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

Kinga Makovi*, Anahit Sargsyan*, Talal Rahwan*
*New York University Abu Dhabi
August 3, 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 [1] to investigate the behaviors of fairness, punishment and trust when people are interacting with bots. 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/or with bots that made pre-determined decisions over the summer of 2019. We are building

on this prior work in this study where we now investigate the impact of manipulating norm-consensus (in the form of injunctive norms, i.e., what one should do), and their relationship to the trust one gains when being seen as behaving according to the consensus. We are conducting a within-person design where we invite participants from the first experiment conducted a year ago, and give them norm-consensus information. We expect that the information-treatment we administer will increase the trust-gain gap between those who behave according to the consensus-behavior compared to those who do not behave in such a manner as compared to the setting that participants made decisions a year ago, receiving no such information. Out of the 3 experimental conditions, we anticipate such effects where prior measurement shows lower levels of believed consensus. In the third condition, where we give participants truthful information on the basis of prior measurement we do not have clear expectations.

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 **three** 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 by adding the following sentence: "Recently we conducted a study with other MTurk workers who told us what they believe Player 1 should do in Stage 1 of this activity. **An overwhelming majority (93%)** of MTurk workers told us that Player 1 (BOT/Mturk worker) **should share** with Player 2 (BOT/MTurk worker).". In this experiment all participants are in the role of Player 4, and they all are matched with Player 1, and the manipulation reflected the treatment they were assigned to.

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

Table 1: Experimental conditions.

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 one way. 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. 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 1 summarizes the three conditions.

Common to all three conditions is that all participants, then, 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 all participants have finished the sections with the comprehension check questions, they are told about the identity of Player 1, Player 2, and Player 3, and the fact that they have been assigned to the role of Player 4 (treatment assignment). Then, they make their decision in the trust game using the strategy method. Specifically, they select how much, if any portion of

their allowance of 100 cents to send to Player 1, which we triple, and allow Player 1 to send all, some or none back. As in the original experiment by Jordan et al. [1], we use the strategy method and ask each respondent how much would they send if Player 1 shared, as well as when she did not share.

After participants have made their decisions, 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. We also ask about participants' recall abut the original study, i.e., how similar they found this experiment to the original study, allowing them to choose "*I don't know as I do not remember the details of the previous study.*" We also ask participants if they found the information we shared with them (1) surprising, and (2) true.

3 Recruitment

We will carry out an online survey experiment with a convenience sample of approximately 900 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. Specifically, we implement screening procedures by selecting the "ballot box stuffing" option on Qualtrics, grouping HITs in TurkPrime. 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. Most importantly we use a within-person design. Specifically, we are inviting the MTurk workers who have participated in the original experiment one year ago. We incentivize their participation by (1) increasing the show-up payment to $2.50, (2) emphasizing the additional bonus they may earn up to $3.00, and (3) distributing a lottery of $50.00 to five randomly selected participants.

4 Hypotheses

We have two main hypotheses.

> H1: The trust-gain of Player 1 when sharing over not sharing will be larger in condition 2 with the information-manipulation (calculated from data collected in the present experiment) compared to their trust-gain without the information manipulation (calculated from data collected in the past experiment conducted over the summer of 2019).
>
> H2: The trust-gain of Player 1 when sharing over not sharing will be larger in condition 3 with the information-manipulation (calculated from data collected in the present experiment) compared to their trust-gain without the information manipulation (calculated from data collected in the past experiment conducted over the summer of 2019).

We will calculate these differences using a within-person design.

We have two secondary hypothesis: the trust gain of Player 1 in condition 1 will no longer be statistically significantly different from that in condition 2 (H3a) and condition 3 (H3b).

As exploratory analysis we will also analyze the difference in trust gain difference between the original experiment and the one presented here, which we denote as TG_d. Specifically, we will run the following regression:

$$TG_d = b_0 + \mathbf{bX} + \varepsilon, \tag{1}$$

where $\mathbf{X}$ is a matrix of the following variables: gender, age-group, race (majority/minority), income (binned), education, believed the treatment (binary), level of surprise about the treatment. Importantly, we asked respondents how they feel about both bots and people making

money in the activity, as well as their general feelings about bots and people. For the treatments that involved a bot (conditions 2 and 3) we will re-estimate the model in equation (1) and add these variables to the regression.

To assess the robustness of the results we will replicate the main analyses for (1) people who had full comprehension; (2) for people who correctly recalled the treatment, i.e., who was a bot in their condition; (3) people who have not remembered the details of the original study, or who had remembered it incorrectly and claimed that the study was "completely different." We will also assess, especially on the basis of the exploratory analysis how big of a role sample attrition plays in our estimates.

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.