Fairness and Trust in Human Bot Interaction—Investigating the Consequence of Norm Consensus on Trust Allocation

Kinga Makovi*, Anahit Sargsyan*, Talal Rahwan*
*New York University Abu Dhabi

July 14, 2020

1 Introduction

As machines become increasingly intelligent, many roles that have been traditionally taken by people are now supplemented or augmented by machines and algorithms. Such roles require various levels of interaction between people and machines/algorithms that make decisions, thus research into human-machine interaction is timelier than ever. Recently, research on this topic has adopted methods from behavioral economics and behavioral psychology to study how people react and interact with bots by employing incentivized games. This interdisciplinary approach is gradually recognized and increasingly adopted by the academy and industry to understand the underlying mechanisms that drive differences from platform to platform, and to forecast each platform's performance.

We build on recent work by Jordan and colleagues to investigate the behaviors of fairness, punishment and trust when people are interacting with bots [1]. We carried out an online experiment that includes a two stage game following [1]; a third-party-punishment game followed by a trust game with approximately 3500 participants who interacted with one another and bots that made pre-determined decisions. We are building on this prior work in this study where we now investigate the level of consensus around behaviors, and their relationship to the trust one gains when behaving according to the consensus view. In short, we have two main expectations (1) the higher the level of consensus about a behavior, the higher the level of trust; and (2) the differences in trust gain can partially or fully account for the differences between trust gains across experimental conditions that manipulate the signalled identity of interaction partners.

2 Experimental protocols

In our experiment, we denote the Helper, Receiver, Punisher, and Sender in the third-party-punishment game and the trust game as Player 1, Player 2, Player 3, and Player 4, respectively. Each participant entering the experiment takes part in one of **seven** experimental conditions. The detailed rules for both stages are explained in the pre-analysis plan we submitted for the first experiment (gated until 10/01/20), and can be found in [1] as well. In this experiment the instructions are identical to those in the previous study we conducted, and were only modified for grammar to make it clear to participants that they are not making decisions in the two-stage game in this study, rather, answer questions about how others behaved previously.

Common to all conditions is the first stage (the third-party-punishment game) of the game that all participants read the rules of, and answer 4 comprehension check questions about it. Those who fail to answer at least 3 of the 4 questions correctly, are given a second chance to answer the questions, while being able to consult the instructions. Those who, after the second attempt, do not answer at least 3 of the 4 questions correctly are unable to earn their bonus in the experiment, but, they are able to finish the study.

Experimental conditions differ in two ways. First, they differ in the type (either person or bot) of the participants in the first stage; the player type is signalled by illustrations of people or bots representing algorithms that make decisions. Second, they differ in the identity of the player who is matched to another person in the second stage, i.e., who participates in the trust game. This could be either Player 1 (the Helper), or Player 3 (the Punisher). In other words,

Condition	Player 1	Player 2	Player 3	Pair of Player 4
1	Person	Person	Person	Player 1
2	Bot	Person	Person	Player 1
3	Person	Bot	Person	Player 1
4	Person	Person	Person	Player 3
5	Bot	Person	Person	Player 3
6	Person	Bot	Person	Player 3
7	Person	Person	Bot	Player 3

Table 1: Experimental conditions.

participants, depending on who plays the trust game, read instructions about either Player 1 or Player 3 participating in the trust game with Player 4 (who is always a person, never a bot). In each of these conditions, participants are told, Player 4 receives information about the other player's sharing or punishment behavior in the first stage. Table 2 summarizes the seven conditions.

Common to conditions 1–3 are the instructions participants read about the second stage in which Player 1 is paired with Player 4 in the trust game. Common to conditions 4–7 are the instructions participants read about the second stage in which Player 3 is paired with Player 4 in the trust game. Common to all seven conditions is that all participants read the rules of the trust game, and answer 4 comprehension check questions. Those who fail to answer at least 3 of the 4 questions correctly are given a second chance to answer the questions, while being able to consult the instructions. Those who, after the second attempt, do not answer at least 3 of the 4 questions correctly are unable to earn their bonus in the experiment, but they are able to finish the study.

After the comprehension check questions, all participants are asked to answer a series of questions about a particular treatment listed in Table 2. When they answer these questions, they see a visual summary of the treatment they are assigned to, which visually communicates the identity of participants, whether they are bots or people. Specifically, we ask participants:

1. to make guesses about the behavior of Player 1 (in conditions 1–3) or Player 3 (in conditions 4–7);
2. to tell us what Player 1 (in conditions 1–3) or Player 3 (in conditions 4–7) should do in their opinion;
3. to make guesses about how other MTurk workers responded to the previous question, i.e., what Player 1 (in conditions 1–3) or Player 3 (in conditions 4–7) should do.

For example, MTurk workers who are participating in the condition where Player 1 is a person and is paired with Player 4 in the trust game (this applies to conditions 1 and 3), will see the following on their screen: "*You will now make a guess about how other MTurk workers participating in this study answered the previous question: "What do you think, should the person in the role of Player 1 share?" We have asked many MTurk workers, and therefore, we know how they answered to this question.*" When they make their decision, they see: "*If your answer is correct, you will earn $0.75 as a bonus, in addition to your show up fee of $2.00. If you have answered at least 3 questions of the 4 from the comprehension questions in both blocks you will be able to earn your bonus.*" As exemplified here, the guesses participants make are incentivized to encourage thoughtful reporting. To continue the example, we use the answer options of: (1) 0 in 10 MTurk workers said that the person in the role of Player 1 should share (0%); (2) 1 in 10 MTurk workers said that the person in the role of Player 1 should share (10%); ...; (11) 10 in 10 MTurk workers said that the person in the role of Player 1 should share (100%). Obviously, the question wording is updated to reflect the treatment, i.e., if we are asking about MTurk workers in the role of Player 1 or Player 3, and the respective behavior they engage in, i.e., sharing and punishment.

In other words, for each MTurk worker we have a measure of their empirical expectations, and the level of consensus they expect about a specific behavior; as well as what they believe is normative in the eyes of the average person; and their personal views about what one should do.

After participants made their guesses, they answer a number of follow-up questions, including a question that asks them which player(s) in their specific treatment was a bot (to gauge their attention); and questions about their demographic background. Most importantly, they are then invited to make hypothetical decisions in the trust game itself, as Player 4. We follow [1], and our prior study referenced earlier, and use the strategy method, i.e., for each respondent we have a measure of the trust gain for Player 1 (in conditions 1–3) when she switches from sending over not sending; and Player 3 (in conditions 4–7) when she switches from punishing over not punishing.

3 Recruitment

We will carry out an online survey experiment with a convenience sample of approximately 2100 adults (for each condition, approximately 300 participants) from the Unites States. Participants will be recruited through Amazon Mechanical Turk (MTurk) by using the services of TurkPrime. Participants will only be allowed to enter the experiment once, and we will not allow people into the experiment who participated in the original study we are building upon. Specifically, we implement screening procedures by selecting the "ballot box stuffing" option on Qualtrics, grouping HITs in TurkPrime, excluding the MTurk IDs of previous participants. Additionally, we keep track of worker IDs of participants who completed the study; if they attempted to enter the study again, they will be directed to an end-of-survey message.

4 Hypotheses

We will estimate the following regressions:

$$Y_i = \alpha + \beta_1 X_{1i} + \varepsilon, \tag{1}$$

$$Y_i = \alpha + \beta_2 X_{2i} + \varepsilon, \tag{2}$$

where Y_i is the trust-gain for sending (in conditions 1–3) or punishing (in conditions 4–7) over not-sending and not-punishing, respectively, measured as the difference of self-reported trust in the hypothetical decisions of participants in the role of Player 4; X_{1i} is the level of consensus in one's empirical expectations; and X_{2i} in the level of consensus about normative expectations.

We anticipate that β_1 and β_2 are positive, i.e., the higher the level of consensus in behavior and norms, the larger the trust gain when engaging in the majority behavior, and, when engaging in the normative behavior. We also anticipate that these results hold when including a vector of individual-level covariates in the regression (specifically: gender, race, income, and region). We further anticipate that the results hold when controls are included for experimental condition; i.e., within each condition, the same results obtain. We also anticipate that the results hold (1) when restricting the sample to only those who had full comprehension, and answered all 8 of the comprehension check questions; (2) when restricting the sample to only those who answered the attention check questions correctly.

We will conduct exploratory analysis to learn which regression (empirical or normative expectations) produces a superior model fit. We will also explore how a typology of participants based on high/low consensus on empirical expectations, and high/low normative consensus drive trust gain. While we anticipate that X_1 and X_2 will be highly correlated, we will reestimate the regression including both covariates to see if nonlinearities occur, and if the interaction of X_1 and X_2 improves model-fit. Note that these analyses would only be possible if no multicollinearity issues occur.

References

[1] Jillian J. Jordan, Moshe Hoffman, Paul Bloom, and David G. Rand. Third-party punishment as a costly signal of trustworthiness. *Nature*, 530:473–476, 2016.