# HRI paper (not actual title)

## [Possible subtitle?]

Erin Paeng
info
info
info
info

Prof. Jim
info
info
info
info

## ABSTRACT

Normal social interactions often require some degree of trust; without it, the world would be immensely more complicated to navigate. Game theory offers valuable insight on how we operate in a social setting requiring trust. In this study, we asked the question: how do humans interact with robots differently than with other humans in situations requiring trust? 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 successful 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 the emergence of trust and cooperation between agents. More specifically, we wish to explore: do people interact with an agent they perceive as a robot in a quantitatively or qualitatively different way than they interact with an agent they perceive as human? By examining a human agent's level of trust in robots relative to their trust in other humans, we hope to identify:

*Do trust and cooperation emerge differently between two humans than between a human and a robot?*

While this experiment uses game theory to explore comparative trust and cooperation, we believe results found in this experiment lend valuable insight on the nature of interaction between humans and robots. Fields beyond game theory can benefit from studies that reveal, at least to some degree, how humans perceive and subsequently interact with robots compared to with humans.

## 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 [20], or a willingness to take risks and reveal vulnerabilities [13]. 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 [7].

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 [18]. Hancock et al. [11] 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. [3] have studied how a robot's presence, whether virtual or physical, affects trust in interactions. Yagoda et al. [22] 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 [22] [5], 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 [19], 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 [2]. However, an important point should be made here: there is no real consensus on the causal relationship between trust and cooperation. IPD conflates the two in its game structure, which makes it difficult to test one or the other independently. Yamagishi et al. [23] 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. In this work, we adopt Yamagishi's definition of trust as "an act that voluntarily exposes oneself to greater positive and negative externalities by the actions of the other(s)" [23], which is its definition in trust game literature as well [10]. Furthermore, we adopt Yamagishi's definition of cooperation as "an is an act that increases the welfare of the other(s) at some opportunity cost where the former is greater than the latter" [23]. The abandoned potential gain (opportunity cost) is the defining point of cooperation. 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.It is this game model that we employed to compare the emergence of cooperation and trust in human-human and human-robot interaction.

While game theory has been used in HRI to explore social behavior [14] [6] [17], 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 emergence of trust and cooperation between human-human and human-robot games. To our knowledge, no such comparative analysis using a game theoretic approach has been conducted before.

## 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 the 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.

The canonical structure of IPD, however, conflates the concepts of cooperation and trust. A popular game to study trust separately in game theory was introduced by Berg et al. [4] as the Investment Game (IG), and can be described as follows: subjects in rooms A and B are given equal endowments. Subjects in room A must decide how much of their endowment to send to their anonymous counterpart. This amount is tripled then given to the partner in room B. The subject in room B must choose how much of the money to return to their counterpart in room A. In this way, while both parties have opportunities to increase their wealth, but the success of their strategy to do so relies heavily on the level of trust between the two players.

Prisoner's Dilemma with Variable Dependency (PD/D) is a modification of PD, 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. Furthermore, results gained from CE have shown remarkably different levels of trust and cooperation in PD/D compared to that in PD; given its success in measuring trust and cooperation independently in game theoretic scenarios, it was deemed an appropriate experimental setup for our study. CE is essentially a combination of the canonical PD and 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:

[insert matrix]

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:

[inset matrix]

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 given 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 how trust emerges in human-robot interaction compared to in human-human interaction. 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.

3. Finally, to ensure long-lasting relationships between humans and robots, we must explore the unfortunate cases when trust is broken, and how trust and cooperation re-emerge following that event.

# 4. EXPERIMENTAL SETUP

## 4.1 Method

### 4.1.1 AMT for behavioral studies

There are several thoughts to consider when using AMT to conduct behavioral research. On one hand, research indicates that data collected from participants sampled through AMT compare well to that collected through traditional sampling; furthermore, AMT provides more diverse population than undergraduate college students and more representative of the Internet-using population [9] [16]. However, there has been growing concern that experienced Turkers have encountered very similar studies multiple times, and are thus immune to the "gut reaction" response that many researchers seek [8]. Furthermore, Turkers are exposed to common experimental paradigms more than researchers realize— more than half have been previously exposed to Prisoner's Dilemma and Ultimatum Game scenarios [8].

Our experiment relies on the perception dyadic interaction. Summerville et al. explored pseudo-dyadic interaction[1] through AMT and found through four individual studies that Turkers responded to "real" partners in a qualitatively similar manner to those in a laboratory setting [21]. Of particular interest to us is their second study, which used the popular Dictator Game scenario between a "human" recipient and "no recipient". The results gained here were comparable to those of laboratory experiments in which the participant was interacting with an actual person. When surveyed afterward, very few if any were suspicious of computerized interaction without probing; however, with probing, up to a third expressed suspicion. In general, they were more suspicious when the nature of the partner was a focal point of the study; hence, a cover story or additional steps to imitate true dyadic interaction seems to be especially important when using AMT. Given that Summerville et al. were successful in performing pseudo-dyadic studies with lab-comparable results, we replicated some aspects of their experimental setup.

Given this knowledge surrounding AMT's usage in experiments of social behavior, we concluded that it was an acceptable medium through which to conduct this study.

[1]A dyad is defined in sociology as a group of two people. Hence, pseudo-dyadic interaction would be a mock interaction between two people

### 4.1.2 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 $0.75 in bonuses. We modeled our high-level experimental setup after previous studies that have also successfully utilized AMT for HRI social experiments [15] [21].

### 4.1.3 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.

*Computerized coin entrustment.* The following is a description of the algorithm used to calculate coin entrustment in each round. In general, the algorithm tended towards more cooperative behavior and encouraged higher entrustment by readily exhibiting greater trust. Pseudo-code of the algorithm is presented below; in it let $M_n^c$ and $E_n^c$ represent the computer's move and entrustment respectively in round $n$ and $M_n^p$ and $E_n^p$ represent the player's move and entrustment respectively in round $n$ :

A design choice was made here to set the minimum coin entrustment to 1. In making this decision between a minimum of 0 or 1 coins, we looked to IG's and IPD/D's structure. In both, a minimal amount of 0 was allowed; in IG, as in our experiment, transferring 0 has no impact on the payoff for either player, and the receiver of this entrustment could choose to keep or return all 0 [4]. Yamagishi conducted IPD/D experiments with both minimums. However, we were concerned that, following a defection by one player, a continuous cycle of defections and zero entrustments may become the status quo. This behavior is counterproductive to our goal of studying the emergence of trust and cooperation. Hence, we chose to retain the 1 minimum in our experiment.

*Computerized cooperation or defection.* The decision to keep or return coins followed the principle behind the optimum strategy in Yamagishi's paper [23]. Called a cautiously cooperative strategy, it never defects and adjusts only the level of coin entrustment to reflect trust. Because we want to explore both the initial emergence of trust and cooperation

**Algorithm 1** Computerized coin entrustment

```
if M^c_{n-1} or M^p_{n-1} were defections then
  return 1
else if E^c_{n-1} or E^p_{n-1} > 8 then
  return 10
else
  x ← exponential average of E^c_i ∈ i in range(0, n)
  if E^p_{n-1} - E^p_{n-2} > 1 then
    adder = 3
  else
    adder = 1
  end if
  return select random(E^p_{n+1}+adder, x+adder)
  if E^c_{n-1} - E^p_{n-1} > 2 then
    return E^p_{n-1} + 3
  else if E^c_{n-1} == E^c_{n-2} then
    return E^c_{n-1} + 2
  end if
end if
```

and its re-emergence after a betrayal of trust, our strategy defects on round 8 if the participant has not already defected in the previous rounds. The advantage of this strategy is that it is entirely deterministic, assuring each player is exposed to the same strategy and thus controlling the experiment.

*Game play.* CE was played for 16 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. These wait times were calculated by the following algorithm; in it, let $t$ be the time between the two most recent button clicks and $p$ be the wait time executed after the first click:

**Algorithm 2** Wait time calculation

```
if t − p > 2 seconds then
  wait 1 second
else
  y ← rand(0, t − p)
  wait ← select random(y, y + 2sec, y + 3.5sec, y + 4sec,
  0.5sec)
  return select random(wait, 0)
end if
```

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. Do qualities attributed to humans and robot differ? [1]
3. What level of trust do people have in robots as a concept compared to their trust in humans? [12]

*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 [1] 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 [12] "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] K. Arras and D. Cerqui. Do we want to share our lives and bodies with robots. *A 2000-people survey, Technical Report Nr. 0605-001 Autonomous Systems Lab Swiss Federal Institute of Technology*, 2000.

[2] R. Axelrod and W. D. Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

[3] W. Bainbridge, J. Hart, E. S. Kim, B. Scassellati, et al. The effect of presence on human-robot interaction. In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, pages 701–706. IEEE, 2008.

[4] J. Berg, J. Dickhaut, and K. McCabe. Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142, 1995.

[5] D. R. Billings, K. E. Schaefer, J. Y. Chen, and P. A. Hancock. Human-robot interaction: developing trust in robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 109–110. ACM, 2012.

[6] F. Broz. *Planning for human-robot interaction: representing time and human intention*. ProQuest, 2008.

[7] J. Casper and R. R. Murphy. Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 33(3):367–385, 2003.

[8] J. Chandler, P. Mueller, and G. Paolacci. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, 46(1):112–130, 2014.

[9] M. J. Crump, J. V. McDonnell, and T. M. Gureckis. Evaluating amazonâĂŹs mechanical turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410, 2013.

[10] P. Dasgupta. Trust as a commodity. *Trust: Making and breaking cooperative relations*, 4:49–72, 2000.

[11] 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.

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

[13] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80, 2004.

[14] K. W. Lee and J.-H. Hwang. Human–robot interaction as a cooperative game. In *Trends in Intelligent Systems and Computer Engineering*, pages 91–103. Springer, 2008.

[15] 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.

[16] W. Mason and S. Suri. Conducting behavioral research on amazonâĂŹs mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[17] M. B. Mathur and D. B. Reichling. An uncanny game of trust: social trustworthiness of robots inferred from subtle anthropomorphic facial cues. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 313–314. ACM, 2009.

[18] B. M. Muir. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27(5):527–539, 1987.

[19] R. B. Myerson. Game theory: analysis of conflict. *Harvard University*, 1991.

[20] J. Rotter. A new scale for the measurement of interpersonal trust. *Journal of personality*, 1967.

[21] A. Summerville and C. R. Chartier. Pseudo-dyadic âĂIJinteractionâĂİ on amazonâĂŹs mechanical turk. *Behavior research methods*, 45(1):116–124, 2013.

[22] R. E. Yagoda and D. J. Gillan. You want me to trust a robot? the development of a human–robot interaction trust scale. *International Journal of Social Robotics*, 4(3):235–248, 2012.

[23] 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.