Split or Steal: Are Robots More Trustworthy Than Humans?

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.

1 Scenario

Our scenario is built around the *Split or Steal* game, a version of the prisioner's dilemma [1] where 2 players dispute a fictional monetary prize each round. The players have the option to either "split" or "steal" the prize. The key difference is that there exists place for communication prior to the decision taking action. Table 1 shows how the prize is shared according to the chosen actions.

The human participant will meet and play with a series of actors which can be either human actors or a virtual robots. Each session will be divided in two stages. First, the controlled actor will present itself to the human participant, who will respond with an introduction of themselves. Then, a sequence of 5 *Split or Steal* rounds will be played. The first phase of each round consists of a period where the controlled actor communicates to the participant its intention of either split or steal that round. This statement may or may not correspond to their intention. After both players cast their choice, a revelation of the choices is made and the possible prizes are distributed.

		Controlled Actor (A)	
		Split	Steal
Human Participant (H)	Split	H: 50% A: 0%	P: 0% A: 100%
	Steal	H: 100% A: 0%	P: 0% A: 0%

Table 1 Payoff table for the *Split or Steal* game

2 Research Questions & Hypothesis

Our study is aimed at answering the research question: **How do trust levels in a social dilemma vary depending on the opponent (human vs. virtual robot) and their propensity to lie?**

We hypothesize that virtual robots will maintain their trust levels more consistently, even as their propensity to lie increases, when compared to humans. Our intuition reasons with human grudge likely being more intense when the target is another human compared with a non-living being. We also hypothesize that the initial trust values with humans will be higher when compared with the robots, due to humans being more able of bonding with other humans when compared with a robot.

3 Study Design

3.1 Conditions

Our study will be conducted through Zoom calls. This means that, as a condition, our participants must have access to an internet connected computer.

3.2 Participants

We plan on recruiting a sample of 10 human participants. The demography of these will largely consist of university students from IST.

We will also have one controlled participant for each class (i.e. wizard-of-oz virtual robot and human actors) and, inside each class, for each playing strategy.

3.3 Materials

Our study will require a virtual robot, which will serve as the base for creating the different robot participants. We intend to use Webots [2] as the simulation environment, and a simulated version of SoftBank Robotics' Nao [1] as the base of our virtual robots.

[1] https://cyberbotics.com/doc/guide/nao

3.4 Methodology

Each human participant will meet $2N$ controlled participants, where N is the number of different strategies for the controlled participants. Strategies can be defined, for instance, by a fixed sequence of actions with a certain percentage of split and steal decisions, or by a mimicking plan with an initial predefined decision for the first round.

We will divide each game into two stages: 1) the introductory stage, where both participants briefly introduce themselves; 2) the playing stage, where a sequence of the split or steal games are played.

At the end of the each stage, we will collect data useful for measuring and comparing the trust levels of human participants in relation with their opponents. Section 4 provides further detail on how this collection will be performed.

4 Evaluation

Based on the experiments performed, we evaluate the participant's trust. Prior to evaluating the trust, we have to state a clear definition of it. Trust in HRI literature has many definitions, each of them adapted to the particular environmental context. To simplify the reasoning about trust, we use the Wagner et al.'s [3] nomenclature of a trustee and trustor. The trustor is our participant, whose actions we do not control and intend to measure. The trustee is the robot or a human actor, part of the experiment whose actions we control. Correspondingly, in this paper, we define *trust as the degree of the expectation of a participant towards the robot.* In other words, the robot is more trustworthy if its actions are assumed to be more predictable.

Furthermore, to estimate the trust of a participant, we are going to use a mix of subjective and objective trust measurements.

In Human-robot trust research, there are two main methods for measuring trust: subjective and objective. *Subjective trust measurement* techniques involve assessing experiment participants' answers to the questionnaires. We define questions specific to our research as they may be more contextualized to the actual experiment, rather than using a predefined trust scale such as Human-Robot Trust Scale [4] and Trust in Automation Scale [5]. Thus, we believe our custom questions will yield improved accuracy.

The questionnaire is handed-out to the participants at the end of the first stage of each interaction. An arbitrary participant may arrive with an already formed opinion regarding robots, and therefore their trust is biased by their previous experiences. So for that matter, we built this trust baseline. After the experiment, a new questionnaire is performed. The intention is to obtain another point of measure regarding the trust of the participant towards the trustee. It can be used independently, as well as for analysing the trust level evolution when coupled with the initial questionnaires. We will use seven-point likert scales as the answer template for each question.

On the other hand, the subjective trust measurement is naturally less accurate due to their lack of dynamics of trust-building [6] and the tendency of the participants to speculate about themselves. Therefore, to enforce the accuracy of subject trust measurement, we perform an *objective trust measurement* on participants at the time of the experiment. In this type of trust measurement, we will analyze how participants interact with the trustees and infer, from experimental data, their trust evolution during the experiment. For that reason, we will perform a *behavioral change analysis*, where we can observe and measure participants' behavior when they are naturally interacting with different trustees [7].

References

[1] Poundstone, W.: Prisoner's Dilemma/John Von Neumann, Game Theory and the Puzzle of the Bomb. Anchor, ??? (1993)

[2] Michel, O.: Cyberbotics ltd. webots™: professional mobile robot simulation. International Journal of Advanced Robotic Systems **1**(1), 5 (2004)

[3] Wagner, A.R., Arkin, R.C.: Recognizing situations that demand trust. In: 2011 RO-MAN, pp. 7–14 (2011). IEEE

[4] Schaefer, K.E.: Measuring trust in human robot interactions: Development of the "trust perception scale-hri". In: Robust Intelligence and Trust in Autonomous Systems, pp. 191–218. Springer, ??? (2016)

[5] Jian, J.-Y., Bisantz, A.M., Drury, C.G.: Foundations for an empirically determined scale of trust in automated systems. International journal of cognitive ergonomics **4**(1), 53–71 (2000)

[6] Baker, A.L., Phillips, E.K., Ullman, D., Keebler, J.R.: Toward an understanding of trust repair in human-robot interaction: Current research and future directions. ACM Transactions on Interactive Intelligent Systems (TiiS) **8**(4), 1–30 (2018)

[7] Jayaraman, S.K., Creech, C., Robert Jr, L.P., Tilbury, D.M., Yang, X.J., Pradhan, A.K., Tsui, K.M.: Trust in av: An uncertainty reduction model of av-pedestrian interactions. In: Companion of the 2018 ACM/IEEE International Conference on Human-robot Interaction, pp. 133–134 (2018)