Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Human-Robot Trust and Cooperation Through a Game Theoretic Framework

Erin Paeng, Jane Wu, and **James C. Boerkoel Jr.**
Human Experience & Agent Teamwork Laboratory (http://cs.hmc.edu/HEAT/)
Harvey Mudd College, Claremont, CA
{epaeng, jhwu, boerkoel}@g.hmc.edu

Introduction

Trust and cooperation are fundamental to human interactions. How much we trust other people directly influences the decisions we make and our willingness to cooperate. It thus seems natural that trust be equally important in successful human-robot interaction (HRI), since how much a human trusts a robot affects how they might interact with it. As a result, considerable research has been done to explore factors that influence trust in HRI, particularly using game theory (Lee and Hwang 2008; Mathur and Reichling 2009). Unfortunately, these approaches lack a comparative analysis of trust and cooperation as distinct qualities.

We propose using a coin entrustment game (explained later), a variant of prisoner's dilemma, to measure trust and cooperation as separate phenomenon between human and robot agents. With this game, we test the following hypotheses: (1) Humans will achieve and maintain higher levels of trust when interacting with what they believe to be a robot than with another human; and (2) humans will cooperate more readily with robots and will maintain a higher level of cooperation. This work contributes an experimental paradigm that uses the coin entrustment game as a way to test our hypotheses. Our empirical analysis shows that humans tend to trust robots to a greater degree than other humans, while cooperating equally well in both.

Experimental Paradigm

To measure both cooperation and trust separately, we use the Coin Entrustment (CE) game, a variant of Iterative Prisoner's Dilemma (IPD) that attempts to separately measure trust and cooperation (Yamagishi et al. 2005). In each of an undisclosed number of game rounds, each player begins with 10 coins and must make two decisions: (1) How many of their coins to entrust to the other player; and (2) whether to keep or return the coins that were entrusted to them by the other player. Coins that are returned double in value. The payoff matrix for a single round of this game is expressed in Table 1 and depends on both Player A's entrustment x, and player B's entrustment y. C refers to a player cooperating (returning coins), while D refers to defection (keeping coins). In CE, the level of trust between players is measured by the number of coins entrusted, and the level of cooperation by keeping or returning the opponent's coins.

		Player B: C	Player B: D
Player A	C	$2x, 2y$	$-x, 2y + x$
	D	$2x + y, -y$	$y - x, x - y$

Table 1: Coin Entrustment Payoff Matrix.

Experimental Design

Our experiment was run on Amazon's Mechanical Turk (AMT), a crowd-sourcing site, with a total of 229 players. Participants navigated through three sections: a consent page with game instructions, 16 rounds of the CE game (the number of rounds is unknown to the player), and a post-game survey. While all participants played against the same strategy algorithm, we told half that their opponent was a robot and the other half that their their opponent was a human. Descriptive characteristics about both agents were left undisclosed, as the experiment's intent was to explore how people's perceptions about robots vs. humans influenced their trust and cooperation.

Gameplay

Algorithm 1 describes how we compute the number of coins to entrust in each round (recall that each player begins each round with 10 coins). In general, our algorithm tended towards more cooperative behavior and encouraged higher entrustment by readily exhibiting greater trust. Our strategy is based on a Pavlovian model, where entrustments are based on the strategy's payoff in the previous round.

The decision to keep or return coins followed the Tit for Two Tats (TFTT) strategy used in IPD literature. Our algorithm cooperates on the first round, and defects only if the opponent has defected twice in a row. To explore both the initial emergence of trust and cooperation and its re-emergence after a betrayal of trust, our strategy also defects on round 8 if it has not already defected in the previous rounds. Our strategy is entirely deterministic and controls for variance in strategy, leaving perceived opponent type as the only manipulated variable.

Algorithm 1: Coin Entrustment

Input : PREVIOUSPAYOFF, the total number of coins won in the previous round. ENTRUSTMENT, the number of coins entrusted by our opponent in the 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
    ENTRUSTMENT = 10+(PREVIOUSPAYOFF-10)/1.5
    entrust $\min(\text{ENTRUSTMENT}, 10)$

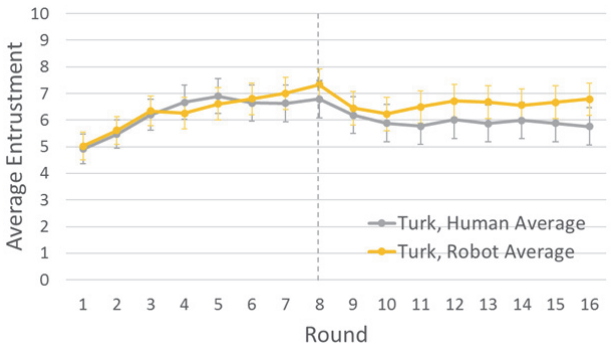

Figure 1: Average entrustment (measured in the number of coins entrusted) per round.

Post-game survey

After the game concluded, we presented participants with a survey that asked: (1) What motivated participants when playing the game (to assess whether people who played against robot *vs.* human opponents were differently motivated)?, (2) Do qualities attributed to humans and robot differ? (Arras and Cerqui 2000), and (3) How does participants' trust in robots compare to their trust in humans? (Jian, Bisantz, and Drury 2000). The survey was presented twice to each participant, addressing perceptions about humans and robots separately. A participant that played against a robot first answered questions about robots, then were asked to imagine a game involving the a human opponent and answer the same questions (and vice versa).

Summary of Results

We collected results along three dimensions: trust—the number of coins entrusted per round; cooperation rate—the rate at which players choose to return, rather than keep, their opponent's coins; and qualitative perceptions about the opponent—measured by our post-game survey. In Figure 1, we see how a participant's trust in their opponent progresses through the game. 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 in across the rounds. This difference is particularly acute after defection occurs in round 8—humans tended to regain trust in their robot opponents, but not human opponents. We also measured the cooperation rate per round, which we calculate as the cumulative ratio with which the participant cooperated versus defected. We found that participants cooperated at nearly identical rates regardless of opponent type, indicating people tended to adjust the number of coins they entrust rather than punish their opponent by keeping their coins. We were thus unable to reject the null hypothesis for our second hypothesis.

Our post-game survey highlighted differences in people's perceptions about robots and humans. Participants perceive that humans posess greater intelligence, rationality, sympathy, humanity, faculty for feelings and sensations, and life, while robots are more perfect, precise, and reliable.

Discussion

In this paper, we explored how people trust and cooperate with robots differently than they do with humans using a game-theoretic framework. Our experiment confirms that people develop trust faster and to a greater extent with robots than they do with humans in this particular game scenario. In the future, we would like to extend our explorations to a wider variety of interaction domains. We would also like to develop a deeper understanding of why such trends emerge. For instance, participants who played against a robot report coin maximization as their primary motivator, while those who played against a human were primarily motivated by victory over their opponent. This points to a possible dynamic in inter-personal relationships that is missing from human-robot interactions—the desire for social dominance.

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*.

Jian, J.-Y.; Bisantz, A. M.; and Drury, C. G. 2000. Foundations for an empirically determined scale of trust in automated systems. *Int'l J. 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.

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.

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.