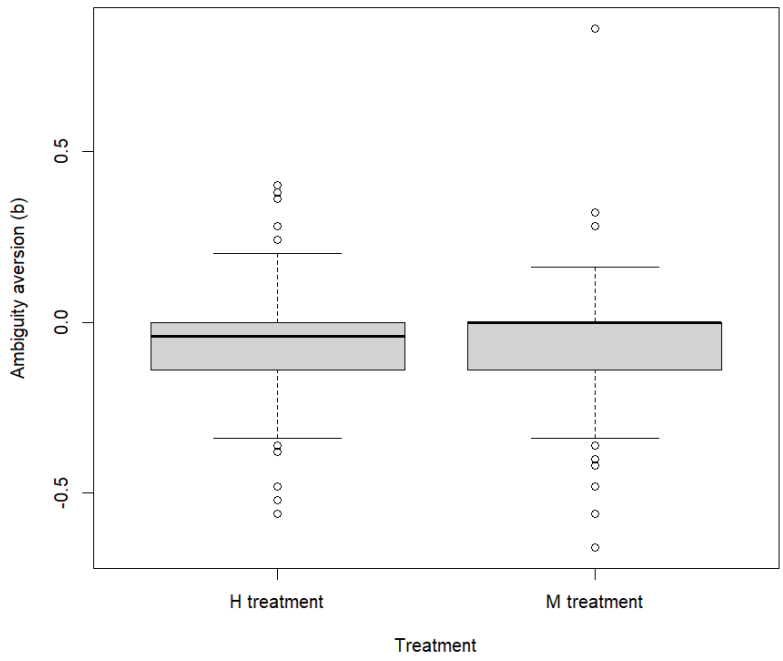

Fig. 3 Boxplots for the ambiguity aversion index (b) in H and M treatments.

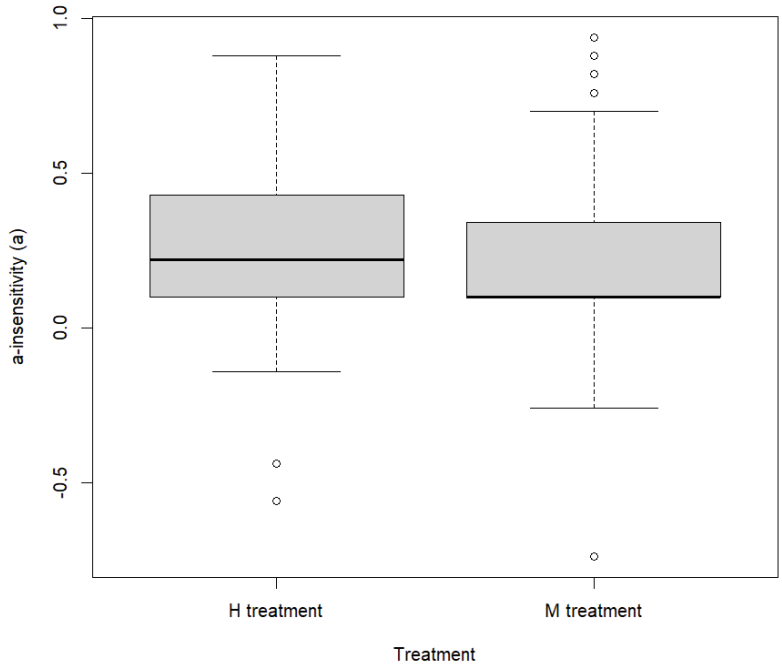

Fig. 4 Boxplots for the a-insensitivity index (a) in H and M treatments.

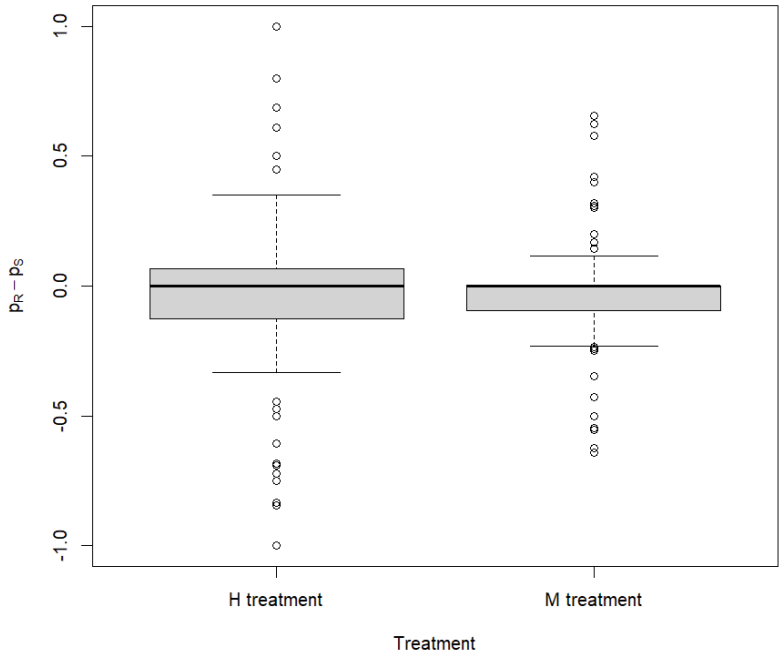

Fig. 5 Boxplots for $p_R - p_S$ in H and M treatments.

Moreover, subjects' trustee decisions are significantly correlated with their expectations regarding the counterpart's trustee decision. Specifically, the Spearman's rank correlations are as follows: for corr(*TrusteeR*, $p_R - p_S$), $\rho_S = 0.417$, p < 0.001; and for corr(*TrusteeS*, $p_R - p_S$), $\rho_S = -0.338$, p < 0.001. This evidence supports the findings of Li et al. (2019) on self-similar reasoning, suggesting that individuals form their expectations about the other's trustee decision based on their own decision, as illustrated in Fig. 6.

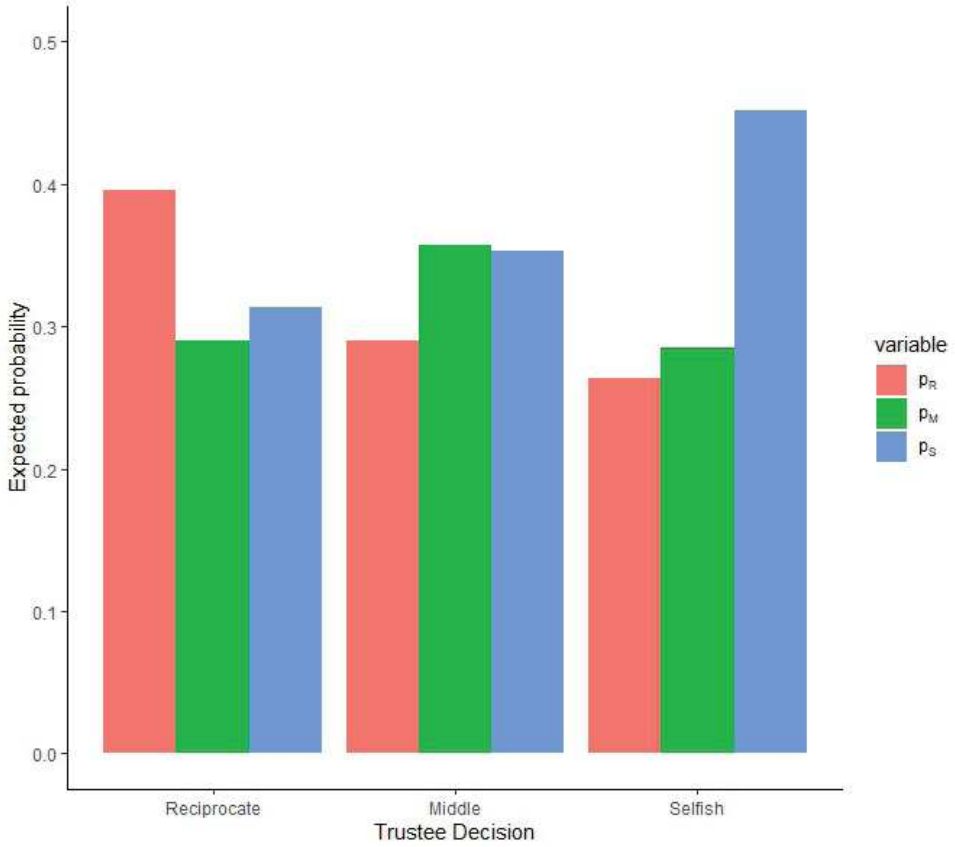

Fig. 6 Subjects' average ambiguity-neutral first-order beliefs about counterparts' trustee decision, given their own trustee decision.

We also find suggestive evidence of a positive correlation between $p_R - p_S$ and the ambiguity aversion index (p = 0.077), as well as a negative correlation between $p_R - p_S$ and the a-insensitivity index (p = 0.014).

5.2 Hypothesis 1

Having confirmed that subjects' ambiguity aversion, a-insensitivity, and ambiguity-neutral expectations are independent of the treatment assignment, we now proceed to analyze the effect of the treatment assignment on subjects' propensity to trust. Before doing so, we first analyze the trustee decision to ensure it is balanced between treatments and that there are no significant interactions with other regressors.

5.2.1 Trustee Decision analysis

We begin by briefly summarizing the evidence suggesting the presence of self-similar reasoning. Figure 6 shows that subjects' expectations regarding a counterparty's trustee decision (e.g., p_S) are higher when the subject made the same trustee decision (*Selfish*). The potential presence of self-similar reasoning is further supported by Model 3 in Table 4. However, it is important to note that this evidence is purely correlational and not causal.

As shown in Table 5, there is a significant correlation between $p_R - p_S$ and the trustee decision made by subjects; considering this evidence in conjunction with that found in Model 3 in Table 4 suggests that there could be a relationship of interdependence between subjects' trustee decision and their expectations on others' trustee decision.[32] Furthermore, in Table 5 we find other evidence that we need to account for.

First, the random assignment of subjects to the H and M treatments, which occurred after subjects made their trustee decisions and were informed that they would be assigned to one of the treatments with equal probability, did not result in perfect balance of the trustee decision covariate across treatments. Our analyzable sample included 87 subjects assigned to the H treatment (Reciprocate = 42, Middle = 20, Selfish = 25) and 93 subjects assigned to the M treatment (Reciprocate = 44, Middle = 13, Selfish = 36). Consequently, there is

[32]It's likely that a subject's behavior may influence her expectations on others, but also that how she thinks others will behave may influence what she choose to do. The latter may be due to adhering to social norms and/or to adapt her conduct towards others based on the conduct she expects them to have towards her. This second point will be analyzed in more depth in Section 6.2.

	Logit model - Trustee Decision	
	$TrusteeR$	$TrusteeS$
$Male$	-0.19	1.06****
	(0.35)	(0.29)
H *treatment*	0.06	-0.65**
	(0.30)	(0.32)
Ambiguity aversion	-0.96	1.14
	(1.06)	(1.00)
a-insensitivity	0.83	-1.58**
	(0.66)	(0.73)
$p_R - p_S$	4.44****	-4.77****
	(0.95)	(0.92)
const	-0.14	-0.82**
	(0.38)	(0.42)

Table 5: 180 observations, Logit models with standard errors clustered by experimental session.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$; $^{****}p < 0.001$

an imbalance between the two treatments in the percentage of subjects who chose Selfish, with 28.74% in the H treatment (28.70% in the full sample of 108 subjects assigned to the H treatment) and 38.71% in the M treatment (38.81% in the full sample of 108 subjects assigned to the M treatment).

Second, there is a notable gender difference in trustee decisions. Our analyzable sample includes 80 male subjects (Reciprocate = 35, Middle = 8, Selfish = 37) and 100 female subjects (Reciprocate = 51, Middle = 25, Selfish = 24). The difference is particularly pronounced with regard to the Selfish option: while only 24% of female subjects chose this option (23.97% in the full sample), 46.25% of male subjects selected it (47.37% in the full sample). A two-sided Fisher Exact test for a 3×2 contingency table reveals that these distributions are significantly different ($p < 0.01$). Consistent with the literature reviewed

by Croson and Gneezy (2009), but not with Van Den Akker et al. (2020), we find that men are significantly more likely to choose *Selfish* than women.[33]

Finally, we observe a significant difference in the level of a-insensitivity based on the option chosen for the trustee decision. Subjects who chose Selfish (61 subjects) have an average a-insensitivity level of 0.19 (median = 0.1), while those who chose Middle (33 subjects) have an average of 0.27 (median = 0.22), and those who chose Reciprocate (86 subjects) have an average of 0.26 (median = 0.22). In other words, subjects with higher discriminatory power between different levels of likelihood tend to choose the Selfish option more often, while those with lower discriminatory power are more likely to choose the Middle or Reciprocate options.

These findings highlight the importance of controlling for subjects' trustee decisions when analyzing their propensity to trust. Failing to do so would introduce omitted variable bias, potentially distorting the regression parameters for the gender dummy, the treatment dummy, the ambiguity-neutral first-order beliefs, and the a-insensitivity index.

5.2.2 Betrayal and ambiguity aversion

Based on the findings presented so far, we now proceed with analyzing the effect of treatment assignment on subjects' propensity to trust using the binary logit model in Table 6.

In addition to confirming the results found by Li et al. (2019) regarding the impact of ambiguity attitudes and a-neutral first-order beliefs on the propensity to trust,[34] Models 1 and 2 in Table 6 reveal a significant negative effect of the H treatment on subjects' propensity to trust (51.72% of subjects choose Trust in the H treatment versus 66.67% in the M treatment). This provides strong evidence in support of **H1**.

[33] This result remains statistically significant when considering the full sample of 216 subjects ($p < 0.001$).

[34] As Li et al. (2019) found, we observe that (I) an increase in ambiguity aversion is associated with a decrease in the propensity to trust, (II) an increase in $p_R - p_S$ is associated with an increase in the propensity to trust, and (III) that an increase in a-insensitivity is linked to a reduction in the propensity to act based on a-neutral first-order beliefs.

	Logit model - Trustor Decision	
	Model 1	Model 2
$Male$	1.33***	1.23****
	(0.42)	(0.35)
$H\ treatment$	-0.78**	-0.79**
	(0.40)	(0.35)
$Ambiguity\ aversion$	-2.40**	-2.16*
	(1.03)	(1.12)
$a\text{-}insensitivity$	-0.65	-0.75
	(0.93)	(0.85)
$p_R - p_S$	4.14***	4.09**
	(1.61)	(1.60)
$a\text{-}insensitivity \times (p_R - p_S)$	-6.86*	-6.94**
	(3.52)	(3.21)
$TrusteeR$	0.70	-0.41
	(0.66)	(0.62)
$TrusteeS$	-1.70****	-1.44****
	(0.49)	(0.40)
$const$	1.25*	1.07**
	(0.68)	(0.54)
$Log\ Likelihood$	-96.14	-100.81
AIC	210.28	219.61

Table 6: Logit models with standard errors clustered by experimental session. Model 1 is estimated based on the observations of subjects who failed less than 2 monotonicity tests (175 subjects), Model 2 is estimated based on the observations of subjects who failed at most 2 monotonicity tests (180 subjects).
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$; $^{****}p < 0.001$

Result 1: Consistent with **H1**, ceteris paribus, subjects are significantly less likely to trust a human trustee than a Machine selection.

This evidence remains robust when using the subsample of subjects selected by Li et al. (2019) (Model 1 from Table 6), which includes only subjects who failed fewer than two monotonicity tests and who satisfy the conditions $\overline{m}_C > \overline{m}_S$ and $m_{ij} + m_{ik} \geq m_j + m_k$ for $i, j, k = Reciprocate, Middle, Selfish$. The results are consistent consistent with the presence of betrayal aversion and with subjects associating a direct intention of betrayal towards them with the Selfish choice made by a human partner, and not (or less) with the selection of the Selfish option performed by a Machine.

We confirm the importance of ambiguity attitudes and ambiguity neutral beliefs, as found by Li et al. (2019). However, in contrast to Li et al. (2020), we also identify a significant treatment effect consistent with betrayal aversion. In our environment, both ambiguity attitudes and betrayal aversion appear to coexist and contribute to explaining the propensity to trust.

In addition to Result 1, two other noteworthy findings regarding the decision to trust emerge.

(I) There is a significant gender difference in the propensity to trust (more on this in Section 6.1).

(II) Ceteris paribus, subjects who chose the Selfish option for the trustee decision are less likely to trust.

Regarding finding (II), to address any concerns about the potential overlap between the variables $p_R - p_S$ and $TrusteeR(S)$ in conveying information about the propensity to trust, we estimated the Variance Inflation Factors (VIFs) for the parameters of Model 2 in Table 6.[35] As shown, there are no issues with multicollinearity. Although the two variables are

[35]The VIFs for the regressors included in the model are as follows: $Male = 1.17$, $M\ treatment = 1.07$, $Ambiguity\ aversion = 1.09$, $a\text{-}insensitivity = 1.17$, $p_R - p_S = 4.66$, $a\text{-}insensitivity \times (p_R - p_S) = 4.74$, $TrusteeR = 1.93$, $TrusteeS = 2.19$.

interdependent, they provide distinct information regarding the propensity to trust.

5.3 Hypotheses 5

To confirm that betrayal aversion is likely driving the lower trust observed in the H treatment, we analyzed the emotions experienced by subjects during the trustor decision.

The analyzable sample for testing **H5a** and **H5b** consists of 169 observations. Given the nature of the variables that measure the emotions experienced by the subjects—continuous in (0, 1) for individual emotions and continuous in (-1, 1) for Valence—we test **H5a** and **H5b** using Beta regression models, where the dependent variable is the average of the Valence or one of the emotions that contribute to determine the Valence, as experienced by the subject during the trustor decision. This analysis controls for the pre-stimulus emotional state at the individual level, as is commonly done in psychology research (Höfling et al., 2020).[36] This control is crucial because we lack information about the emotional state with which subjects enter the laboratory, as well as about the effects on their emotional state of what happened from the start of the experiment until just before the trustor decision. Failing to control for these factors would lead to biased estimates due to uncontrolled heterogeneity in individual emotionality.

As shown in Table 7, we find no difference in the Valence experienced by subjects during the trustor decision, regardless of the nature of the trustee decision maker.

Result 2a: The level of Valence manifested by subjects is not significantly different between treatments.

Regarding the components that contribute to determining Valence, we find no significant correlation between the H treatment and the expressed levels of happiness, sadness, or scare. However, consistent with the relevant literature, we observe a significant positive correlation

[36]The pre-stimulus emotional state is the average emotional state during the 10 seconds preceding the trustor decision, when subjects observed a waiting screen. Specifically, this is the average of each of the seven emotions and Arousal during this 10-second period.

	Beta Regression - Emotions during the Trustor Decision				
	Valence	Happiness	Sadness	Anger	Scare
Male	-0.03	-0.33*	-0.15	0.14	-0.49****
	(0.04)	(0.18)	(0.12)	(0.20)	(0.09)
H treatment	-0.05	-0.00	-0.09	0.30***	0.11
	(0.04)	(0.08)	(0.15)	(0.12)	(0.12)
TrusteeR	0.03	-0.26	-0.19	0.01	-0.12
	(0.08)	(0.19)	(0.16)	(0.22)	(0.21)
TrusteeS	-0.01	-0.17	0.17	0.01	0.05
	(0.08)	(0.22)	(0.23)	(0.27)	(0.23)
Ambiguity aversion	0.03	0.48**	-0.43*	0.08	0.39
	(0.12)	(0.22)	(0.22)	(0.36)	(0.33)
a-insensitivity	0.05	-0.19	-0.01	-0.30	0.47*
	(0.06)	(0.24)	(0.18)	(0.20)	(0.25)
$p_R - p_S$	-0.03	0.07	0.23**	-0.30	0.29**
	(0.08)	(0.24)	(0.09)	(0.22)	(0.14)
const	-0.07	-3.76****	-2.76****	-3.02****	-4.73****
	(0.12)	(0.39)	(0.36)	(0.20)	(0.35)
Pre-stimulus emotions	Yes	Yes	Yes	Yes	Yes

Table 7: 169 observations. Beta Regression models with standard errors clustered by experimental session. Since the variable Valence is defined in the interval (-1, 1), it was transformed into $y' = (y - a)/(b - a)$ where $a = -1$ and $b = 1$.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$; $^{****}p < 0.001$

($p = 0.009$) between the H treatment and the anger experienced during the trustor decision. Ceteris paribus, subjects appear to feel significantly more anger when the trustee decision maker is human, as compared to when it is Machine.

Result 2b: The level of sadness, scare and happiness manifested by subjects is not significantly different across treatments. Ceteris paribus, the subjects assigned to the H treatment manifested a significantly higher level of anger than subjects assigned to the M treatment.

This evidence aligns with the negative correlation between the H treatment and the propensity to trust. The data suggest the following interpretation: during the trustor decision, the subjects reflect on the possible scenarios to which the choice to trust could lead; in the H treatment, reflecting on the prospect in which the human trustee decision maker chooses the Selfish option, they feel angry at the idea that the latter could carry out the aforementioned action towards them, perceived as a betrayal, and are less inclined to trust. On the contrary, in the M treatment this prospect is not (or less) perceived as a betrayal, hence the subjects feel less anger and are more inclined to trust with no (or lower) non-monetary betrayal cost associated with the prospect of being subjected to the selection of the Selfish option by Machine.

As a further check, we verify if there is a treatment effect also on the remaining two emotions measured by Noldus' FaceReader, that are disgust and contempt. As shown in Table 8, we find evidence that in the H treatment, in addition to anger, subjects experienced a higher level of contempt. This is partially in line with previous evidence of an association between the hostility triad (anger, contempt and disgust) and perceived betrayal (e.g. Tuzovic et al., 2022).

	Beta Regression - Disgust and Contempt	
	Disgust	Contempt
$Male$	-0.17	0.03
	(0.12)	(0.06)
$H\ treatment$	0.07	0.19***
	(0.12)	(0.07)
$Ambiguity\ aversion$	0.09	-0.13
	(0.40)	(0.38)
$a\text{-}insensitivity$	-0.10	-0.20**
	(0.37)	(0.09)
$p_R - p_S$	-0.19	0.39*
	(0.33)	(0.21)
$TrusteeR$	0.10	-0.10
	(0.24)	(0.17)
$TrusteeS$	0.02	0.04
	(0.27)	(0.13)
$const$	-5.20****	-3.88****
	(0.40)	(0.27)
$Pre\text{-}stimulus\ emotions$	Yes	Yes

Table 8: 169 observations. Beta Regression models with standard errors clustered by experimental session.
$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$; $^{****}p < 0.001$

5.4 Hypotheses 6-7

Unfortunately, it is not possible to test **H6-7** due to sample size limitations. The number of subjects who chose to trust, were randomly selected for the role of trustor, and were exposed to the Selfish option is only 20 (15 in the H treatment and 5 in the M treatment).

6 Additional findings

In this section, we present additional findings that were not included among the pre-registered hypotheses but are believed to be of interest to the reader.

6.1 Gender differences in trustor and trustee behavior

As reported in Section 5.2.2, we find that men are more likely to trust than women (68.75% vs 52%). This finding is consistent with the literature surveyed by Croson and Gneezy (2009) and Van Den Akker et al. (2020). This difference is also confirmed by a two-tailed Fisher Exact test ($p = 0.03$) when analyzing the subsample for which a-neutral expectations regarding the other trustee's decision and ambiguity attitudes are estimable.[37] Here, we report two other gender differences.

First, we report the following evidence regarding the intensity of scare experienced by the subjects during the trustor decision: male subjects appear to have felt less scare than female subjects (Table 7, $p < 0.001$). This finding is consistent with the lower propensity to trust observed among female subjects compared to male subjects.

Second, we report a gender difference regarding the strategic nature of the trustor decision. By comparing the trustor decision with the expected value of the Trust choice,

[37] In contrast, when looking at the complete dataset (64 women out of 121 chose to trust, 56 men out of 95 chose to trust), this difference is not significant. However, the 36 additional subjects exhibited choices that were logically inconsistent, making it impossible to estimate their expectations and ambiguity attitudes. Given that ambiguity attitudes are a determinant of trust, we believe including these subjects in the analysis confounds the results due to their highly inconsistent or non-computable ambiguity attitudes and a-neutral expectations.

calculated as $\mathbb{E}[Trust] = p_R \times 15 + p_M \times 10 + p_S \times 8$, we classify a subject as strategic if she chooses Trust when $\mathbb{E}[Trust] \geq 10$ or Distrust when $\mathbb{E}[Trust] < 10$. Conversely, we classify a subject as non-strategic if she chooses Trust when $\mathbb{E}[Trust] < 10$ or Distrust when $\mathbb{E}[Trust] \geq 10$.[38] In the subsample of 180 subjects with correctly estimable a-neutral expectations about the other's trustee decision, we find that 57 women are classified as strategic (48 who chose Trust with $\mathbb{E}[Trust] \geq 10$ and 9 who chose Distrust with $\mathbb{E}[Trust] < 10$), while 43 women are classified as non-strategic (39 who chose Trust with $\mathbb{E}[Trust] < 10$ and 4 who chose Distrust with $\mathbb{E}[Trust] \geq 10$). Among men, 61 are classified as strategic (53 who chose Trust with $\mathbb{E}[Trust] \geq 10$ and 8 who chose Distrust with $\mathbb{E}[Trust] < 10$), and 19 are classified as non-strategic (17 who chose Trust with $\mathbb{E}[Trust] < 10$ and 2 who chose Distrust with $\mathbb{E}[Trust] \geq 10$). A two-tailed Fisher Exact test reveals a significant gender difference ($p < 0.01$): the proportion of men classified as strategic (76.25%) is significantly higher than the proportion of women classified as strategic (57%). This finding aligns with the results of Buchan et al. (2008) in an investment game: men trust more than women, women are more trustworthy than men, and the relationship between expected return and trusting behavior is stronger among men than women.

6.2 Beliefs' manipulation in order to morally justify the Selfish choice

Revisiting the evidence in Table 3 and Fig. 6, if there were no cognitive biases other than self-similar reasoning, we would expect $(p_R|TrusteeR) = (p_S|TrusteeS)$, meaning that subjects who choose Reciprocate and those who choose Selfish should believe that others are similar to them to the same extent. However, as shown in Fig. 6, a different pattern emerges: $(p_R|R) = 39.59\%$ and $(p_S|S) = 45.19\%$. This suggests that subjects who chose Selfish believe that others are more similar to them than those who chose Reciprocate. Assuming

[38] In the full subsample of 180 subjects with estimable a-neutral expectations, only one subject, a woman, has $\mathbb{E}[Trust] = 10$.

that most subjects in our sample can be classified as *homo moralis*—the only evolutionarily stable human category according to Alger and Weibull (2013)—it follows that the majority of them should exhibit a convex combination of selfish and moral preferences. As a result, even subjects who choose Selfish are likely to have some preference for morality and to derive disutility from believing they are engaging in an immoral act, such as betrayal. However, a selfish subject could maximize her economic gain while minimizing the non-monetary disutility associated with committing an immoral act by convincing herself that the other person, in her position, would choose the Selfish option as well. Thus, she would perceive herself not as a betrayer committing an immoral act, but as merely adhering to a social norm. A subject who chooses Reciprocate, on the other hand, would not gain from such belief manipulation.[39]

To verify this aspect, we build a $Self\text{-}similarity_i$ index defined in $[0, 1]$ capturing the distance perceived by a subject between her own and another trustee decision:

$$Self - similarity_i = 1 - \frac{p_j \cdot D_{ij} + p_k \cdot D_{ik}}{max\{D\}} \tag{4}$$

where $\{i, j, k\} = \{Reciprocate, Middle, Selfish\}$, $D_{ij} = \sqrt{(X_i - X_j)^2 + (Y_i - Y_j)^2}$ where (X_i, Y_i) is the payoff pair $(Payoff_{Trustor}, Payoff_{Trustee})$ resulting from trustee decision i, and $max\{D\}$ is the maximum possible distance given the available options for the trustee decision, i.e., $D_{RS} = \sqrt{(15-8)^2 + (15-22)^2}$. Based on a two-tailed Wilcoxon rank sum test (W = 1373, p < 0.001), we find that $Self\text{-}similarity_R$ (mean = 0.52, median = 0.47) is significantly smaller than $Self\text{-}similarity_S$ (mean = 0.61, median = 0.53). We therefore find evidence that subjects who chose $Selfish$ believe that others are significantly more similar to them than subjects who chose *Reciprocate*, consistent with the literature on motivated

[39]Gangadharan et al. (2024), through a Donation game, show that non-donors significantly underestimate the percentage of donors (when beliefs are elicited in a way that does not reduce the salience of self-image utility), while donors accurately estimate this percentage, regardless of the belief elicitation method. They argue that donors preserve their self-image by donating and thus have no incentive to distort their beliefs about others' behavior, while non-donors have an incentive to underestimate the percentage of donors in order to protect their self-image.

reasoning (Gino et al., 2016).

Finally, we conduct a test to assess the accuracy of the subjects' beliefs relative to their actual choices. Specifically, we compute the sum of squared errors (SSE) between the ambiguity-neutral expectations and the actual observed proportions (103, 39 and 74 subjects chose Reciprocate, Middle and Selfish respectively). The SSE is measured as $SSE = (p_R - p_R^{True})^2 + (p_M - p_M^{True})^2 + (p_S - p_S^{True})^2$. Based on a two-tailed Wilcoxon rank sum test (W = 4946.5, p $<$ 0.01), we find that the estimates from subjects who chose Reciprocate (mean = 0.07, median = 0.04) are significantly closer to the true proportions than those from subjects who chose Middle or Selfish (mean = 0.13, median = 0.05). These results are consistent with the findings of Gangadharan et al. (2024).

7 Conclusions

We study the determinants of the decision to trust using the modified Trust game of Li et al. (2019). We allow the respondent to be either a human, or a machine that statistically replicates human behavior. We ask whether a cost of being betrayed by another human, betrayal aversion, still emerges after controlling for ambiguity attitudes and sources of uncertainty.

Baillon et al. (2018) and Li et al. (2019, 2020) recently developed a methodology to measure individuals' ambiguity attitudes and demonstrated that both ambiguity attitudes and the source of uncertainty influence the decision to trust in ways that could be misinterpreted as a betrayal cost. This challenges the conclusions of previous studies that did not control for these factors and raises questions about the existence of betrayal aversion itself.

In this study, we apply the novel methodology to control for subjects' ambiguity attitudes, allowing the trustee to be either a human or a machine that selects a response stochastically based on the distribution of choices made by the population of human trustees. This second treatment aims to keep the source of uncertainty (human decisions) unchanged while reducing the trustor's perception of the selfish response as an intentional act of betrayal. Additionally,

we measure emotions during the trustor decision using the most recent methodologies.

We confirm that ambiguity attitudes are a crucial determinant of the decision to trust and must be controlled for. However, we also find that subjects trust significantly less when the trustee is a human rather than a machine, suggesting the presence of additional betrayal costs associated with interacting with a human.

Moreover, we find that when deciding whether to trust, subjects experience significantly more anger—an emotion that previous studies have linked to anticipated or experienced betrayal costs—when interacting with humans rather than with machines.

Both the lower propensity to trust and the increased experience of anger when interacting with humans are consistent with betrayal aversion being a significant determinant of trusting behavior, alongside ambiguity attitudes.

In addition to the main findings, this study uncovers several additional results. First, we find gender differences consistent with previous literature: males are more likely to trust than females, and, in line with this, they experience lower levels of scare during the trustor decision. Furthermore, females are more trustworthy than males, and males tend to make the trustor decision in a more strategic manner. Second, we find that selfish subjects, ceteris paribus, are less likely to trust than others and have a stronger belief that others are similar to them, which aligns with the literature on motivated reasoning and self-serving beliefs' manipulation.

Understanding the motivations behind individuals' trust decisions is a key research objective, as effective policies aimed at achieving specific goals (e.g., increasing vaccination uptake or boosting stock market participation) rely on these motivations. The identification of betrayal aversion suggests that simply reducing perceived ambiguity or improving expectations may not be enough. Policymakers must also account for the psychological costs associated with the perceived source of an outcome. Future research could explore alternative treatments that mitigate betrayal aversion while maintaining a human trustee, or investigate whether certain topics elicit stronger betrayal aversion than others.

A Appendix A

In this Appendix, starting from the results derived and reported in Appendix C of Li et al. (2019), we present the derivation of a condition that relaxes $m_{ij} \geq m_i$ for all $\{i, j, k\} = \{Reciprocate, Middle, Selfish\}$—which would imply the exclusion from the analyzable sample of subjects characterized by correctly estimated and analyzable values of a, p_R, p_M and p_S—that is necessary and sufficient to guarantee the inclusion in the analyzable sample of all subjects with p_R, p_M and p_S ranging from 0 to 1 and which sum to 1. After a series of algebraic manipulations, Li et al. (2019) define the equations OC.5, OC.6, and OC.7, which can be generalized and consolidated into the equation (3). Additionally, they demonstrate that $1 - a = 3(\overline{m}_C - \overline{m}_S)$, and substitute this term into equation (3), resulting in the following expression:

$$p_i = \frac{m_{ij} + m_{ik} - m_j - m_k}{2(\overline{m}_C - \overline{m}_S)} \tag{5}$$

They ensure that $p_i > 0$ through a priori assumptions. Specifically, $\overline{m}_C > \overline{m}_S$ guarantees that the denominator is positive, while $m_{ij} \geq m_i$ ensures that the numerator is nonnegative. The assumption $\overline{m}_C > \overline{m}_S$ is strictly necessary to also guarantee that $a < 1$, which is a required condition for the a-insensitivity index to yield meaningful values. On the other hand, it is straightforward to verify that the assumption $m_{ij} \geq m_i$ for $\{i, j\} = \{R, M, S\}$ is sufficient, though not necessary, to ensure that the numerator remains nonnegative. By examining equation (5), it becomes clear that violating $m_{ij} \geq m_j$ does not necessarily result in a negative numerator, as Li et al. (2019) have empirically shown by including subjects who failed a monotonicity test in the analyzable sample.

The solution we propose and adopt is to avoid assuming an excessively restrictive condition, such as $m_{ij} \geq m_i$ for $\{i, j\} = \{R, M, S\}$, a priori. Instead, we work downstream of the problem, assuming only the minimum conditions necessary to guarantee the nonnegativity

of the numerator, namely:

$$m_{ij} + m_{ik} \geq m_j + m_k \; for \; \{i, j, k\} = \{R, M, S\} \tag{6}$$

Upon further reflection on this condition, it becomes apparent that it is not merely a mathematical constraint to ensure a nonnegative numerator. It also reflects a requirement for the minimum logical consistency of the matching probabilities declared by the subjects. Since the events $\{R, M, S\}$ are mutually exclusive, it follows that m_{ij} can be interpreted as $m_{ij} = \tilde{m}_i + \tilde{m}_j$ and m_{ik} as $m_{ik} = \tilde{m}_i + \tilde{m}_k$. These expressions can then be substituted into equation (6):

$$2\tilde{m}_i + \tilde{m}_j + \tilde{m}_k \geq m_j + m_k \; for \; \{i, j, k\} = \{R, M, S\} \tag{7}$$

This last equation has a clear interpretation in terms of the minimum logical consistency of the matching probabilities, both directly and indirectly declared by a subject. Specifically, the sum of the matching probabilities for the single events j ($\tilde{m}_j$) and k ($\tilde{m}_j$), which are indirectly declared through the matching probabilities of event compositions m_{ij} and m_{ik}, plus the term $2\tilde{m}_i$, representing the matching probability of event i as indirectly declared through m_{ij} and m_{ik}, must be no less than the sum of the matching probabilities directly declared for events j and k.

B Appendix B

To assess whether it is likely that prosocial motivations influenced the difference in the trustor decision between the two treatments, we examine whether there are differences in the strategic nature of subjects' choices to trust across the treatments. This analysis categorizes subjects based on their expectations of the potential gain from choosing to trust, as detailed in Section 6.1. The key intuition is that if the M treatment had induced subjects to trust due to prosocial motivations, this would have occurred regardless of their expectations about

$\mathbb{E}[Trust]$. However, our analysis reveals no evidence supporting this hypothesis. Specifically, among subjects with $\mathbb{E}[Trust] \geq 10$, 41 out of 73 (56.16%) chose to trust in the H treatment, and 64 out of 84 (71.43%) chose to trust in the M treatment. Conversely, among subjects with $\mathbb{E}[Trust] < 10$, 4 out of 14 (28.57%) chose to trust in the H treatment, while 2 out of 9 (22.22%) chose to trust in the M treatment.

Acknowledgements

Funding: this project has been funded by European Union - Next Generation EU, UPB SpagnoloG22 Prin CUP: E53D23006140006.

References

Abdellaoui, M., Baillon, A., Placido, L., and Wakker, P. P. (2011). The rich domain of uncertainty: Source functions and their experimental implementation. *American Economic Review*, 101(2):695–723.

Aimone, J., Ball, S., and King-Casas, B. (2015). The betrayal aversion elicitation task: An individual level betrayal aversion measure. *PloS one*, 10(9):e0137491.

Aimone, J. A. and Houser, D. (2012). What you don't know won't hurt you: a laboratory analysis of betrayal aversion. *Experimental Economics*, 15:571–588.

Aimone, J. A., Houser, D., and Weber, B. (2014). Neural signatures of betrayal aversion: an fmri study of trust. *Proceedings of the Royal Society B: Biological Sciences*, 281(1782):20132127.

Algan, Y. and Cahuc, P. (2010). Inherited trust and growth. *American Economic Review*, 100(5):2060–2092.

Algan, Y., Guriev, S., Papaioannou, E., and Passari, E. (2017). The european trust crisis and the rise of populism. *Brookings papers on economic activity*, 2017(2):309–400.

Alger, I. and Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302.

Alsharawy, A., Dwibedi, E., Aimone, J., and Ball, S. (2022). Vaccine hesitancy and betrayal aversion. *Annals of Biomedical Engineering*, 50(7):794–804.

Arrow, K. J. (1972). Gifts and exchanges. *Philosophy & Public Affairs*, pages 343–362.

Baillon, A., Huang, Z., Selim, A., and Wakker, P. P. (2018). Measuring ambiguity attitudes for all (natural) events. *Econometrica*, 86(5):1839–1858.

Baron, J. (1992). The effect of normative beliefs on anticipated emotions. *Journal of personality and social psychology*, 63(2):320.

Battigalli, P. and Generoso, N. (2024). Information flows and memory in games. *Games and Economic Behavior*, 145:356–376.

Benndorf, V., Müller, S., and Rau, H. A. (2024). The effects of betrayal aversion on effort provision when incentives are fragile. *Management Science*.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142.

Bloom, N., Sadun, R., and Reenen, J. V. (2012). Americans do it better: Us multinationals and the productivity miracle. *American Economic Review*, 102(1):167–201.

Bohnet, I., Greig, F., Herrmann, B., and Zeckhauser, R. (2008). Betrayal aversion: Evidence from brazil, china, oman, switzerland, turkey, and the united states. *American Economic Review*, 98(1):294–310.

Bohnet, I. and Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, 55(4):467–484.

Breaban, A. and Noussair, C. N. (2018). Emotional state and market behavior. *Review of Finance*, 22(1):279–309.

Buchan, N. R., Croson, R. T., and Solnick, S. (2008). Trust and gender: An examination of behavior and beliefs in the investment game. *Journal of Economic Behavior & Organization*, 68(3-4):466–476.

Butler, J., Giuliano, P., and Guiso, L. (2016). Trust and cheating. *The Economic Journal*, 126(595):1703–1738.

Butler, J. V. and Miller, J. B. (2018). Social risk and the dimensionality of intentions. *Management Science*, 64(6):2787–2796.

Capra, C. M. (2004). Mood-driven behavior in strategic interactions. *American Economic Review*, 94(2):367–372.

Cribari-Neto, F. and Zeileis, A. (2010). Beta regression in r. *Journal of statistical software*, 34:1–24.

Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic literature*, 47(2):448–474.

Cubitt, R., Gächter, S., and Quercia, S. (2017). Conditional cooperation and betrayal aversion. *Journal of Economic Behavior & Organization*, 141:110–121.

Delle Foglie, L. and Papa, S. (2024). Cognitive and mnemonic abilities in a trust game. *Economics Letters*, page 111810.

Diecidue, E., Gueciloueur, A., and Xia, Q. (2024). Trusting the algorithm: A decision under ambiguity. *Available at SSRN*.

Doyle, L. and Schindler, D. (2019). μ cap: connecting facereader™ to z-tree. *Journal of the Economic Science Association*, 5(1):136–141.

Ekman, P. and Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Engelmann, J. B., Meyer, F., Ruff, C. C., and Fehr, E. (2019). The neural circuitry of affect-induced distortions of trust. *Science advances*, 5(3):eaau3413.

Evans, A. M. and Krueger, J. I. (2017). Ambiguity and expectation-neglect in dilemmas of interpersonal trust. *Judgment and Decision Making*, 12(6):584–595.

Farjam, M. (2019). On whom would i want to depend; humans or computers? *Journal of Economic Psychology*, 72:219–228.

Farolfi, F., Chang, L.-A., and Engelmann, J. B. (2021). *Trust and Emotion: The Effects of Incidental and Integral Affect*, page 124–154. Cambridge University Press.

Fehr, E. (2009). On the economics and biology of trust. *Journal of the european economic association*, 7(2-3):235–266.

Fetchenhauer, D. and Dunning, D. (2012). Betrayal aversion versus principled trustfulness—how to explain risk avoidance and risky choices in trust games. *Journal of Economic Behavior & Organization*, 81(2):534–541.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Gangadharan, L., Grossman, P. J., and Xue, N. (2024). Belief elicitation under competing motivations: Does it matter how you ask? *European Economic Review*, 169:104830.

Gershoff, A. D. and Koehler, J. J. (2011). Safety first? the role of emotion in safety product betrayal aversion. *Journal of Consumer Research*, 38(1):140–150.

Giannetti, M. and Wang, T. Y. (2016). Corporate scandals and household stock market participation. *The Journal of Finance*, 71(6):2591–2636.

Gino, F., Norton, M. I., and Weber, R. A. (2016). Motivated bayesians: Feeling moral while acting egoistically. *Journal of Economic Perspectives*, 30(3):189–212.

Guiso, L., Sapienza, P., and Zingales, L. (2004). The role of social capital in financial development. *American economic review*, 94(3):526–556.

Guiso, L., Sapienza, P., and Zingales, L. (2008). Trusting the stock market. *the Journal of Finance*, 63(6):2557–2600.

Höfling, T. T. A., Gerdes, A. B., Föhl, U., and Alpers, G. W. (2020). Read my face: automatic facial coding versus psychophysiological indicators of emotional valence and arousal. *Frontiers in psychology*, 11:1388.

Houser, D., Schunk, D., and Winter, J. (2010). Distinguishing trust from risk: An anatomy of the investment game. *Journal of economic behavior & organization*, 74(1-2):72–81.

Humphrey, S. J. and Mondorf, S. (2021). Testing the causes of betrayal aversion. *Economics Letters*, 198:109663.

Knack, S. and Keefer, P. (1997). Does social capital have an economic payoff? a cross-country investigation. *The Quarterly journal of economics*, 112(4):1251–1288.

Koehler, J. J. and Gershoff, A. D. (2003). Betrayal aversion: When agents of protection become agents of harm. *Organizational Behavior and Human Decision Processes*, 90(2):244–261.

Kugler, T., Ye, B., Motro, D., and Noussair, C. N. (2020). On trust and disgust: Evidence from face reading and virtual reality. *Social Psychological and Personality Science*, 11(3):317–325.

La Porta, R., Lopez-de Silanes, F., Shleifer, A., and Vishny, R. W. (1996). Trust in large organizations.

Li, C., Turmunkh, U., and Wakker, P. P. (2019). Trust as a decision under ambiguity. *Experimental Economics*, 22:51–75.

Li, C., Turmunkh, U., and Wakker, P. P. (2020). Social and strategic ambiguity versus betrayal aversion. *Games and Economic Behavior*, 123:272–287.

Schlösser, T., Dunning, D., and Fetchenhauer, D. (2013). What a feeling: the role of immediate and anticipated emotions in risky decisions. *Journal of Behavioral Decision Making*, 26(1):13–30.

Schniter, E., Shields, T. W., and Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology*, 78:102253.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54.

Tuzovic, S., Mulcahy, R., and Russell-Bennett, R. (2022). A hostile tale of disclosure and betrayal: Business perceptions of offshoring services. *Industrial Marketing Management*, 102:74–88.

Van Den Akker, O. R., van Assen, M. A., Van Vugt, M., and Wicherts, J. M. (2020). Sex differences in trust and trustworthiness: A meta-analysis of the trust game and the gift-exchange game. *Journal of Economic Psychology*, 81:102329.

RECENT PUBLICATIONS BY *CEIS Tor Vergata*

Firms' Legality and Efficiency: Evidence from Public Procurement
Elisabetta, Chiara Latour
CEIS *Research Paper*, 592 February 2025

Green Ambiguity
Marco Carli
CEIS *Research Paper*, 591 February 2025

Who Decides Matters: Female Representation and Academic Career Advancement
Marianna Brunetti, Annalisa Fabretti and Mariangela Zoli
CEIS *Research Paper*, 590 December 2024

Multivariate Rough Volatility
Ranieri Dugo, Giacomo Giorgio and Paolo Pigato
CEIS *Research Paper*, 589 December 2024

With a Little Help from Nurseries. Childcare Services and Mothers' Employment in Italy
Chiara Puccioni and Daniela Vuri
CEIS *Research Paper*, 588 December 2024

SMEs Performance in Public Procurement and the Italian Legality Rating
Andrea Fazio, Erminia Florio and Gustavo Piga
CEIS *Research Paper*, 587 December 2024

A Reinforcement Learning Algorithm For Option Hedging
Federico Giorgi, Stefano Herzel and Paolo Pigato
CEIS *Research Paper*, 586 December 2024

The Long-Run Effects of R&D Subsidies on High-Tech Start-Ups: Insights From Italy
Christoph Koenig, Letizia Borgomeo and Martina Miotto
CEIS *Research Paper*, 585 October 2024

With a Little Help From the Crowd: Estimating Election Fraud with Forensic Methods
Christoph Koenig
CEIS *Research Paper*, 584 October 2024

Biases and Nudges in the Circular Economy: A Review
Luca Congiu, Enrico Botta and Mariangela Zoli
CEIS *Research Paper*, 583 October 2024

DISTRIBUTION

Our publications are available online at www.ceistorvergata.it

DISCLAIMER

The opinions expressed in these publications are the authors' alone and therefore do not necessarily reflect the opinions of the supporters, staff, or boards of CEIS Tor Vergata.

COPYRIGHT

MEDIA INQUIRIES AND INFORMATION

For media inquiries, please contact Barbara Piazzi at +39 06 72595652/01 or by e-mail at piazzi@ceis.uniroma2.it. Our web site, www.ceistorvergata.it, contains more information about Center's events, publications, and staff.

DEVELOPMENT AND SUPPORT

For information about contributing to CEIS Tor Vergata, please contact at +39 06 72595601 or by e-mail at segr.ceis@economia.uniroma2.it