The 2016 AAAI Fall Symposium Series:
Artificial Intelligence for Human-Robot Interaction
Technical Report FS-16-01

Trust and Cooperation in Human-Robot Decision Making

Jane Wu, Erin Paeng
Human Experience & Agent Teamwork Lab
Harvey Mudd College, Claremont, CA
{jhwu,epaeng}@hmc.edu
www.cs.hmc.edu/HEAT/

Kari Linder, Piercarlo Valdesolo
Moral Emotions and Trust Lab
Claremont McKenna College, Claremont, CA
{klinder16,pvaldesolo}@cmc.edu
www.valdesolo.com/meat-lab/

James C. Boerkoel Jr.
Human Experience & Agent Teamwork Lab
Harvey Mudd College, Claremont, CA
boerkoel@hmc.edu
www.cs.hmc.edu/HEAT/

Abstract

Trust plays a key role in social interactions, particularly when the decisions we make depend on the people we face. In this paper, we use game theory to explore whether a person's decisions are influenced by the type of agent they interact with: human or robot. By adopting a coin entrustment game, we quantitatively measure trust and cooperation to see if such phenomena emerge differently when a person believes they are playing a robot rather than another human. We found that while people cooperate with other humans and robots at a similar rate, they grow to trust robots more completely than humans. As a possible explanation for these differences, our survey results suggest that participants perceive humans as having faculty for feelings and sympathy, whereas they perceive robots as being more precise and reliable.

Introduction

Trust is fundamental to day-to-day human interactions, allowing us to rely on and cooperate with others. Trust has proven to be equally important in many human-robot interaction (HRI) applications (Bainbridge et al. 2008; Hancock et al. 2011; Haring, Matsumoto, and Watanabe 2013; Muir 1987; Yagoda and Gillan 2012), and will only become more important as the shift towards using robots as teammates, rather than just manipulated tools, continues.

This paper sets the foundation for understanding how to build robots capable of cultivating trust in HRI applications. We use game theory to study the emergence of trust and cooperation between agents. Further, we explore how differences in trust impact human-robot and human-human decision making, and whether trust influences the level of cooperation and rationality in those decisions. We also explore how trust and cooperation re-emerge after a robot violates trust. Finally, we explore how participants' motivations and perceptions shift when partnering with humans *vs.* robots. We pose the following hypotheses:

Hypothesis 1 *Humans will achieve and maintain higher levels of* trust *when interacting with what they believe to be a* robot *than with another* human.

Hypothesis 2 *Humans will* cooperate *more readily and consistently when interacting with what they believe to be a* robot *than with another* human.

We suspect that when an agent is perceived as rational (i.e., a robot), it will prompt people to adopt more rational behavior themselves. The game setting we use requires both trust *and* cooperation to optimize performance. Hence, if both parties are rationally optimizing their expected payoffs, we expect a more trustful and cooperative relationship to emerge rather than one biased by emotions or prejudices.

This paper contributes a comprehensive background that discusses the importance of trust in HRI and establishes the game-theoretic foundations of both trust and cooperation. We contribute an experimental paradigm that uses the Coin Entrustment game as a way to test our hypotheses using Amazon Mechanical Turk and in-person lab experiments. Finally, our empirical exploration of our hypotheses allows us to conclude that over the course of the game, humans begin to trust robots to a greater degree than other humans, while cooperating equally well with both.

Background

In this section, we explore the importance of trust in HRI, review game-theory inspired explorations of trust, and discuss related efforts in previous HRI work.

Trust in HRI

Due in part to increasing coexistence, human-robot trust and factors influencing interactions involving trust have been the subject of several recent research efforts. This increasing attention necessitates an examination of what trust means in the context of decision-making in HRI. Trust, for instance, can denote the expectation of an outcome based on a communicated promise (Rotter 1967), or a willingness to take

risks and reveal vulnerabilities (Lee and See 2004). Muir states that trust serves a vital role in the proper use of machines, and notes that an individual's trust for a mechanism is influenced by factors similar to those that influence interpersonal relationships. Reliable behavior builds trust, while betrayal undermines it (Muir 1987). Hancock et al. (Hancock et al. 2011) 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 influence trust most dramatically, implying trust may be most improved by altering a robot's performance. Bainbridge et al. (Bainbridge et al. 2008) investigated how the virtual or physical presence of a robot affects trust in interactions. Furthermore, Haring et al. explored how physical appearance and behavior of a life-like andriod robot impact the level of trust as measured through proximity and an "in-person" economic trust game (Haring, Matsumoto, and Watanabe 2013). Yagoda et al. (Yagoda and Gillan 2012) developed an HRI specific trust-metric that incorporates dimensions related to the human, robot, environment, system, and task. In this paper, we expand on this previous work by measuring trust and cooperation using a game theoretic approach.

Game-theoretic Definitions of Trust

Game theory is a well-studied mathematical field that explores strategic decision making (Myerson 1991) and requires cooperation and trust between agents. We adopt Yamagishi's definition of **trust** as *"an act that voluntarily exposes oneself to greater positive and negative externalities used by the actions of the other(s)"* (Yamagishi et al. 2005). This is the definition in trust game literature (Dasgupta 2000). Furthermore, we also adopt Yamagishi's definition of **cooperation** as *"an act that increases the welfare of the other(s) at some opportunity cost where the former is greater than the latter"* (Yamagishi et al. 2005).

Related Work

There is a rich history of using game theory to study decision making in HRI. For example, Lee lists games, such as Twenty Questions, as an effective approach to understanding trust. Games that reveal how personal payoff influences players' behavior have also been shown to be effective proxy for understanding human-robot cooperation (Lee and Hwang 2008). Marthur *et al.*, used a one-shot Investment Game (IG) along with facial tracking to conclude that the expected wagers were higher when playing against mechanical robots than against humanoid robots (Mathur and Reichling 2009). Trust can be heightened by programming robotic partners to exhibit cues predictive of trustworthy economic behavior in humans (Desteno 2012). To our knowledge, the approach we take is novel in that it attempts to understand *both* trust *and* cooperation as *separate* phenomena.

Experimental Paradigm

We use the Coin Entrustment (CE) game, a variant of the prisoner's dilemma proposed by Yamagishi et al. (Yamagishi et al. 2005), as the foundation for our experimental paradigm. CE is not only simple to understand and straightforward to play, but has also been shown to successfully measure trust and cooperation independently (Yamagishi et al. 2005). Our use of the CE game facilitates the exploration of trust development in human-robot decision making, as well as correlations between trust and effective cooperation. In addition, to ensure long-lasting relationships between humans and robots, CE allows us to explore the unfortunate cases when trust is broken (e.g., either due to a mechanical or logical error or due to an intentionally exploitative decision by the robot), and how trust and cooperation re-emerge.

Game Procedure

CE is an iterative game with multiple rounds, each of which involves the exchange of coins between two players. At the start of each round, both players begin afresh with 10 coins. First, each player commits a number of coins (1-10) to entrust to the other player, and the amounts are revealed to each player simultaneously.

Then, each player decides whether to keep the coins entrusted or return them to their partner. When returned, coins double in number. Again, these decisions are revealed simultaneously. The player's score per round is the number of coins in his/her possession at the end of the round. For instance, if A entrusted 3 coins to B, who in turn entrusted 5, and both players chose to return their opponent's coins, A would end the round with 13 coins ($7 + 3 \times 2$), while B would end the round with 15 coins ($5 + 5 \times 2$). If A instead chose to keep B's coins, A would end the round with 18 coins ($7 + 3 \times 2 + 5$), and B would end the round with a mere 5 coins. This process continues for a pre-determined number of rounds; however, the exact number of rounds is undisclosed to either player.

Experimental Method

This section describes our experimental setup, participants, game setup, and algorithms. We introduce the term **human condition** to refer to the game played against a perceived human opponent, and the term **robot condition** to refer to the game played against a perceived robot opponent.

Participants

Our study recruited participants from two main sources—Amazon's Mechanical Turk and college students, and was approved by our local Institutional Review Board.

Amazon's Mechanical Turk Our experimental design involves the use of Amazon's Mechanical Turk (AMT). Research indicates that data collected from participants sampled through AMT compare well to that collected through traditional human experiments. Furthermore, AMT provides more diversity than our convenience population of college students and is more representative of the general Internet-using population (Crump, McDonnell, and Gureckis 2013; Mason and Suri 2012). To mitigate concerns of possible bias among experienced Turkers (Chandler, Mueller, and Paolacci 2014; Crump, McDonnell, and Gureckis 2013; Mason and Suri 2012), we modeled key aspects of our setup after previous studies that have successfully utilized AMT for

HRI social experiments (Malle, Scheutz, and Voiklis 2015; Summerville and Chartier 2013). Our experiment relies on the perception dyadic interaction[1]. Summerville et al. explored pseudo-dyadic interaction through AMT and found that Turkers responded to "real" partners in a qualitatively similar manner to those in a lab setting (Summerville and Chartier 2013). In general, participants were more suspicious when the nature of their 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. We describe how we implement these ideas in the Gameplay section.

230 participants were recruited from Amazon's Mechanical Turk (AMT) to complete an online experiment. They were compensated $0.25 for a 15-minute study, with opportunities to earn an up to an additional $0.50 based on their performance (average winnings per round).

Lab Experiment We also ran our experiment in a physical lab setting by recruiting 32 undergraduate participants (16 for each condition), incentivizing participation with class credits and a chance for a gift card based on performance. All participants accessed the same webpages used for AMT. Participants in the human condition were told they would play a person in a different room. This method was preferred over including a human confederate, which would have introduced additional social biases. Participants in the robot condition were told they would play our Alderbaran NAO robot, which stood on the table next to the computer and spoke its moves out loud in each round.

Each participant was brought to the lab, which contains four isolated computer cubicles. Human condition participants were brought in groups of 1-3, and robot condition participants were brought in individually. A researcher explained the game mechanics, and participants interacted with the same strategy algorithm as previous game setups.

Gameplay

All participants were taken through the same three steps: consent, gameplay, and a qualitative survey[2]. Upon consenting to participate, the participant proceeded to an instruction screen (Figure 1).

For both AMT and in-lab studies, the experiment was implemented using a web-based interface in which participants used text boxes and buttons to indicate their decisions. All participants played CE for 16 rounds. Finally, all participants completed the same web-based survey to collect qualitative perceptions about their experiences.

Agents For the AMT experiments, descriptive characteristics about both agents were left undisclosed, as the experiment's intent was to explore how people's internal perceptions about robots and humans impact trust and rational cooperation, following the lead of Summerville et al.(Summerville and Chartier 2013),. The opponents were described

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

[2]The team will publicly share all experimental materials.

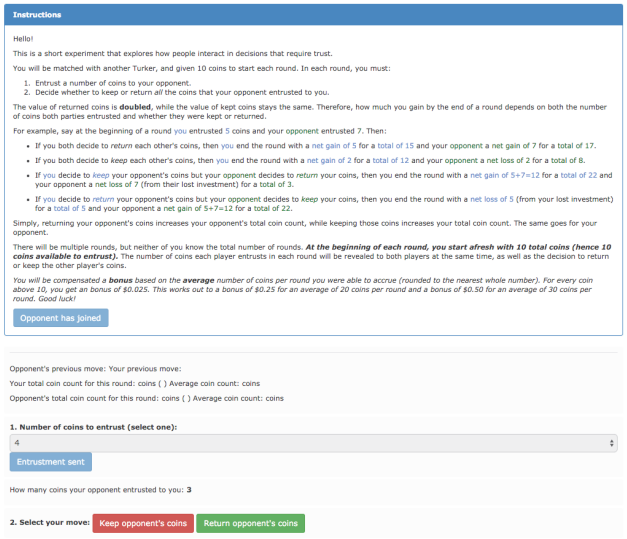

Figure 1: Screenshot of main instruction and gameplay page

as "a robot opponent" or "a human opponent" in the instructions, and thereafter referred to as "your opponent". In the lab experiment, participants were told they would play another human in a different lab (human condition) or the NAO robot (robot condition). In all experiments, opponents were implemented using exactly the same deterministic algorithm (described next); perceived agent type was the only manipulated variable.

Algorithmic Coin Entrustment Algorithm 1 describes how we compute the number of coins to entrust in each round. In general, our algorithm tends towards higher entrustment by readily exhibiting increasing trust. Our strategy is based on a Pavlovian model—in each round, it bases its entrustment on its payoff in the previous round. The algorithm always begins by entrusting 3 coins in the first round. If either player defected (kept their opponent's coins) in the previous round, then the algorithm entrusts 1 coin (trust was betrayed). If the algorithm's payoff in the previous round was greater than 0 (entrusted coins were returned), then it entrusts more coins in this round.

Algorithm 1: Coin Entrustment

Input : previousPayoff: net coin gain in previous round.
Output: The number of coins to entrust.
if *first round* **then**
  entrust 3
else
  if *either player defected in the previous round* **then**
    entrust 1 ;
  else if previousPayoff > 0 **then**
    entrustment = $\lceil 10 + (\text{previousPayoff} - 10)/1.5 \rceil$;
    entrust $\min(\text{entrustment}, 10)$;
  else
    entrust $\max(1, 10 + \text{previousPayoff})$;

We made a design choice to set the minimum coin entrustment to 1, rather than 0. A zero entrustment leads to ambiguity about the difference between the cooperate and defect decisions, since both lead to no coins being returned. As a result, a continuous cycle of defections and zero entrustments often becomes the status quo. Selecting 1 coin as the minimum keeps such decisions concrete and permits clearer interpretations of trust and cooperation.

Algorithmic Cooperation The decision to keep or return coins followed the Tit-for-Two-Tats (TFTT) strategy. To encourage the possibility of trust, TFTT was favored over the similar Tit-for-Tat strategy, where the computerized agent defects in response to a single defection. As with entrustment, our cooperation algorithm tends towards more cooperative behavior. The computerized agent cooperates in the first round, defecting only if the human has defected twice in a row. Our strategy also purposely defects in the eighth round if the participant (and hence computer algorithm) has not already defected in the previous rounds. This permits us to explore both the initial emergence of trust and cooperation and their re-establishment after a betrayal of trust.

Mimicking Human Play To enhance the believability of a "human" opponent, we exploit the strategies described in Summerville et al., chosing to use wait times to enhance believability (Summerville and Chartier 2013). First, players were prompted with dialogs displaying *"Waiting for more players to join the queue..."* for several seconds to idicate the selection of an opponent from a larger group. Additionally, a *"Waiting for opponent's move..."* indicator was used between rounds to simulate decision-making time.

Wait times were calculated by Algorithm 2. Here we use t to represent the time between the participant's two most recent button clicks and p to represent the previous wait time as calculated by the algorithm. We first want to check if the participant took an atypically long time to make his/her decision; if so, we want to wait a shorter amount of time (hence, returning 1 second). Next, we flip a random coin—if it comes up heads, we compute a random amount of time to wait, otherwise we return our answer immediately. To ensure that our wait times are relatively believeable, we uniformly sample a value y from the range 0 to $(t-p)$, which represents the "lag" between the participant's move and our algorithms most recent move. This ensures that the amount of time we tend to wait is on the same scale as the human participant. We then extend y by adding an additional 0, 0.5, 2, 3.5, or 4 seconds, selected randomly.

Wait times are imposed each time we algorithmically make a decision in the human condition. Such wait times were excluded from games involving the "robot" opponent.

Post-game survey After the game concluded, we presented a survey targeting the following questions:

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

Algorithm 2: Wait Time Calculation

Input : t: time (seconds) between participant's two most recent clicks. p: previously computed wait time.
Output: A wait time, in seconds.
if $t - p > 2$ *seconds* **then**
  return 1 second ;
else if *randBoolean()* **then**
  $y \leftarrow randDouble(0, t - p)$;
  return $y+$ randomlySelect$(0, 0.5, 2, 3.5, 4)$
else
  return 0;

2. Do qualities attributed to humans and robot differ? (Arras and Cerqui 2000)
3. What level of trust do people have in robots compared to their trust in humans? (Jian, Bisantz, and Drury 2000)

To address the first question, participants were asked to select one of 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"*. For the second question, they were asked to identify qualities they believed apply to the opponent (human or robot) (Arras and Cerqui 2000) from the following: intelligence, faculty for sensations, sympathy, perfection, humanity, faculty for feelings, precision, life, and reliability. Last, they were asked to rate the following phrases related to the agent's trustworthiness on a seven-point Likert scale (Jian, Bisantz, and Drury 2000). The descriptors pertinent to robots were: *"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 questions pertinent to humans replaced all instances of "robots" with "human".

The survey was presented twice to each participant, addressing each type of opponent (human and robot) separately. Any player who played against a (perceived) robot first answered questions about robots; they were then asked to imagine playing the same game against a human opponent, and answer the same questions. A participant that played against a (perceived) human answered these questions in reverse order (pertinent to humans first, then robots).

Results

Our experiments measure trust and cooperation to observe how participants play CE differently when presented with what they believe to be a human or a robot opponent. We define *trust* as the number of coins a player entrusts to their opponent—the more coins a player entrusts (i.e., increasing risk), the more trust is presumed to exist between the players. Similarly, we define *cooperation* as a participant's decision to return or keep (cooperate or defect) his/her opponent's coins. Both of these are consistent with the intended design of the CE game (Yamagishi et al. 2005).

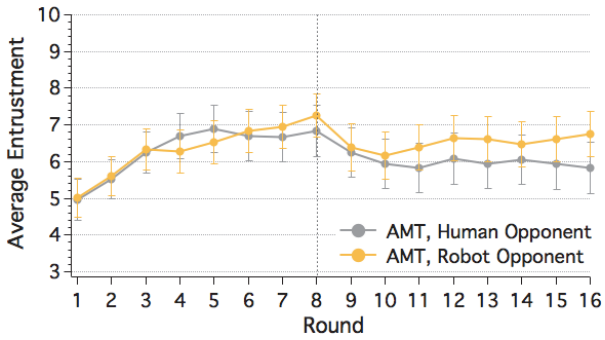

Figure 2: Average coin entrustment per round (AMT)

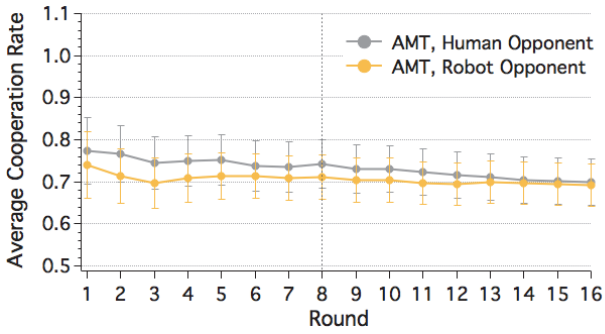

Figure 3: Average cooperation per round (AMT).

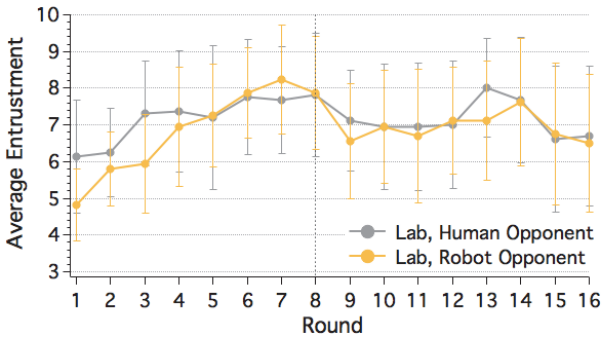

Figure 4: Average coin entrustment per round (Lab)

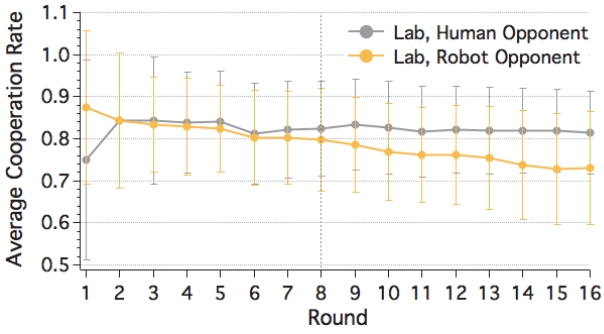

Figure 5: Average cooperation per round (Lab)

If Hypothesis 1 is true, we would expect the number of coins entrusted to the robot opponent to be greater than that entrusted to the human opponent. If Hypothesis 2 is true, we would expect participants to maintain a higher rate of cooperation with the robot agent. The null hypothesis in each case is that no difference exists between the conditions—humans trust and cooperate with humans and robots to the same degree. We explore each of these hypotheses in the subsequent subsections and conclude by discussing the perceived qualities attributed to each opponent type. We analyze the results from the AMT and lab experiments separately.

Trust

We measure **trust** in terms of the number of coins a participant is willing to put at risk (entrust). We explore the (re)emergence of trust in two phases of the game. First, we see how trust initially develops (before defection), and second, we explore whether trust is impacted by our defection in the eighth round. In the AMT study, participants developed initial trust more quickly with a robot than with a human (see Figure 2); while both entrusted initially 5.0 coins, the robot condition peaked at 7.2 coins entrusted compared to the human condition's 6.8 coins entrusted immediately before programmed defection. Error bars represent the 95% confidence intervals across all our results.

Additionally, through the course of the game, the average entrustment to a "robot" opponent increasingly deviates from the amount entrusted to a "human" opponent. We used a mixed ANOVA to evaluate our results, with the between-subjects factor being the opponent type and the within-subjects factor being the 16 rounds. Our ANOVA confirms that opponent type leads to statistically significant differences in coins entrusted across all rounds, with $F(15, 3405) = 1.804, p < .05$. Therefore, we confirm our first hypothesis that players trust robots more than humans across all rounds.

Cooperation

In each round, the participants decided whether to return (cooperate) or keep (defect) the coins entrusted to them. We calculate the **cooperation rate** as the ratio of times a participant cooperated versus defected as each round progressed. Again, we analyze the AMT and lab results separately. In the AMT study, average cooperation rates for both conditions were nearly identical (Figure 3). Furthermore, the cooperation rate changed very little over the rounds, suggesting that participants responded to defection in the eighth round by reducing their trust (coin entrustment) rather than by defecting themselves in the next round. The results also suggest that participants cooperated with a "human" opponent just as readily as a "robot" opponent. In sum, our results do *not* support our second hypothesis, as we are unable to reject the null hypothesis using our ANOVA.

Lab Trust and Cooperation

Our lab results provide both a supplement to our AMT study and an interesting perspective on how the presence of a physical robot can affect a participant's trust and cooperation levels. In Figure 4, we see that in both conditions trust is developed in the first eight rounds and lost after programmed de-

fection. While we cannot reject the null hypothesis, our lab entrustment results seem to reinforce trends seen on Turk.

However, average cooperation rates from the lab yield more interesting differences from AMT (Figure 5). In the first round, cooperation for the human condition was 13% below the robot condition value, yet the averages converged immediately in the second round. Additionally, after the eighth round the robot condition's average fell to 8% below that of the human condition, whereas the values began to converge in the AMT study.

Teammate Perceptions

Next we turn to the question of *why* we see the trends that we do. Each participant was asked to respond to survey questions that explored their motivations, reasons for trust, and perceptions of their opponents. We compare the robot and human conditions for each of these questions below.

Motivation For the AMT study, participants were most often motivated by coin maximization (and thus monetary bonuses). The motivation to beat their opponent was higher in the human condition (24% as oppsoed to 9%). However, when asked to imagine the game with a robot opponent, 31% in the human condition said they would want to beat their opponent; when participants in the robot condition were asked about an imaginary human opponent, 26% were motivated by victory.

For the lab study, 94% of participants stated coin maximization as motivation, with 0% motivated by beating the humanoid robot. Participants who played against a human opponent were most motivated by coin maximization, even when imagining a robot opponent (69% for both cases).

In the AMT study, we find that humans are most motivated by a victory-defeat scenario when matched with a human opponent (real or imagined), and most motivated by maximizing score when faced with a robot opponent (real or imagined). This points to a possible dynamic in interpersonal relationships that is missing from human-robot interactions in the game—that is, the desire for social dominance.

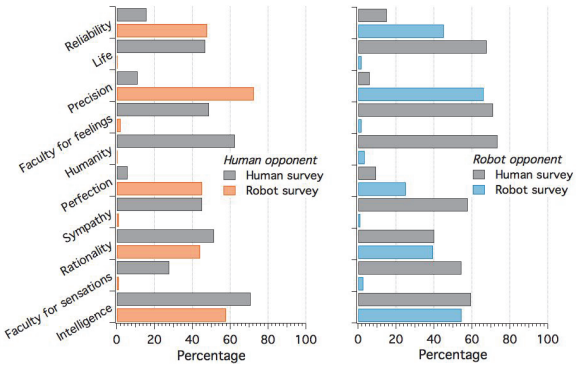

Figure 6: Survey results about perceived human *vs.* robot qualities depending on actual opponent type (AMT)

Trust In both AMT and lab studies, participants reported on average that they found humans to be slightly more trustworthy than robots, (difference of 0.3 in AMT and 0.6 in the lab on a 1-7 scale). These results suggest an implicit bias to associate trust with humans over robots, but this bias may not significantly affect a participant's actions during the CE game, as shown in our results.

Agent qualities In both studies, a majority of participants ($> 50\%$) thought *humans* have faculty for feelings, sympathy, humanity, and life. On the other hand, participants thought of *robots* as precise (>50%) and reliable (46% in AMT and 59% in the lab). Interestingly, while participants in the lab robot condition played against a humanoid robot, rather than simply a computer, their agent quality results show trends that match the Turk robot condition results, i.e. greater association of perfection, precision, and reliability with robots than humans.

Discussion

In this paper, we explored how people trust and cooperate with robots differently than with humans using a Coin Entrustment game—a framework designed to separately measure emergence of trust and cooperation. Furthermore, our game-theoretic definitions of trust and cooperation allow us to simplistically model them in real world HRI applications. By defining quantitative metrics for these two phenomena, we can begin to measure the importance of trusting perceived agents (entrusting coins) and cooperation (returning an opponent's coins), two key elements of successful human-robot interactions.

In our AMT study, we confirmed our hypothesis that in repeated interactions with a robot, a human may grow to trust a robot teammate *more than* a human teammate. We recall that, following programmed defection, participants accelerated their trust in the computerized robot opponent more quickly than in their human opponent. However, we were unable to confirm our hypothesis that people would cooperate with robots more quickly and fully than with humans. The AMT results suggest that humans use trust, in the form of coins, rather than cooperation, to hedge against human players, which they may view with less certainty and more skepticism. Therefore, while trust varied more widely over the rounds, cooperation stayed relatively consistent when the participants played against a computerized opponent. Yet, our lab results show that cooperation with our NAO robot fell more significantly after trust was purposely broken, suggesting a distinction between playing a computerized opponent and a humanoid robot. Trust, in the form of coin entrustment, was similar between the human and robot opponents, suggesting that how participants choose to hedge their bets- whether by cooperating or trusting less- depends on the game circumstnaces. The AMT study found that participants altered trust, while the lab study showed greater behavioral distinction in cooperation. Finally, participants' motivations changed depending on their perceived opponent—in the AMT study, participants were most motivated to win against a human, and maximize their score when playing against a robot. In the future, we would like to extend our explorations to a wider variety of interaction domains, perhaps introducing a robot as a collaborator, rather than an opponent. We also hope to gain a better understanding

of the trends we see, with an emphasis on how trust and cooperation are both used to navigate the complexities of human-robot interaction, as we found differences between computerized and android opponents. In particular, it would be useful to explore the nature of reciprocity in the scope of human-machine trust, as well as different ways of exposing interdependency between agents. Additionally useful would be a broader analysis of how trust and cooperation independently translate into action in other scenarios. We hope continued examination of trust and cooperation in HRI can be applied to improve the design of human-robot systems.

References

Arras, K., and Cerqui, D. 2000. 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*.

Bainbridge, W.; Hart, J.; Kim, E. S.; and Scassellati, B. 2008. The effect of presence on human-robot interaction. In *Proc. of RO-MAN 2008*, 701–706.

Chandler, J.; Mueller, P.; and Paolacci, G. 2014. Nonnaïveté among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods* 46(1):112–130.

Crump, M. J.; McDonnell, J. V.; and Gureckis, T. M. 2013. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one* 8(3):e57410.

Dasgupta, P. 2000. Trust as a commodity. *Trust: Making and Breaking Cooperative Relations* 4:49–72.

Desteno, Breazeal, F. P. B. D. L. 2012. Detecting the trustworthiness of novel partners in economic exchange. *Association for Psychological Science*.

Hancock, P. A.; Billings, D. R.; Schaefer, K. E.; Chen, J. Y.; De Visser, E. J.; and Parasuraman, R. 2011. 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.

Haring, K. S.; Matsumoto, Y.; and Watanabe, K. 2013. How do people perceive and trust a lifelike robot. In *Proc. of the World Congress on Engineering and Computer Science*, volume 1.

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

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

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

Malle, B. F.; Scheutz, M.; and Voiklis, J. 2015. Sacrifice one for the good of many? people apply different moral norms to human and robot agents. In *Proc. of HRI-2015*, 117–124.

Mason, W., and Suri, S. 2012. Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods* 44(1):1–23.

Mathur, M. B., and Reichling, D. B. 2009. An uncanny game of trust: social trustworthiness of robots inferred from subtle anthropomorphic facial cues. In *Proc. of HRI-2009*, 313–314.

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

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

Rotter, J. 1967. A new scale for the measurement of interpersonal trust. *Journal of Personality*.

Summerville, A., and Chartier, C. R. 2013. Pseudo-dyadic "interaction" on amazon's mechanical turk. *Behavior Research Methods* 45(1):116–124.

Yagoda, R. E., and Gillan, D. J. 2012. 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.

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