HRI paper (not actual title)

[Possible subtitle?]

Erin Paeng
info
info
info
info

Prof. Jim
info
info
info
info

ABSTRACT

Humans are irrational, sometimes surprisingly so. Game theory, while its value as a way to measure irrationality has been contested, nonetheless offers valuable insight on how we operate in a social setting requiring trust. In this study, we asked the question: if humans act irrationally when interacting with other humans, does their behavior change when interacting with a being they perceive as "rational"? [1] This question was explored through the classic game theoretic game prisoner's dilemma.
—List results here —

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*
Not sure what these are, will complete later

General Terms

Game theory, rationality, human-robot interaction

Keywords

Complete later

1. INTRODUCTION

Trust is fundamental to human interaction. It allows us to rely and cooperate with others, which was crucial to our advancement as a species. Hence, it seems natural that trust would be equally important in human-robot interaction. With growing social intimacy between humans and robots, research in the nature of trust between these agents has spread as well. More specifically, the continuing advancement of robots is elevating their status from manipulated tools to teammates and social partners- in essence, they are posed to adopt more "human" roles in interaction. As these agents assume independence and responsibility, trust becomes increasingly important to facilitate a natural and successful experience.

To motivate the study of trust, we present the following scenario: suppose a fire erupts in an occupied building, and both firemen and robots are sent in to guide people to safety. The robots can communicate with each other and thus have a more comprehensive picture of the situation as they search the area; however, they are unfamiliar as SAR agents. When presented with conflicting leads to safety, do people trust the robots or firemen more? They have no experience with either specific agent before, but may trust each agent differently.

In this paper, we use game theory to study trust and cooperation between agents. More specifically, we wish to explore: are people more eager to trust or cooperate more readily when interacting with an agent they perceive as rational? By examining a human agent's level of trust in robots relative to their trust in other humans, we hope to identify:

1. If humans trust humans more than robots without previous experience with the agent;

2. How quickly they are able to gain and sustain a cooperating strategy.

Furthermore, if humans do indeed trust robots more without previous experience and are able to obtain a stable cooperating strategy, we aim to learn:

1. If robots are perceived as more rational which thus results in more cooperative or trusting interactions;

2. If there are observable behaviors that indicate rising or falling trust in an agent.

2. BACKGROUND

As a result of increasing coexistence, considerable research has been done to explore factors that influence trust in human-robot interactions. As with many constructs, there are a variety of definitions for trust. It can be the expectation of an outcome based on communicated promise [3], or a willingness to take risks and reveal vulnerabilities [4]. Trust is a fundamental part of human interaction, and is subsequently importance in human-robot interaction - particularly considering robots are increasingly designed to cooperate with human agents in difficult or dangerous situations [5].

[1] What's the consensus on third person versus first person? Is there convention I should follow?

Muir stated that trust serves an important function in the proper usage of machines, and notes that an individual's trust for machines is influenced by factors similar to those between individuals. For instance, a reliable behavior increases trust while betrayal and subsequent loss of trust must be rebuilt [6]. Hancock et al. [1] have published a meta-analysis of factors affecting trust in human robot interaction and categorized these factors based on a survey of existing literature. They found that robot characteristics and performance were the largest influences on trust, implying trust may be most improved by altering a robot's performance. Bainbridge et al.[2] have studied how a robot's presence, whether virtual or physical, affects trust in interactions. Yagoda et al. [7] developed an HRI specific trust-metric which incorporates dimensions related to the human, robot, environment, system, and task.

Considerable research has tried to summarize possible factors that influence trust in HRI [7,8], but we are interested in measuring trust through a game theoretic approach. Game theory is a well-studied mathematical model that explores strategic decision-making [9], and strategy may involve cooperation or some degree of trust within agents. A canonical example of such a strategy may be found in the iterative Prisoner's Dilemma (IPD), upon which Axelrod founded his theory of the evolution of cooperation [10]. However, an important point should be made here: there is no real consensus on the causal relationship between trust and cooperation. However, IPD conflates the two in its game structure, which makes it difficult to test one or the other independently. Yamagishi et al. [3] developed a variant of prisoner's dilemma which involved variable dependence (PD/D)- that is, the level of cooperation was separated from how much a player chose to trust his opponent. It is this game model that we employed to compare both cooperation and trust between human-human and human-robot games.

While game theory has been used in HRI to explore social behavior [12, 13, 14], we propose using PD/D to measure trust and willingness to cooperate as separate phenomenon between human and robot agents. In this way, we can conduct a comparative analysis on the evolution of trust and cooperation between human-human and human-robot games. By further observing exhibited behaviors, we hope to identify physical indicators of changing trust (as indicated by the game) and use it to inform the robot of its current trustworthiness.

3. EXPERIMENTAL PARADIGM

Prisoner's Dilemma is one of the best known games of strategy in social science. In the traditional game, the police have arrested two suspects and are interrogating them in separate rooms with no communication between suspects. However, because the prosecutors lack evidence to convict both on the maximum sentence, they hope to send both to prison on a lesser charge. Both prisoners are offered the same bargain simultaneously, and each prisoner must choose to: betray the other by confessing that the other committed the crime or cooperate and remain silent. The offer is as follows: if both betray the other, each gets a moderate sentence (punishment payoff P); if both remain silent, each receives a lesser charge (reward payoff R); and if one betrays the other, the betrayer will be set free while the cooperator goes to prison on the maximum sentence (temptation payoff T and sucker's payoff S). In the iterative Prisoner's Dilemma (IPD), the game is played a random number of rounds with the requirement that $2R > T+S$ to ensure that mutual cooperation is the dominant strategy. This game structure is simple and allows for some measure of flexibility; ie, the researcher may choose a different storyline, change the payoff matrix, etc. It is further accommodating in that we may add elements to the story and observe the physical behavior of participants during gameplay.

The canonical structure of IPD, however, conflates the concepts of cooperation and trust. Unfortunately, there is no consensus on the definition of trust for scientists in fields that study it. We define it as the act of voluntarily exposing oneself to positive and negative externalities by the actions of others, as it is defined in trust game literature [15]. We define cooperation as the act of incurring personal opportunity cost to increase the welfare of others, where the latter is greater than the former. The abandoned potential gain (opportunity cost) is the defining point of cooperation. These are the definitions that Yamagishi et.al. adopted, and whose game play we model closely. By adopting a game that studies trust and cooperation independently, we can study the impact of agent variation on their development through the game and identify discrepancies that emerge.

Prisoner's Dilemma with Variable Dependency (PD/D) is a modification of the canonical game, developed by Yamagishi et al. While his paper listed several variants operating on the idea of distinguished trust and cooperation, we selected the Coin Entrustment (CE) variant as it was relatively simple to understand and had straightforward game play. CE is essentially a combination of the canonical PD and Investment Game (IG); the same choices of cooperating or defecting exist, but are supplemented by an additional trust element. In the game each player begins with 10 coins, and in each round must make two decisions:

1. How many of his coins to entrust to the other player;
2. Whether to keep or return the coins that were entrusted to him.

Each round is divided into two sub-rounds; in the first sub-round, each player decides the number of coins to entrust, and the amount is revealed to each player simultaneously. In the second sub-round, knowing how many coins their opponent had entrusted, each player decides whether to keep or return those coins and these decisions are revealed simultaneously. This process continues on for a pre-determined number of rounds; however, the number of rounds is undisclosed to either player.

As in PD, the payoff structure includes punishment (P), temptation (T), sucker (S), and reward (R) payoffs, expressed in the matrix below:

CE is unique in that the exact amount for these payoffs depends on the number of coins mutually entrusted. If a player chooses to return an entrustment, that entrustment is doubled in value when returned to the truster. If a player chooses instead to keep an entrustment, it is added to their personal score. Let quantity x be Player A's entrustment in Player B and quantity y be Player B's in Player A. The payoff matrix for this transaction is below:

Therefore, potential gain is governed by the trusted instead of the truster, unlike in traditional IG in which the coin value is tripled before being entrusted to the other player.

We provide the following justifications for the CE game structure:

1. To build robots capable of cultivating trust, we must first understand the nature of the current trust between humans and robots. Specifically, we are studying the emergence of trust in an agent with which the human participant has no previous experience.

2. To nurture cooperation between humans and robots, it is helpful to understand to what extent trust is involved in effective cooperation. Often cooperation is necessary before the opportunity to build trust has presented itself, in which case the results from this experiment are helpful.

4. EXPERIMENTAL SETUP

4.1 Method

4.1.1 Participants

200 participants were recruited from Amazon's Mechanical Turk (AMT) to complete an online experiment. They were compensated a base amount of $0.25 for a n-minute study, with opportunities to earn an additional $0.50 in bonuses. We modeled are high-level experimental setup after previous studies that have also successfully utilized AMT for HRI social experiments [2]. Research indicates that data collected from participants sampled through AMT compare well to that collected through traditional sampling; furthermore, AMT is more diverse than undergraduate college students and more representative of the Internet-using population [16, 17]. Hence, we concluded AMT was an acceptable medium through which to conduct this study.

4.1.2 HIT Structure

Agents. Descriptive characteristics about both agents were left undisclosed, as the experiment's intent was to explore people's general opinions about robots compared humans, and how those internal conceptions impact trust and rational cooperation. Furthermore, this design choice allows for future detail specification and serves as a comparative baseline. The opponents were described as "a robot opponent" and "a human opponent" in the instructions, and thereafter referred to as "your opponent". The agent type was the only manipulated variable in the study.

Game play. CE was played for n rounds in each HIT assignment. While opponents were described as "robot" or "human", both agents were actually just strategy algorithms. To enhance the "human" experience for games against a human opponent, artificial wait times were randomly executed to indicate that the human opponent was thinking. Such wait times were excluded from games involving the "robot" opponent. After the game concluded, a survey was presented that attempted to gain insight on the following questions:

1. What motivated participants when playing the game, and are there differences in motivation between playing against human and robot opponents?

2. Which qualities of the opponent agent impacted how quickly trust and cooperation were developed?

3. What level of trust do people have in robots as a concept compared to their trust in humans?

Procedures and measures. The experiment comprised of two main parts, presented in a web browser as three separate pages. Upon consenting to participate, the Turker was taken to the game page with instructions for CE and buttons to indicate selections. It was made clear that the total number of rounds was unknown to either player, and that decisions made in each round would be revealed simultaneously.

When the game was complete, an alert prompted the participant to the second page, which listed the survey questions. Participants were asked three main questions. They were first asked to select the option from the following that best reflected their motivation for the game: "beating my opponent", "maximizing my earnings, "helping my opponent", "finishing the game as quickly as possible", and "other". Second, they were asked to identify qualities that applied to the opponent agent [18] from the following qualities: intelligence, faculty for sensations sympathy, perfection, humanity, faculty for feelings, precision, life, and reliability. Lastly, they were asked to rate the following phrases related to the agent's trustworthiness on a seven-point Likert scale [19]: "robots are deceptive", "robots behave in an underhanded manner", "I am confident in robots", "robots have integrity", "robots are dependable", "robots are reliable", "I can trust robots", and "I am familiar with robots".

The survey was essentially presented twice to each participant, addressing each agent (human and robot) separately. A Turker that played against a robot first answered questions about robots; they were then asked to imagine a game involving the opposite agent (a human) and answer the same questions. A Turker that played against a human answered these questions in reverse order (pertinent to humans first, then robots).

**** Include picture of the game setup

5. REFERENCES

[1] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(5):517–527, 2011.

[2] B. F. Malle, M. Scheutz, and J. Voiklis. Sacrifice one for the good of many? people apply different moral norms to human and robot agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 117–124. ACM, 2015.

[3] T. Yamagishi, S. Kanazawa, R. Mashima, and S. Terai. Separating trust from cooperation in a dynamic relationship prisoner's dilemma with variable dependence. *Rationality and Society*, 17(3):275–308, 2005.