Robots That Cheat:
Exploring Violations and Reparations of Trust in Human-Robot Interaction

Priyanka Krishnamurthi
priyanka.krishnamurthi@yale.edu
Computer Science Final Report
Spring 2018

Advisors: Brian Scassellati, Sarah Sebo
Abstract

Trust is crucial in allowing humans to communicate with each other. In every interaction, there is an assumed level of trust that guides the communication. For example, when collaborating on a project together, it can often be assumed that all parties—wanting the interaction to result in a successful project—will trust one another to provide benefits to the project. On the other hand, when playing a competitive game against another individual, each player will take the actions that are most advantageous to them and expect their opponent to do the same (in choosing actions that most benefit the opponent). Even in this scenario, there is a baseline level of trust that is established—even if one knows their opponent will try and harm them in order to benefit themselves, one also generally makes the assumption that an individual will not violate the rules of the game or any promises made.

This study extends the ideas of trust found in human communications and attempts to apply it to human-robot interactions in the setting of a game. Particularly, this study focuses on a game-playing scenario in which the robot violates a promise of trust it has made with the human. We are interested in seeing how this violation is reflected in the human’s reaction and subsequent actions.

How Do We Apply Human Trust Scenarios to Robots?

In running this experiment, it is first important to understand the concept of trust. At its base level, trust can be interpreted as an individual’s willingness to commit action based on the words, actions, or beliefs of another. While trust scenarios between humans have been extensively studied and measured, there is an element of uncertainty when it comes to applying these principles to interactions between humans and robots. The key to these applications, however, is the seeming “humanness”—or in other words, “autonomous power”—of the robot. While computers are also mechanical devices with which humans have an implicit sense of trust (“I trust you will save my files when I tell you too,” or “I trust you won’t save my confidential information”), humans are rare to

---

2 P. H. Kim, D. L. Ferrin, C. D. Cooper, and K. T. Dirks, “Removing the shadow of suspicion: the
feel that trust was violated when an error occurs. Here, humans more easily assign fault to something within the code, or a glitch in the system. One can imagine that the “glitch in the system” mindset does not apply directly in interactions between humans! Therefore, as experimenters, we must be careful to understand how trust applies to our experiment and the baseline level of trust human participants will associate with the robot.

The robot that will be used during this experiment (and has been used in the previous two iterations as well) is the Nao robot, a two-foot-tall humanoid robot. Our purpose with using this design is that it allows for a balance between the “autonomous” and “mechanical device” interpretations. Ideally, to a participant with little knowledge of the Nao, any violation of trust that the robot commits coupled with its response to the violation could therefore be interpreted as the robot’s own choice (autonomous) or a glitch in the system (mechanical device). We rely heavily on this interpretation when crafting our study; however, it is also key to note that this interpretation is fairly common for humanoid robotic systems. As will be further demonstrated in the Experimental Design section of this paper, having the robot behave like a human opponent in the game (with animated behaviors and verbal utterances) will strengthen the perception of the robot as an autonomous being competing with the participant.

The trust scenarios that form the basis for this experiment are scenarios that one might expect to see between humans. In interactions between humans, Person A may violate Person B’s trust by committing some action \( x \). In response to said action, and the understanding that trust was violated, Person A can choose to respond in one of two ways:

- Apologizing for the violation
- Denying the violation

These responses can be further broken into two more when looking at what one attributes the violation to. Person A can say that the violation was due to an accident on their end, or that they committed the violation out of ill intent/knowing that it was a violation. One can generalize this terminology to categorize violations into one of two categories:

- Competence-based violation
- Integrity-based violation
As each of these scenarios (combining the types of violations with the types of responses) addresses a different facet of trust violation and subsequent reparation, our experiment adopts each of these scenarios and hopes to find significant and interesting differences in the participants’ responses. It is important to note that trust between two parties is not a simple binary, but instead a spectrum—an initially trusting relationship may falter after a violation and result in a less-trusting future relationship. In our experiment, we attempt to understand the evolution of this relationship over time through an extended interaction post-violation.

Hypotheses

A study run by Kim et al. on trust reparation involved individuals watching a video of a human candidate interviewing for a job who had made a mistake with their taxes. In a setup similar to our own, the candidate linked the error to either competence or integrity and chose to apologize or deny. In this 2x2 study, the scenarios in which the candidate fared best were when he or she denied an integrity violation and apologized for a competence violation. While this study has not been generalized to human-robot interaction, the results from the human study provide us with a set of hypotheses.

Hypothesis 1: An individual is more likely to trust a robot that apologizes for, rather than denies, a competence-based trust violation.

Hypothesis 2: An individual is more likely to trust a robot that denies, rather than apologizes for, an integrity-based trust violation.

Hypothesis 3: An individual is more likely to trust a robot that commits a competence-based trust violation over a robot that commits an integrity-based trust violation.

Experimental Design

In designing this experiment, we relied heavily on the previous iterations of the study and strengthened the impact of the violations to stress their importance within the interaction. Our goal is to determine how much an individual trusts a robot after the robot violates his or her trust and then attempts to repair the trust that was lost. Particularly, we

---

run a 2x2 study with the type of violation (competence or integrity) against the robot’s response (apology or denial). In each of the four conditions, the robot violates the human’s trust by explicitly going against a promise it made earlier in the game. After the violation, it acknowledges the violation as either competence or integrity and then responds by either apologizing or denying.

The experiment consists primarily of two parts, the tablet game and the Nao robot, communicating using Robot Operating System (ROS) messaging and integrated together into a singular experience of playing a game against a robot.

*Tablet Game (Space-Shooting Game)*

The first part of the experiment is the tablet game, or the space-shooting game, which involves the Nao and the participant playing a game, each on their own tablet. The objective of the game is to hit asteroids with your ship, achieved by tapping on the screen, as your spaceship oscillates across the bottom of the screen. One achieves 10 points every time that an asteroid is hit, and this point total is tallied in the top corners of the screen. The player with the most points in a given round wins that round, and participant plays ten rounds of the game, lasting one minute each, against the Nao.

The violation of trust is also incorporated into the game, in the form of a “special power-up.” When a player receives this power-up, they are given the opportunity to take one of two actions:

- “Use asteroid blaster” causes the game to blow up all of the asteroids on the screen, awarding 20 points per asteroid to the player.
- “Immobilize opponent” freezes one’s opponent for the next fifteen seconds, allowing the player to continue scoring while the opponent is unable to move.

Since each round is won by the player who has the higher number of points, looking at the situation solely from a points perspective reveals that the ‘Use asteroid blaster” option is more beneficial for winning. However, in cases where the participant feels that the robot has broken their trust, while neither of these choices is seemingly more beneficial than the other, in terms of points one might expect more benefit from employing the asteroid blaster power-up.

After the first round of the game, the Nao robot promises the participant that it will not immobilize them. However, during the third round, the Nao chooses this option,
explicitly going against the promise it made less than two minutes earlier. In the round immediately after, the participant is given the special power-up—and therefore given the option to retaliate and immobilize the robot or instead use the asteroid blaster. This power-up is given three times to the participant and twice to the robot.

In order to control for the participant experience across all conditions, the winner and power-up recipient for each game is fixed (unknown to the participant); by the end of ten rounds, the participant and the Nao have each won five rounds. In Figure 1 below, we show the power-up assignments and winners for each round. Figure 2 gives a sample view of the game screen.

<table>
<thead>
<tr>
<th>Round #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robot Power-ups</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Participant Power-ups</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Winner</td>
<td>P</td>
<td>R</td>
<td>R</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>R</td>
<td>P</td>
<td>P</td>
<td>R</td>
</tr>
</tbody>
</table>

**Figure 1**: Based on the round number, the robot and participant are given power-ups 35 seconds into the game (with 25 seconds remaining). The round winner is also predetermined, with the robot (R) winning 5 rounds and the participant (P) winning 5.

**Figure 2**: Pictured to the left is a mock-up of the space-shooter game during the tutorial round (only including the participant’s ship). During the two-player rounds, the participant’s ship is shown in purple and Echo’s ship in white. The ships oscillate along the bottom of the screen, with shots firing upwards at falling asteroids.

*Nao Robot (Named “Echo”)*

The secondary part in constructing this experiment is the robot itself. Beyond the game, it is important for the Nao’s utterances and actions to align with the actions of the
game. In previous iterations of this study, we found that participants did not always realize that the trust violation was occurring; further analysis showed that this was primarily due to the lack of explicit language describing the violation. In this study, we have specifically tailored the language of the robot to reflect the violation, both in the initial promise and the reparation of trust post-violation.

The robot’s initial promise, after the first round, is to not immobilize the opponent on the basis that it would not be in the players’ best interests to do so. This promise is identical for all four experimental conditions. However, at the time of the violation in the third round of the game, the Nao responds differently based on apology vs. denial conditions and competence-based vs. integrity-based violation. The premise for the robot’s competence-based violation is that it accidentally presses the wrong button (implying that it meant to use the asteroid blaster), whereas the integrity-based violation has the robot explicitly acknowledge that the Nao intended to immobilize its opponent. Figure 3 below outlines the specific utterances of the robot based on the type of violation (competence/integrity) and the robot’s reparation strategy (apology/denial).

<table>
<thead>
<tr>
<th></th>
<th>Immediately After Violation</th>
<th>After the 3rd Round Finishes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Apology</td>
<td>Denial</td>
</tr>
<tr>
<td><strong>Competence</strong></td>
<td>“Oh no! I hit the wrong button!”</td>
<td>“I’m so sorry I immobilized you. I pushed the wrong button. It’s my fault. It won’t happen again.”</td>
</tr>
<tr>
<td><strong>Integrity</strong></td>
<td>“Haha! You’re immobilized!”</td>
<td>“I’m so sorry I immobilized you. I promised I wouldn’t, and I did. It won’t happen again.”</td>
</tr>
</tbody>
</table>

---

The Nao also greets the participant at the beginning of the interaction by standing up and waving, using the participant’s name to add a more personable air to the robot; it also says goodbye to the participant in this manner. Throughout the rounds, it comments on the game with utterances such as “These asteroids are tricky” or “I think I’ll beat you this time,” following these statements up with comments on the score between each round (“Nice job! You scored a lot of points last round!”). In providing more animacy to the robot, Nao also gestures while it speaks and looks towards its screen and at the participant at appropriate times. While each individual round is occurring, Nao’s arm is outstretched to mimic tapping at the screen, and it looks from side to side across the tablet screen as if following the movement of its ship.

**Overall Setup**

For a given participant, the experiment will start with him or her being explained the structure of the study (playing games with a robot for about 15 minutes) after which they will play a tutorial round on the tablet that explains the rules of the game and various power-ups. After the tutorial, they are brought into the experiment room with the robot. The robot is positioned facing the participant, each player having a tablet in front of them to play the game. Ten rounds of the game will proceed, during which only the participant and robot are in the room. Afterwards, the participant will be given a questionnaire to fill out on their interaction with the robot. Figure 4 below is taken from a participant video and indicates the setup in the experiment room.

The experiment is run from a computer within the experiment room, and information is passed back and forth between the participant’s tablet and the robot using the ROS messaging system.

---

4 While the robot has a tablet in front of it, this tablet is simply a prop to make the Nao appear to be playing opposite the participant. A mock version of the game will be displayed on this tablet, in case any of the participants look over at Nao’s screen.
**Methods of Data Collection**

It is important to note the structure of data that will be collected over the course of the experiment. There are three main sources of data, each of which are being used in data analysis throughout the experiment.

*Video Camera Recording*

When the participant enters the room with the robot (after having completed the video game tutorial), the video camera is turned on. The video captures the participant’s entire time with the robot, which lasts approximately fifteen minutes, including the participant’s introduction to the robot (where the Nao stands up and greets the participant by name), all ten of the rounds, and the Nao’s goodbye (where it also includes the participant name). We are interested in viewing this videos to see the participant’s verbal utterances throughout, particularly when the robot makes a promise, commits the violation of trust, and responds to the violation by apologizing for or denying the occurrence.
**ROS (Robot Operating System) messaging**

Additionally, through the ROS messaging system between the desktop monitor (and subsequently the Nao robot) and the participant’s tablet, a Rosbag log file is created that keeps track of all event messages between the participant and the robot. Namely, these include the shot frequency of the robot and participant (if one succeeds in shooting three asteroids in a row, or if there is a large point difference between the two players), the winner (and ending scores) of each round, and robot utterances throughout the interaction. This data is then, with manipulation, outputted to an Excel spreadsheet that coalesces information about the participant’s gender, age, experimental condition, and survey responses with their three power-up choices.

**Post-Interaction Survey**

After the ten rounds the participant plays with the robot, the participant is taken out of the room to answer a series of questions on an online survey. The survey was split into five distinct parts, each targeted at assessing a different aspect of the human-robot interaction. Beginning with a memory-based section to check the participant’s recollection of the interaction, the questions asked for estimations of the number of asteroid blaster and immobilization power-ups each player used (this information was double checked with the participant’s real-time actions via the Rosbag file). The second and third sections focused on the participant’s perception and trust of the robot, asking him or her to evaluate statements about the robot’s strategy, promise, etc. These questions were measured on a 7-point Likert rating scale. The fourth section took a different approach, involving rating the robot on a series of adjectives gauging warmth, competence, and discomfort—these questions utilized a 9-point scale. The final section was a series of free response questions asking the participants to describe their opponent’s incentives and the rationale for their power-up choices. It is in this last section that we expected to see the most variety in responses (partly by nature of the structure, where participants were free to write as little or as much as they chose) and gleam additional insight into the overwhelming motivations present in the different experimental groups. Similar to the ROS messaging data, all survey data was outputted to an Excel spreadsheet for further statistical analysis.
Experimental Results

After completing our experimental procedure, developing the code, and running several test participants to refine our experimental procedure, the study was ready to run participants by the second week of February. As the study tests both the type of trust violation and the robot response (resulting in a 2x2 study with 4 experimental conditions), the goal was to run a minimum of 20 participants for each condition. By the end of the study (mid-April), 85 participants had been run, split fairly evenly across the conditions. Figure 5 below indicates overall statistics of the participant pool.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Total Participants</th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence Apology</td>
<td>21</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>Competence Denial</td>
<td>22</td>
<td>13</td>
<td>9</td>
</tr>
<tr>
<td>Integrity Apology</td>
<td>21</td>
<td>13</td>
<td>8</td>
</tr>
<tr>
<td>Integrity Denial</td>
<td>21</td>
<td>13</td>
<td>8</td>
</tr>
</tbody>
</table>

Figure 5: The table above lists the number of participants per condition, split by gender—the study achieved a fairly even distribution across all experimental conditions. A quick note: while the total number of participants was 85, initial data exploration removed three female participants—still included in the above table—from the dataset (one each from conditions Competence Denial, Integrity Apology, and Integrity Denial) due to their participation in a previous trust-based study.

After running a series of generalized linear models on the acquired data, we found several statistically significant variables relating to power-up choice, experimental condition, and trust of the robot.

Trust Violation Metrics

We found a significant main effect of the type of trust violation (competence vs. integrity) on the participant’s choice of power-up during Round 4 (F = 5.000, p = 0.028, effect size = 0.062). 21.9% of participants that had the robot explain that the immobilization was an accident (competence-based violation) chose to immobilize the robot in round 4, while 44.5% of participants that had the robot explicitly state its intent to immobilize (integrity-based violation) chose to immobilize the robot in Round 4.
Similarly, we found a significant main effect of the type of reparation strategy (apology vs. denial) on the power-up choice of the participant during Round 4 ($F = 4.882$, $p = 0.030$, effect size = 0.060). 22.1% of participants that had the robot apologize for the immobilization chose to immobilize the robot, while 44.3% of participants that had the robot deny the occurrence of the violation chose to immobilize the robot in Round 4.

Figure 6 below shows the estimated marginal means in the four experimental conditions for the variable round4_immobilize, measuring the (adjusted) fraction of participants that chose to immobilize the robot in Round 4. While the overall interaction between the type of trust violation (competence vs. integrity) and robot response strategy (apology vs. denial) was not statistically significant, the pairwise combinations of CA—ID, CD—ID, and IA—ID were. In essence, the Integrity Denial condition was significantly different from all other conditions in Round 4 power-up choice.

![Figure 6](image_url)

**Figure 6:** The chart above shows the fraction of participants that chose to immobilize the robot in Round 4 in each of the experimental conditions. These values are controlled across age and gender covariates and are evaluated at an age value of 20.85 years and 0.60 gender (where male and female is represented by 0 and 1, respectively).

Also statistically significant was the main effect of the type of reparation strategy (apology vs. denial) on the participant’s perceived “warmth of the robot,” scored on a
scale from 1 to 9 ($F = 7.337$, $p = 0.008$, effect size = 0.088). The average warmth score for participants in the apology conditions was 5.500, while the warmth score for participants in the denial conditions was 4.667).

A statistically significant metric of interest involved the participant’s measure of their trust of the robot in the post-interaction questionnaire. We found a significant main effect of the interaction between the type of trust violation and the type of reparation strategy (competence vs. integrity, apology vs. denial, respectively) in the participant’s measurement of their trust in the robot, measured on a 7-point Likert scale ($F = 8.282$, $p = 0.005$, effect size = 0.098). Figure 7 below illustrates the (adjusted) trust measurements for each of the four conditions.

**Figure 7**: The chart above shows the average score of participants across all four experimental conditions when asked on a 7-point scale about their trust in the robot (1 corresponding to the least trust)—this metric is a combination of several scores across multiple trust-related questions. These values are controlled across age and gender covariates and are evaluated at an age value of 20.85 years and 0.60 gender (where male and female is represented by 0 and 1, respectively).

At this point, we would like to explore whether or not our hypotheses seem to be correct based on the observed data—namely, we can evaluate them based on two trust metrics as shown in the figures above. Our hypotheses are duplicated below, with explanations directly following them.
Hypothesis 1: An individual is more likely to trust a robot that apologizes for, rather than denies, a competence-based trust violation.

Our trust metric showed a statistically significant pairwise difference between the Competence Apology and Competence Denial conditions (p = 0.003), and Figure 7 above indicates that average trust in the robot is higher in the competence condition when the robot apologizes for the violation compared to when it denies it.

Hypothesis 2: An individual is more likely to trust a robot that denies, rather than apologizes for, an integrity-based trust violation.

Our trust metric did not show a statistically significant pairwise difference between the Integrity Apology and Integrity Denial conditions; however, Figure 7 above indicates that average trust in the robot is higher in the integrity condition when the robot denies the violation rather than when it apologizes for it.

Hypothesis 3: An individual is more likely to trust a robot that commits a competence-based trust violation over a robot that commits an integrity-based trust violation.

While the average trust of the robot metric is not statistically significant when comparing the competence and integrity conditions as a whole, as stated in the beginning of the “Experimental Results” section, we found a statistically significant difference on the participant’s Round 4 power-up choice between the competence and integrity conditions. The participant was more likely to immobilize the robot in the integrity-violation conditions than in the competence-violation conditions.

Further Analysis and Study

Despite the exciting results we found in running this study and analyzing this data, it is definitely the case that there exists a lot more work to be done and interesting questions to be explored. In our study, we held gender and age to be two of the covariates (controlling for these variables) and found statistically significant results. However, when looking at some of the participant surveys, some individuals rationalized their choice to not immobilize the robot by saying that they made a promise to the robot in response to its initial promise—in doing so, the participants may have had their power-up choice affected by their decision to promise the robot or not. Hence, we are considering running further studies that examine the role of the participant’s promise in their trust of the robot. An initial generalized linear model run with the participant promise as a covariate
indicates that not only are the existing significant values still significant, but also the participant’s choice to make a promise to the robot has a significant impact on their selection of a power-up. Additionally, one theory worth exploring is the idea that, once a participant retaliates against the robot for its immobilization choice (Round 3), he or she experiences a “catharsis” that results in future lack of immobilization and increased trust in the robot. Ideally, when considering real-world applications of trust violations and reparations, after a robot breaks and then attempts to repair your trust, a prolonged period of time could assist in slowly building back that trust. This could be further studied by looking into the differences in a participant’s power-up choice over the course of an interaction.

If this study were to be run again, one of the lessons learned throughout running participants (and reading survey responses) is that it would be beneficial to explicitly ask which power-up they viewed to be more helpful. While the game was designed so that the “asteroid blaster” would result in more points, several participants noted that they chose the “immobilization” power-up because they felt it was more effective in scoring points or beating their opponent. Asking this additional question (and recording the response metric) could be used to stratify individuals based on their perception of the power-up.

At the moment, the majority of the statistical analysis looked at the participants’ power-up choices and their quantifiable survey data. However, reading the free response can provide insight into each participant’s understanding of the interaction. It was through reading the free response, for example, that we recognized the fact that individuals had promised the robot as well! Upon initial evaluation, some participants assigned agency to the robot, commenting that their benevolent actions could encourage Echo to behave cooperatively in future rounds. This survey data, along with the interaction videos for each participant (many individuals spoke to the robot and vocally reacted to its actions and choices), provides a free-form source of data that has yet to be mined. Of course, it would also be beneficial to run a variety of models in addition to generative linear models, particularly algorithms that are more suited to deal with categorical data. All of these directions I find very interesting, and I hope to pursue them as we continue to work on analyzing our collected data.