Human vs. Robot Trust Levels in the Split or Steal Game

André Silva[1*†], Pedro Matono[1*†], Rodrigo Antunes[1*†] and Viktor Vasylkovskyi[1*†]

[1*]Instituto Superior Técnico, Universidade de Lisboa, Portugal.

*Corresponding author(s). E-mail(s):
andre.a.n.silva@tecnico.ulisboa.pt;
pedro.rovisco.matono@tecnico.ulisboa.pt;
rodrigojbantunes@tecnico.ulisboa.pt;
viktor.vasylkovskyi@tecnico.ulisboa.pt;
†These authors contributed equally to this work.

Abstract

Social robots have a high potential for helping humans in their daily life. However, human parties must deposit, to some extent, trust in them. In this paper, we investigate how trust levels in humans and robots vary in function of their truthfulness. To this end, we setup an experiment around the Split or Steal game and conduct it with 21 participants. Human and robotic actors, who could lie or be truthful, were assessed based on their truthfulness. Our results suggest that robots have overall lower trust levels than their human counterparts, but that lying places them at the same lower level.

Keywords: social robots, social dilemma, trust, lying

1 Introduction

Robots capable of interacting with humans as teammates or partners can help people perform some operations and achieve the desired goals. Such interaction implies that both parties - human and robotic - will expect the other to perform an expected task with the desired certainty. *Such certainty* is a

trust that parties place in each other. Trust is an essential element to consider because the presence or absence of trust certainly impacts the outcome of the interaction. Work performance can also be diminished if humans do not trust robots appropriately.

On the other hand, humans sometimes act dishonestly towards each other, affecting trust. Complementary, when humans keep their promises and act sincerely, they regain trust from each other. A question arises: *What if robots act dishonestly towards their human partners?*

In this paper, we analyze how human trust is affected by robots when they act dishonestly. Moreover, we compare how a human's trust is affected when the robot lies compared to how its trust is affected when another human lies. We perform an experiment around the *Split or Steal* game, a game based on the Prisoner's Dilemma, where the controlled player (i.e. controlled human actor, or a robot) can lie about the decision it intends to make. We evaluate trust levels through the Multi-Dimensional Measure of Trust (MDMT) [1] questionnaire.

Our results show that humans have a tendency towards having lower trust levels for robots than when compared to humans. They also show that the opponent's decisions play a decisive role in the trust levels, regardless of them being a robot or a human.

1.1 Research Questions

In this paper, we aim at answering the following research questions:

- RQ1 - Are robot opponents as thrust-worthy as their human counterparts?
- RQ2 - How does lying or not affect trust levels in robots and humans?

2 Related Work

The Prisoner's Dilemma Game (PDG) has been used as an instrument to study a plurality of human behavioral characteristics since its inception, a major one being trust [2]. Several variations of this dilemma have been created, one of them is the Repeated Prisoner's Dilemma Game (RPDG) which consists of playing several consecutive rounds, this means that the outcome of the previous rounds affects the decision of the player in the subsequent rounds [3].

Sandoval et al. [4] investigated reciprocity in HRI by arranging participants to play the RPDG against both a human and robot agent. "Split or Steal" is another variant, used as the final round of an UK TV gameshow, Goldenballs. Coffey [5] used the episodes of the gameshow as a natural experiment to analyze human response when faced with PD type situations. For our experiment, we applied on a small level the RPDG aspect of consecutive rounds, to the Split or Steal variant.

3 Technical Development

We used a virtual robot based on SoftBank Robotics' Nao [6] to implement our experimental procedure. For generating speech, we choose ResponsiveVoice [7].

The reproducible material for the experiment consists of a video recording of an interaction simulation, composed of the robots visual actions and voice lines. The robot assumes three behaviors: idle; taking position; and actively communicating.

The initial idle position can be seen in Figure 1. The second behavior corresponds to the robot walking closer to the capture point. Figure 2 shows the robot in the actively communicating behavior, where it has blue eyes and performs small hand movements. The robot also assumes short idle moments to indicate to the experiment conductor, in the Wizard of Oz setup, when to pause the video. This mechanism is transparent to the participants, who see no indication of it, and works as a variable-length idle moment.

To conduct the experiment remotely, we used the Zoom platform.

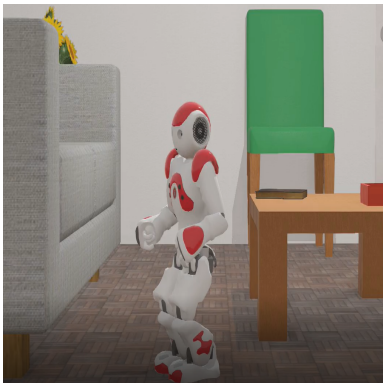

Fig. 1 Caption of the virtual robot in the initial idle position.

4 Methodology

To answer the research questions identified in Section 1.1

Our experiment involves two interactions with controlled actors, one with the virtual robot and another with a human actor. In each interaction, two rounds of the *Split or Steal* game are played. We randomize the order in which the interactions occur (i.e., against a human first, against the robot first).

The controlled actors follow an interaction script, made available in Annex A. They can assume two strategies for the game in hand: 1) always Split; 2) always Steal. The strategy was also randomly chosen for each participant. Nonetheless, the interactions always follow the same script, where they argue for both players to choose Split. The difference is only revealed through the game's conductor when she reveals the results of each round.

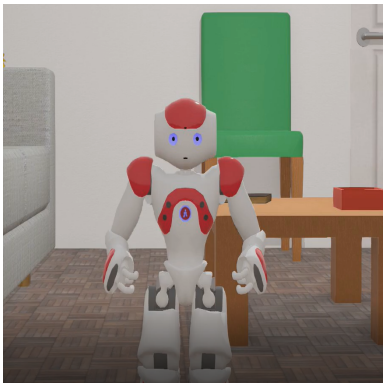

Fig. 2 Caption of the virtual robot while it is talking.

The aim of considering two different strategies is to generate a baseline situation, where the actor remains faithful to its word, and a lying situation, where the actor goes twice against it.

Our sample population's demography is discriminated in Table 1. In total the experiment counted with 21 participants. For 11, the actors always choose *Split*, while for the remaining 10 they always choose *Steal*. Most of the participants did not have previous experience with social robots.

At the end of the experiment, participants were handed the Multi-Dimensional Measure of Trust v1 [1] questionnaire. The participant's decisions were also registered.

Sample	*Total n*	21
	Split n	11
	Steal n	10
Age	18-29	100%
Gender	Male	61.9%
	Female	33.3%
	Non-Binary	4.8%
Previous experience w/ social robots		23.8%

Table 1 Participants' demography

5 Results

Figure 3 plots the trust levels for the robot and human trustees. The MDMT questionnaire [1] is divided into four subscales and an overall scale. Table 2 shows the discriminated results for these five scales and for each of the adopted strategies.

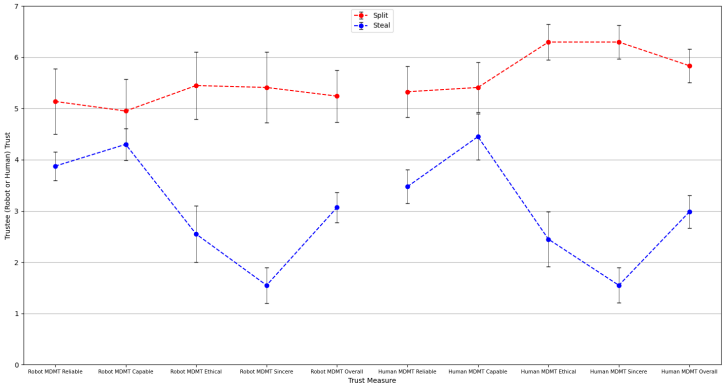

Fig. 3 Trust levels for the robot and human trustees, assessed on the MDMT subscales, and the MDMT overall score. Error bars show SE.

	Robot				**Human**			
	Split		*Steal*		*Split*		*Steal*	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
MDMT Reliable	5,136	2,111	3,875	0,892	5,326	1,652	3,475	1,037
MDMT Capable	4,95	1,957	4,3	0,978	5,409	1,617	4,45	1,418
MDMT Ethical	5,447	2,177	2,55	1,739	6,295	1,145	2,45	1,682
MDMT Sincere	5,409	2,289	1,55	1,104	6,295	1,1	1,55	1,085
MDMT Overall	5,24	1,683	3,069	0,924	5,834	1,085	2,981	1,006

Table 2 MDMT survey results

We find that people indicate more trust in the overall score ($p = 0.002$, $d = 1.58$) for robots who split ($M = 5.240$, $SD = 1.683$, $n = 11$) when compared to robots who steal ($M = 3.069$, $SD = 0.924$, $n = 10$).

We also find the same ($p < 0.001$, $d = 2.72$, $n = 11$) but when comparing humans who split ($M = 5.834$, $SD = 1.085$) when compared to humans who steal ($M = 2.981$, $SD = 1.006$, $n = 10$).

When comparing both humans who split ($M = 5.834$, $SD = 1.085$, $n = 11$) and robots who split ($M = 5.240$, $SD = 1.683$, $n = 11$), we find that people indicate more trust towards humans in the overall score ($p = 0.024$, $d = 0.419$).

However, when comparing humans who steal ($M = 2.981$, $SD = 1.006$, $n = 10$) and robots who steal ($M = 3.069$, $SD = 0.924$, $n = 10$), we fail to reject the null hypothesis ($p = 0.492$, $d = -0.090$).

When comparing all humans ($M = 4.476$, $SD = 1.782$, $n = 21$) with all robots ($M = 4.206$, $SD = 1.742$, $n = 21$), we find that people tend to indicate more trust towards humans in the overall score but not significantly ($p = 0.085$, $d = 0.153$).

6 Discussion

When it comes to RQ1, although not significant, our results suggest that humans have higher overall trust levels regardless of their strategy in the game.

As for RQ2, our results show that in the case of an actor being willing to cooperate with the participant then, human actors are perceived as more trustworthy than robot actors. This conclusion was to be expected as it matches the current scientific literature consensus. However, it is interesting that we failed to reject the null hypothesis regarding actors who lied and used a more individualistic strategy suggesting that, when lied to, the level of trust a person has about an agent is affected the same way regardless of the nature of the actor.

We also note the limitations of our experiments. It would be interesting to see if these results would remain similar with a larger and more representative population sample, as well as by increasing the number of rounds per human-actor interaction and with the actors having more complex level of strategies to pick from, instead of the binary choice of always lying or always being truthful. Another aspect for improving the soundness of the results would be to include another script, where the controlled actor argues for its own Steal decision.

7 Conclusion

Nowadays, robots are being studied with emphasis on human-robot interaction, where they may act as teammates or partners and collaborate with humans to achieve the desired goals. As a result, trust between humans and robots requires special attention since it plays a crucial role in human-robot interactions.

As future work, many more studies should be performed to evaluate better how trust is affected when robots are dishonest. One of the suggestions is to perform the proposed experiences with more participants to provide better accuracy as future work. Moreover, since the experiences took place remotely, we could only rely on the questionnaire from the participants. It could also be interesting to physically perform the experiences to conduct behavioral change analysis and measure participants' behavior when they naturally interact with different trustees. We believe that understanding trust is fundamental in improving human-robot interactions and that our work may serve as a ground point to further investigations.

References

[1] Ullman, D., Malle, B.F.: What does it mean to trust a robot? steps toward a multidimensional measure of trust. In: Companion of the 2018 Acm/ieee International Conference on Human-robot Interaction, pp. 263–264 (2018)

[2] Tedeschi, J.T., Hiester, D.S., Gahagan, J.P.: Trust and the prisoner's dilemma game. The Journal of Social Psychology **79**(1), 43–50 (1969)

[3] Shubik, M.: Prisoner's dilemma: A study in conflict and cooperation. by anatol rapoport and albert m. chammah.(ann arbor: The university of michigan press, 1965. pp. 229. 1.95, paper.). American Political Science Review **61**(1), 173–175 (1967)

[4] Sandoval, E.B., Brandstetter, J., Obaid, M., Bartneck, C.: Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game. International Journal of Social Robotics **8**(2), 303–317 (2016)

[5] Coffey, S.: Split or steal? a natural experiment of the prisoner's dilemma. A Natural Experiment of the Prisoner's Dilemma (October 1, 2009) (2009)

[6] SoftBank Robotics' Nao. https://cyberbotics.com/doc/guide/nao Accessed 2022-02-09

[7] ResponsiveVoice. https://responsivevoice.org/ Accessed 2022-02-09