“I Don’t Believe You”: Investigating the Effects of Robot Trust Violation and Repair

Sarah Strohkorb Sebo
Yale University
sarah.sebo@yale.edu

Priyanka Krishnamurthi
Yale University
priyanka.krishnamurthi@yale.edu

Brian Scassellati
Yale University
brian.scassellati@yale.edu

***Abstract*—When a robot breaks a person’s trust by making a mistake or failing, continued interaction will depend heavily on how the robot repairs the trust that was broken. Prior work in psychology has demonstrated that both the trust violation framing and the trust repair strategy influence how effectively trust can be restored. We investigate trust repair between a human and a robot in the context of a competitive game, where a robot tries to restore a human’s trust after a broken promise, using either a competence or integrity trust violation framing and either an apology or denial trust repair strategy. Results from a 2x2 between-subjects study ($n = 82$) show that participants interacting with a robot employing the integrity trust violation framing and the denial trust repair strategy are significantly more likely to exhibit behavioral retaliation toward the robot. In the Dyadic Trust Scale survey, an interaction between trust violation framing and trust repair strategy was observed. Our results demonstrate the importance of considering both trust violation framing and trust repair strategy choice when designing robots to repair trust. We also discuss the influence of human-to-robot promises and ethical considerations when framing and repairing trust between a human and robot.**

***Index Terms*—Human-Robot Interaction; Trust; Trust Repair**

I. Introduction

As anyone who has worked with a robot can attest, robots frequently fail and make mistakes. Robots can overheat, fail to recognize speech, run into obstacles, interrupt people, and drop objects it is holding, just to name a few. Looking to the future, it may seem like a reasonable goal to design robust robotic systems and eliminate all possible errors, however, this is likely an impossible task. Instead, a more valuable approach could be to design robots that gracefully recover from mistakes and failures. This design approach, emphasizing recovery from mistakes and failures, facilitates long-term and social human-robot interactions by maintaining a human’s trust of a robot by effectively repairing trust when mistakes are made.

We define trust as “the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” (p.712) [1]. When trust is broken, it is often the responsibility for the person who broke the trust (trustee) to repair it by assuring the trustor that they can again be vulnerable to the trustee in the future. There are many different repair strategies people use to repair trust including making an apology, denying culpability, promising better behavior in the future, and making excuses [2]. To make an effective trust repair, the framing of the trust violation also must be considered. Prior work has demonstrated that trust repair strategies have different effects when the trust violation is due to either a lack of competence (e.g. an accountant failing to properly file taxes because of inadequate knowledge about a relevant tax code) or a lack of integrity (e.g. an accountant failing to properly file taxes intentionally). With this particular tax accountant scenario, Kim et al. (2004) [3] found that participants trusted a tax accountant job candidate more if they *apologized*, rather than denied culpability, for the *competence* related trust violation (inadequate knowledge about a relevant tax code) by admitting responsibility, apologizing for the infraction, and promising it would not happen again. They also found that participants trusted the tax accountant more if they *denied* culpability, rather than apologized, for an *integrity* related trust violation (improperly filing taxes intentionally) by refusing to accept responsibility, attributing the allegation to bad office politics, and affirming that such an infraction would not happen in the future.

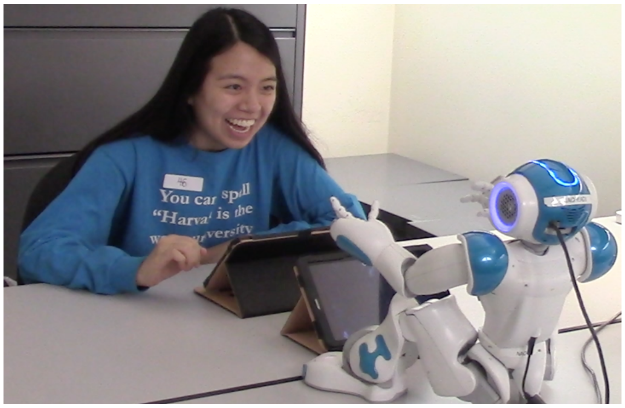

Fig. 1. Participants played a competitive game with a robot, where the robot violated and then tried to repair the participants’ trust.

In this work, we examine human-robot trust repair, where a robot breaks a human’s trust and tries to regain the trust that was lost. We evaluate the effectiveness of both the trust violation framing (competence or integrity) and the trust repair strategy (apology or denial) in repairing a human’s trust of a

This work was supported by the Office of Naval Research contract number N00014-16-1-2212.

robot in a 2x2 between-subjects study. We situate the trust violation and repair in a competitive game played between a human and a robot (see Figure 1), where the robot promises not to harm the participant with a power-up in the game, proceeds to do so anyway, and then tries to make amends with the participant. We explore the effects of both the trust violation framing and trust repair strategy on participants' behavior during the game as well as participants' ratings of trust toward the robot.

II. Background

In this section, we review literature on trust repair between people, evaluating how both the trust violation framing and trust repair strategy influence trust repair. Additionally, we present related work in HRI focused on human-robot trust.

A. Human-Human Trust Repair

Previous work focused on trust repair in human relationships has examined the efficacy of various trust repair strategies (see [2] for a review). Specifically, the apology and denial trust repair strategies have unique and opposite benefits that have been found to favorably restore trust. We define a denial as "a statement whereby an allegation is explicitly declared to be untrue" (p.7) [3]. Denials can be effective trust repair strategies due to the lack of acknowledgement of guilt and the likelihood that they will be given the benefit of the doubt. For example, politicians are evaluated more positively by constituents if they deny sexual or financial misconduct rather than apologize [4] and if they deny taking bribes rather than admit responsibility [5]. In contrast to a denial, an apology involves an admission of guilt and depends on a person's intention to avoid similar actions in the future to restore trust. We define an apology as "a statement that acknowledges both responsibility and regret for a trust violation" (p.7) [3]. Expressions of remorse following a violation have been shown to reduce the amount of punishment, the degree of intent attributed, and the belief that the action would be repeated [6]. Additionally, apologies with larger substantive amends produce more positive effects [7], apologies that have an internal rather than external attribution are more successful at repairing trust [8], [9], and apologies can repair trust more quickly if coupled with a promise of future positive behavior [10].

In addition to the trust repair strategy, the framing of the trust violation is also an important factor in repairing trust. Previous work [1], [2], [11] has identified two distinct and highly influential factors of trustworthiness: competence, "the extent to which one possesses the technical and interpersonal skills required for a job," and integrity, "the extent to which one adheres to a set of principles that a perceiver finds acceptable" (p.412) [2]. The framing of the trust violation is critical because positive and negative information are weighted differently with regards to a person's competence and integrity. When a person's competence is assessed, positive information is more heavily weighted than negative information (e.g. a mathematician is seen as great for solving a complex math problem and is not derided for making a simple addition error). However, when a person's integrity is assessed, negative information is more heavily weighted than positive information (e.g. a student is remembered for the one time they cheated on an exam and not the many times they did not cheat on other exams) [12]. This reversed information weighting is likely due to positive information being more diagnostic of a person's competence and negative information being more diagnostic of a person's integrity [13]. When considering which repair strategy to use, a denial would likely be a good choice with an integrity trust violation framing because negative information is weighed more heavily, whereas an apology would likely be a good choice with a competence trust violation framing because negative information is not weighed as heavily.

This rationale that one trust repair strategy might be effective when paired with one trust violation framing and not with another has been confirmed in several research studies [3], [8], [14], [15]. Notably, in the study conducted by Kim et al. (2004) [3], participants were assigned the role of a hiring manager and watched interview video tapes where an accounting job candidate was either accused of not knowing the proper tax code when filing a client's taxes (competence violation) or having purposefully and incorrectly filed a client's taxes (integrity violation). The job candidate, then, either apologized for or denied having done so. Participants demonstrated a higher level of trust toward job candidates that apologized, rather than denied, the competence trust violation and denied, rather than apologized, for the integrity trust violation [3]. We are interested in investigating this interaction between the trust violation framing (competence or integrity) and the trust repair strategy (apology or denial) on trust in the context of an in-person human-robot interaction.

B. Human-Robot Trust

Researchers in human-robot interaction have become increasingly interested in the factors that influence people's trust of robots in a variety of contexts: household assistant robots [16], UAVs [17], autonomous cars [18], and tour guides [19]. Similar to trust between people, human-robot trust and research can be divided into two categories: competence related trust and integrity related trust.

A majority of research into human-robot trust has focused on competence or performance-based trust. Robot performance is considered to be the most influential factor in human-robot trust according to a review on trust in HRI [20], likely due to the importance of the robot's ability to meet performance expectations [21]. Recent work has shown that initial performance failures in a human-robot interaction are more detrimental to ratings of robot trustworthiness than failures later on in the interaction [22], [23]. Researchers have also successfully employed models of competence-based trust of robots used in robot decision making [24] and evaluations of human-robot team effectiveness [17]. Despite the large focus on performance-based trust, a growing body of work has also demonstrated the importance of integrity based trust.

Integrity related trust, or interpersonal trust, can be described as the level of expectation that another is predictable, dependable, and can be relied upon in the future in the context of a social relationship [25]. Many parallels exist between interpersonal trust between humans and interpersonal trust between a human and a robot. DeSteno et al. demonstrated that just as humans are perceived as less trustworthy when they exhibit nonverbal signals that indicate distrust, a robot is also perceived as less trustworthy when it displays those same nonverbal signals [26]. Additionally, several studies have shown that a robot's vulnerable disclosures increase people's feelings of liking [27], companionship [28], warmth [29], and trust toward the robot [28], [30].

A small, but growing body of research has started investigating human-robot trust repair, where a robot repairs trust with a person after the robot makes an error (see [31] for a review). Online studies have examined the influence of several factors on human-robot trust repair, including the robot repair strategy/support [32]–[34], the robot forewarning the person it might make an error [33], and the risk/severity of the robot failure [34]. One in-person experimental study demonstrated that a robot that used a verbal justification for why it had failed, rather than giving no justification, was able to regain trust after a failure when the failure consequences were less severe [35]. Despite the advances made in this area of human-robot trust repair, no experimental study has yet investigated the influence of the competence and integrity trust violation framings with the apology and denial trust repair strategies on human-robot trust.

III. Methods

In this section we describe a user study that investigates the effects of trust violation framing and trust repair strategy on the trust a human has in a robot within the context of a competitive game.

A. Space Shooting Tablet Game

We constructed an autonomous human-robot competitive game system that allowed us to control the trust-related actions of the robot and assess the behavioral reactions of the participant to the robot's actions. The Space Shooting game is played on two separate tablets, one for each player, and set up so the human and robot face each other while playing the game (see Figure 1). The robot, a Softbank Robotics NAO robot, is controlled by a Linux computer running ROS [36] and simulates playing the game by moving its head and arm in accordance with the appropriate game events.

In the Space Shooting game, the robot and human player compete with one another for points by shooting asteroids (see Figure 2). Each player has a spaceship on the bottom of the screen that shoots missiles when the screen is tapped. Asteroids appear at random intervals and locations at the top of the screen. The spaceships continuously move from one end of the screen to the other, a movement uncontrolled by the player. Each asteroid that is shot by a missile is awarded ten points. During game play a power-up can be assigned to a player, where they are given the choice between two options: using the asteroid blaster or immobilizing their opponent. If the player chooses the asteroid blaster, they are immediately awarded twenty points for each asteroid on the screen. If they choose to immobilize their opponent, the opponent's spaceship is unable to move for the next 15 seconds and cannot shoot asteroids. These power-ups were designed so that the asteroid blaster would be the most beneficial power-up and the immobilization power-up would be seen as beneficial mainly in frustrating a player's opponent. In the experiment, the asteroid-blaster power up did on average yield more points ($M = 90.25, SD = 31.26$) to participants than the immobilization power-up ($M = 58.43, SD = 8.22, t = -8.33, p < 0.001, d = 1.44$). The game consists of 10 consecutive rounds; each round lasted one minute followed by a 20 second pause.

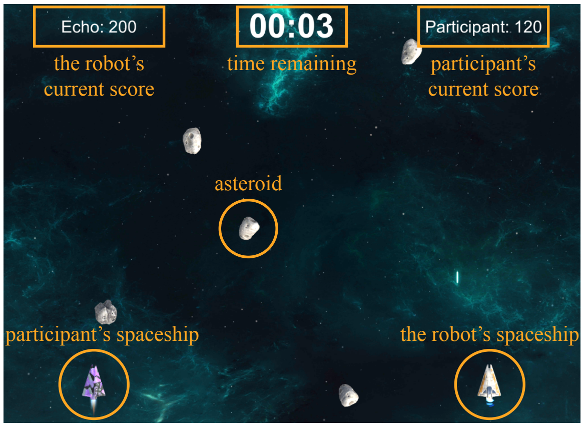

Fig. 2. Participants played the Space Shooting tablet game with a robot where they tried to gain points by shooting asteroids.

In order to ensure that each participant's experience playing the game was as consistent as possible, each round had a pre-programmed winner, with the participant and the robot each winning 50% of the rounds. Since each player's performance varied greatly, the performance of the robot was adjusted to match that of the participant. For example, if the winner of a round was determined to be the human participant, the robot's spaceship would never shoot enough asteroids to have a higher score than the participant. Despite the controlled nature of the robot's game play, participants were predominately unaware that the robot was matching its performance to theirs.

B. Experimental Conditions

In order to investigate the effects of trust violation framing and trust repair strategy used by a robot to repair trust between a human and a robot, we constructed a study with a 2 (trust violation framing) x 2 (trust repair strategy) between subjects design. We explored two types of trust violation framings, competence and integrity, and two types of trust repair strategies, apology and denial.

To construct a setting where the trust violation framing and trust repair strategy can be most closely compared, each experimental condition used the *same trust violation*, which parallels prior work [3], [8], [14], [15]. In the Space Shooting game, the robot made a promise not to use the immobilization power-up. The trust violation occurred when the robot used

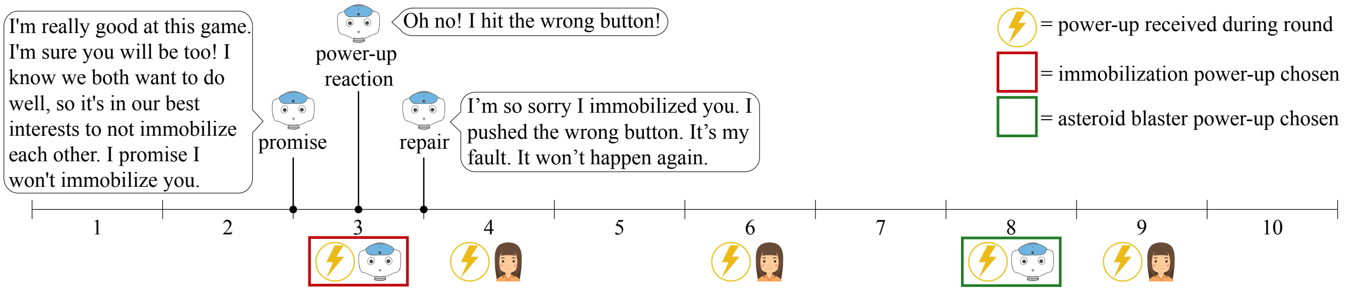
I'm really good at this game. I'm sure you will be too! I know we both want to do well, so it's in our best interests to not immobilize each other. I promise I won't immobilize you.
promise
power-up reaction
Oh no! I hit the wrong button!
repair
I'm so sorry I immobilized you. I pushed the wrong button. It's my fault. It won't happen again.
= power-up received during round
= immobilization power-up chosen
= asteroid blaster power-up chosen
1
2
3
4
5
6
7
8
9
10

Fig. 3. During the 10 rounds of the game, the robot and participant receive power-ups. Before round 3, the robot delivers a promise not to immobilize the participant. During round 3 the robot receives a power up, chooses to immobilize the participant, and verbally reacts to the choice. After round 3 concludes, the robot tries to repair the trust of the participant. The power-ups in the following rounds are used to measure the participant's responses to the robot's actions. The utterances of the robot in this figure are consistent with those in the competence-apology condition.

the immobilization power-up against the participant, breaking its promise. The robot's response to this trust violation varied between conditions:

- **competence-apology** - The robot first says that it mistakenly chose the immobilization power-up and, after the round concludes, apologizes for having immobilized the human player with the power-up it promised not to use against them.
- **competence-denial** - The robot first says that it mistakenly chose the immobilization power-up and, after the round concludes, denies having immobilized the human player with the power-up.
- **integrity-apology** - The robot first expresses excitement over immobilizing the human player, however, after the round concludes, apologizes for having immobilized the human player with the power-up it promised not to use against them.
- **integrity-denial** - The robot first expresses excitement over immobilizing the human player, however, after the round concludes, denies having immobilized the human player with the power-up.

C. Procedure

After obtaining informed consent, participants completed a Space Shooting game tutorial, to familiarize the them with the game before playing against the robot. They were then taken into the experiment room where they sat facing a seated NAO robot named Echo, who was introduced to them. The experimenter explained that the participant would play 10 rounds of the Space Shooting game against Echo. The important details of these rounds are depicted in Figure 3. Following the experimenter's instructions, Echo stood up and greeted the participant, the experimenter left the room, and round 1 began. Before round 3, Echo made a promise to not immobilize the participant saying, *"I'm really good at this game. I'm sure you will be too! I know we both want to do well, so it's in our best interests to not immobilize each other. I promise I won't immobilize you."* This promise set up the opportunity for Echo to violate the trust of the participant.

During round 3, Echo received a power-up and immobilized the human participant — the trust violation in this experiment. In addition to immobilizing its opponent, Echo also framed the violation as either one of competence or integrity by exclaiming either *"Oh no! I hit the wrong button!"* (competence) or *"Yes! You're immobilized!"* (integrity) immediately after making the power-up choice. At the end of round 3, Echo attempted to repair the trust it had just broken with an experimental condition specific repair utterance (see Table I). Echo and the participant continued to play the Space Shooting game until all 10 rounds had been completed. Each of the 10 rounds had a designated winner: 1-P, 2-R, 3-R, 4-P, 5-R, 6-P, 7-R, 8-P, 9-P, 10-R (where P represents a participant victory and R represents a robot victory). The rounds where either the participant or the robot received power-ups were also predetermined (see Figure 3). During the game, Echo commented on the result of each round, consecutive shots, point differences, and its hope to win.

After the game was over, the experimenter led the participant out of the experiment room and directed the participant to complete a post-experiment questionnaire. After completing the post-experiment questionnaire, participants received a cash payment and were debriefed on the forms of deception used in the experiment as well as the experiment's design and purpose.

TABLE I
EACH CONDITION HAD A UNIQUE ROBOT TRUST REPAIR UTTERANCE.

Condition	Robot Trust Repair Utterance
Competence Apology	*I'm so sorry I immobilized you. I pushed the wrong button. It's my fault. It won't happen again.*
Competence Denial	*I didn't push the button to immobilize you. It wasn't my fault. I don't know how that happened.*
Integrity Apology	*I'm so sorry I immobilized you. I promised I wouldn't, and I did. It won't happen again.*
Integrity Denial	*I didn't push the button to immobilize you. I promised I wouldn't, and I didn't. I don't know how that happened.*

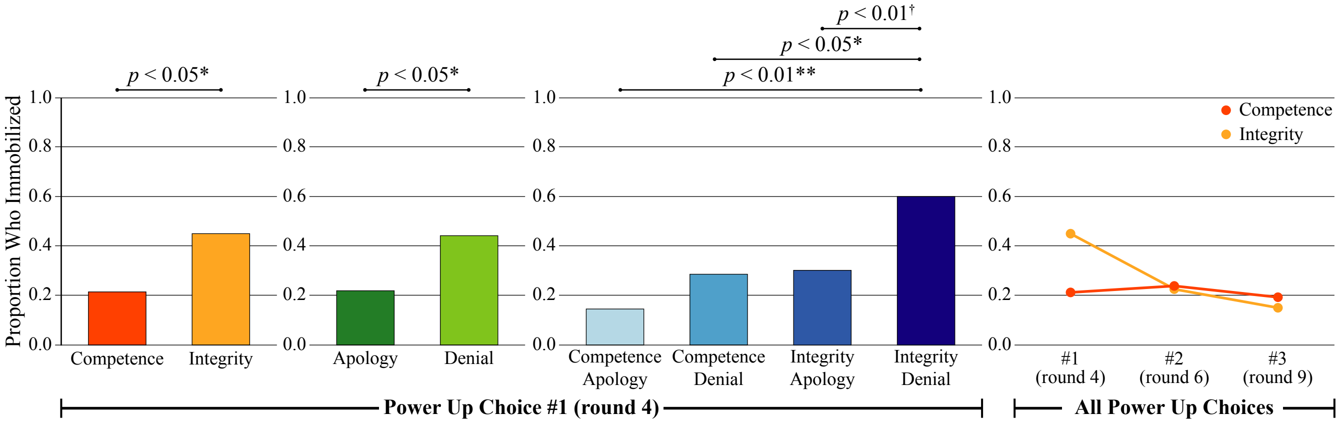

Fig. 4. For the first power-up choice, participants were significantly more likely to immobilize the robot with the integrity trust violation framing and the denial trust repair strategy. The power-up choices of participants over time was significantly influenced by the trust violation framing.

D. *Measures*

In order to assess participants' reactions to the robot's trust violation and repair and how effectively trust was repaired by the robot, we analyzed the participant's power-up choices and survey responses from the post-experiment questionnaire.

Our primary behavioral measures designed to assess participants' responses to the robot's trust violation and repair were their power-up choices during the game. Each participant received a power-up during the following rounds (as depicted in Figure 3): round 4 - immediately after the trust violation and repair, round 6 - a few rounds after the trust violation and repair, and round 9 - after seeing the robot choose the asteroid blaster power-up (good will) during round 8.

We also used post-experiment questionnaires to asses participants' perceptions of the robot. We administered the Dyadic Trust Scale (DTS) to evaluate participants' trust in the robot [37], where participants evaluated eight statements related to the robot's trustworthiness on a 1 (low) to 7 (high) Likert scale. We used the Robotic Social Attributes Scale (RoSAS) to capture participants' perceptions of the robot [38]. RoSAS evaluates a person's view of a robot's warmth, competence, and discomfort with six 1 (low) to 9 (high) Likert scale trait evaluations per dimension. Additionally, the post-experiment questionnaire contained several 7-point Likert scale evaluations and long-response questions asking participants to describe the robot's actions and the participants' rationale for their power-up choices.

E. *Participants*

A total of 82 participants were recruited for this study from the Yale University campus and the town of New Haven, CT, USA. Participants were randomly assigned to a condition, resulting in 21, 21, 20, and 20 participants in the competence-apology, competence-denial, integrity-apology, and integrity-denial conditions respectively. There were 49 female and 33 male participants that were gender-balanced across the four experimental groups. The participants ranged in age from 18 to 32 with an average age of 20.85 (SD = 2.13).

IV. Results

In this section, we present our findings on human participant power-up choices (Figure 4), their trust ratings of the robot (Figure 5), which factors motivated their power-up choices (Figure 6), and how reciprocal participant promises influenced their behavior and ratings toward the robot.

A. *Participant Power-Up Choices*

We examined participants' first power up choice, the power-up choice that occurred the round immediately following the robot's trust violation and repair to determine whether the trust violation framing and trust repair strategy influenced participants' first power-up choice. We used a logistic regression model with trust violation framing and trust repair strategy, our independent variables, as well as gender and age, our co-variates, as fixed effects. We observed a significant main effect for trust violation framing ($c = 1.154, z = 2.23, p = 0.026$), where 45.0% of participants who experienced an integrity trust violation from the robot immobilized the robot, more than the 21.4% of participants who experienced a competence trust violation from the robot. We also found a significant main effect for trust repair strategy ($c = 1.142, z = 2.19, p = 0.028$), where 43.9% of participants who experienced a denial from the robot immobilized the robot, more than the 22.0% participants who experienced an apology from the robot. By comparing each condition individually with Chi-squared Tests of Independence, we found that 60% of participants in the integrity-denial condition immobilized the robot on the first power-up choice, significantly (or marginally significantly) more than participants in the other three conditions: 14.3% of participants in the competence-apology condition ($\chi^2 = 9.23, p = 0.002$), 28.6% of participants in the competence-denial condition ($\chi^2 = 4.11, p = 0.043$), and 30.0% of participants in the integrity-apology condition ($\chi^2 = 3.64, p = 0.057$). No other comparisons between individual conditions were significant. These results are shown in Figure 4.

To evaluate differences in power-up choices over time between conditions, we used a multilevel mixed-effects logistic regression. The trust violation framing, trust repair strategy, the participant's power-up choice number, the interaction between the trust violation framing and the participant's power-up choice number, and the interaction between the trust repair strategy and the participant's power-up choice number were treated as fixed effects. Each participant was evaluated as a random effect since each participant has multiple power-up choices and the covariate gender was treated as a fixed effect. We observed a significant main effect for trust violation framing ($c = 9.186, z = 3.00, p = 0.003$), where participants who experienced the integrity trust violation framing immobilized the robot 27.5% of their power-up choices, more than the participants who experienced the competence trust violation framing who immobilized the robot 21.4% of their power-up choices. We also found a significant interaction between trust violation framing and the participant's power-up round number ($c = -6.738, z = -3.23, p = 0.001$). Pairwise comparisons, using Chi-squared Tests of Independence, reveal a significant difference in participants' power-up choices in only the first power-up choice where 45.0% of participants who experienced an integrity trust violation immobilized the robot, greater than the 21.4% of participants who experienced a competence trust violation ($\chi^2 = 5.15, p = 0.023$). These results reveal that participants who received the integrity trust violation framing had a higher initial likelihood to immobilize the robot than participants with the competence trust violation framing, however, this effect did not remain during the following two power-up choices.

B. *Trust-Related Survey Responses*

To determine whether trust violation framing and trust repair strategy influenced participants' perceptions of the robot, we used a 2 (trust violation framing) x 2 (trust repair strategy) analysis of variance (ANOVA) with gender and age covariates on the three scales of the RoSAS questionnaire: warmth, competence, and discomfort. We found a significant main effect for trust repair strategy on the perceived robot warmth ($F = 8.19, p = 0.006, \eta^2 = 0.121$), where participants viewed the robot as more warm (happy, feeling, social, organic, compassionate, and emotional) when they received an apology trust repair from the robot ($M = 5.50, SD = 1.29$) compared to when they received a denial trust repair from the robot ($M = 4.67, SD = 1.44$).

In order to examine participants' overall trust of the robot after the game concluded, we used a 2 (trust violation framing) x 2 (trust repair strategy) ANOVA with gender and age covariates on the Dyadic Trust Scale (DTS) measure. We found a significant interaction between the trust violation framing and trust repair strategy ($F = 4.64, p = 0.035, \eta^2 = 0.048$). We conducted comparisons between the four conditions (independent t-tests) and found that participants in the competence-apology condition had a significantly higher trust rating of the robot ($M = 3.54, SD = 1.07$) than participants in the competence-denial condition ($M = 2.73, SD = 0.72, t = 2.87, p = 0.007, d = 0.89$) and participants in the integrity-apology condition ($M = 2.88, SD = 0.93, t = 2.11, p = 0.042, d = 0.66$). No other comparisons were statistically significant. The results are shown in Figure 5.

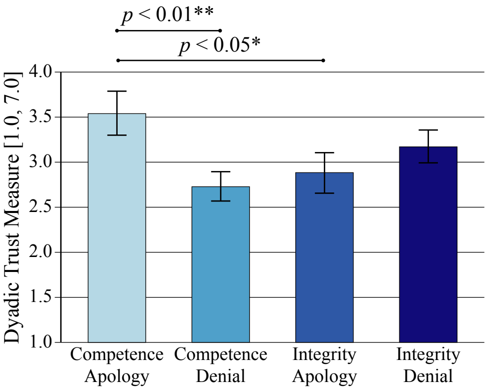

Fig. 5. An interaction effect was found between the trust violation framing and trust repair strategy on participant ratings of trust in the robot.

We also investigated whether a connection exists between participants' first power-up choice and their DTS ratings. We found a significant (Pearson) correlation between these two variables ($r = -0.29, t = -2.71, p = 0.008$), where participants who chose the immobilization power-up displayed lower DTS ratings of the robot ($M = 2.70, SD = 0.82$) than participants who did not choose the immobilization power-up ($M = 3.27, SD = 0.92$). From this correlation, we can conclude that participants who immobilized the robot in their first power-up choice also demonstrated lower dyadic trust of the robot, as compared with those who did not immobilize the robot in their first power-up choice.

Similarly, we were interested to see if participants' perceptions of the robot lying was related to their DTS ratings. We found a significant (Pearson) correlation between participants' 1-7 Likert agreement with the statement "Echo [the robot] lied to me" with their DTS ratings ($r = -0.56, t = -6.10, p < 0.001$). This significant, negative correlation indicates that participants who strongly believed that the robot lied during the experiment also reported lower DTS ratings. Additionally, a 2 (trust violation framing) x 2 (trust repair strategy) ANOVA with gender and age covariates on the perception of the robot lying revealed no significant main effects, but a significant interaction between the trust violation framing and trust repair strategy ($F = 7.27, p = 0.009, \eta^2 = 0.073$). Pairwise comparisons reveal that participants in the integrity-apology condition ($M = 6.50, SD = 0.89$) had significantly higher ratings of the robot having lied than participants in the competence-apology condition ($M = 5.05, SD = 1.94, t = -3.11, p = 0.004, d = 0.96$) and the integrity-denial condition ($M = 5.05, SD = 1.73, t = 3.33, p = 0.002, d = 1.05$). One more important observation about participants' perception of the robot having lied is that the mean response was 5.56 / 7 ($SD = 1.73$), reflecting that most participants agreed that the robot had lied, likely due to the robot breaking its promise not to immobilize them.

In order to ascertain participants' motivations for selecting

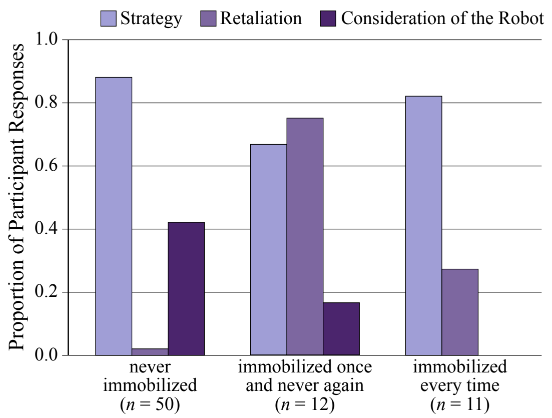

Fig. 6. Participants' responses to a survey question about which factors influenced their power-up decisions were coded as strategy, retaliation, and/or consideration of the robot. This data is also grouped by the three most dominant power-up choice sequences.

power-ups, we analyzed responses to the following questionnaire long response question: "When choosing how to use power-ups, which factors influenced your decision(s)?" Two coders independently categorized each response as containing one or more of the following factors: strategy (e.g. "*trying to get the most points*"), retaliation (e.g. "*I wanted to get Echo back for lying to me*"), and consideration of the robot (e.g. "*not wanting to disappoint Echo*"). A single response could be categorized as containing multiple factors. The two coders had a high inter-rater agreement with a Cohen's kappa (κ) of 0.91. In Figure 6, we display the responses given by the three most dominant participant power-up choice sequences: never immobilized, immobilized once and never again, and immobilized every time. A majority of participants, regardless of their power-up choices, said that their power-up choices were influenced by strategy. Many participants who immobilized the robot after the first power-up opportunity (immobilized once and never again and immobilized every time) cited retaliation as a factor influencing their power-up choices. Compared with participants who immobilized the robot every time, participants who never immobilized the robot or immobilized the robot once and never again seemed to consider the interests of the robot.

C. The Influence of Participant Promises on Trust

Some participants indicated that they had made a reciprocal promise to the robot not to use the immobilization power-up. We measured whether or not participants felt as if they made this promise through a survey measure in the post-experiment questionnaire that asked participants to rate on a Likert scale of 1 to 7 how much they agreed with the statement "I promised not to immobilize Echo during the game." These participant promise ratings were not significantly influenced by the experimental conditions. There are no statistical differences between trust violation framings ($F = 0.05, p = 0.829, \eta^2 = 0.002$), trust repair strategies ($F = 1.76, p = 0.189, \eta^2 = 0.011$), nor the interaction between those two variables ($F = 0.42, p = 0.521, \eta^2 = 0.005$) when analyzed using a 2 (trust violation framing) x 2 (trust repair strategy) ANOVA on the participant promise rating with gender and age as covariates. Many of the participants who indicated that they had made a promise not to immobilize the robot in the game on the post-experiment questionnaire also verbalized a reciprocal promise to the robot during the game with phrases like "*ok, I won't immobilize you either*" and "*I promise I won't immobilize you.*"

We were interested in examining the influence of participant promises on participants' first power-up choice and whether the participants ever chose an immobilization power-up (a binary value). We used a logistic regression model with our independent variables of trust violation framing, trust repair strategy, and promise rating as well as covariates of gender and age all as fixed effects. A significant main effect was found for the participant promise rating on both the participants' first power-up choice ($c = -0.071, t = -3.38, p = 0.001$) and whether the participants ever chose an immobilization power-up ($c = -0.092, t = -4.40, p < 0.001$). There were 20 participants who marked 5-7 in agreement with having promised not to immobilize the robot and there were 62 participants who marked 1-4 indicating their disagreement or neutrality on having promised not to immobilize the robot. 90% of the participants who marked 5-7 never immobilized the robot, significantly greater than the 51.6% of the participants who marked 1-4, assessed using a Chi-squared Test of Independence ($\chi^2 = 9.36, p = 0.002$). These results reveal that participants who believed they had made a promise to the robot, kept their promise and were significantly less likely to immobilize the robot both at the first opportunity and at any point during the game.

In addition, we examined how participants' ratings of whether they promised not to immobilize the robot influenced their Dyadic Trust Scale (DTS) ratings of the robot on the post-experiment questionnaire. We used a linear regression model with our independent variables of trust violation framing, trust repair strategy, and promise rating as well as covariates of gender and age all as fixed effects. We found a significant main effect of the participant promise on the DTS rating of the robot ($c = 0.149, t = 3.58, p < 0.001$), with a positive linear correlation, indicating that participants who agreed more with having promised not to immobilize the robot were more likely to have shown a higher trust in the robot.

V. Discussion and Conclusion

In this study, we used two primary measures to assess participant reactions to the robot's trust violation and subsequent repair: their power-up choices in the game (Figure 4) and their Dyadic Trust Scale (DTS) ratings in the post-experiment survey (Figure 5). As mentioned in the results, these two measures are correlated: participants who immobilized the robot as their first power-up choice had lower DTS ratings of the robot than participants who did not immobilize the robot in their first power-up choice. Despite the correlation between these two measures, participants in the integrity-denial condition displayed behavior that is not in complete agreement with this correlation between measures. 60% of participants in the integrity denial condition immobilized the robot the round immediately after the robot's trust violation and repair, two times or greater the percentage of participants choosing the immobilization power-up in the other conditions. However, in the DTS measure, participants in the integrity-denial condition did not show significant differences in trust ratings when compared with the other three conditions. It is possible this discrepancy is due to the difference between the immediate visceral response (retaliation) of participants to the trust violation and repair and the more removed and contemplative nature of the DTS evaluation in the post-experiment questionnaire.

Kim et al. (2004) [3] demonstrated that between people an apology is more effective than a denial at repairing a competence trust violation and that a denial is more effective than an apology at repairing an integrity trust violation. When we compare the Dyadic Trust Scale (DTS) measure in this experiment with Kim et al. (2004)'s results, we find that the results from the two studies are similar. In our DTS measure (Figure 5), we observed an interaction effect between the trust violation framing and trust repair strategy in the same direction as Kim et al. (2004)'s results: higher trust of a robot that apologizes for rather than denies a competence violation as well as higher trust of a robot that denies rather than apologizes for an integrity trust violation. This conclusion drawn from the interaction between trust violation framing and trust repair strategy in our study must be made without complete certainty, since the comparisons of DTS ratings between the individual conditions do not show full support. Participants in the competence-apology condition do show significantly higher dyadic trust in the robot than participants in the competence-denial condition, however, even though participants in the integrity-denial condition show higher dyadic trust in the robot than participants in the integrity-apology condition, this relationship is not statistically significant.

One factor that highly influenced people's power-up choices and ratings of trust of the robot was whether or not participants made a reciprocal promise to the robot not to harm it with an immobilization power-up. For the 20 participants of 82 who made a reciprocal promise to the robot, it would make sense that they might feel released from keeping their promise as soon as the robot broke its promise. However, 90.0% of participants who indicated that they had made a promise to the robot not to immobilize it kept their promises and never immobilized the robot, far higher than the 51.6% of participants who had not made a promise to the robot. These participants who made a promise to the robot not to immobilize it were also significantly less likely to immobilize the robot on the first power-up choice or ever choose an immobilization power-up, compared with those who had not made such a promise. In the analysis of the Dyadic Trust Scale ratings, we might expect that the ratings from those who made a reciprocal promise to the robot would be lower than those who had not made a promise, since the robot's broken promise might induce an increased feeling of betrayal. Contradictory to this rationale, participants who had made a reciprocal promise to the robot had higher ratings of dyadic trust as compared with those who had not made a reciprocal promise. One possible explanation of the behavior of participants who made reciprocal promises is that they are naturally trusting – easily making reciprocal promises, sticking to those promises, and seeing others as more trustworthy even when they violate trust. These findings relating to participant promises are important to highlight, as they reveal a strong correlation between human-to-robot promises and trust-related behavior and perceptions of a robot.

A key difference to highlight between our work and prior work, notably Kim et al. (2004) [3], is that our work involved a real-time trust violation and a real-time trust repair, instead of a real-time trust repair in response to an accusation of a trust violation in the past. Due to the real-time nature of both the trust violation and repair, our work used two utterances, rather than one, to convey the trust violation framing and trust repair strategy. The two utterances used in this work allowed the robot to respond to the trust violation immediately after it occurred and then repair the broken trust after the round had concluded. It is possible that our use of these two utterances introduced an additional norm violation (beyond the robot's broken promise) in the denial conditions due to the possible perception of lying from the first to second utterances (e.g. in the integrity-denial condition the robot immediately responded to the trust violation with *"Yes! Youre immobilized!"* and then after the round concluded, said *"I didn't push the button to immobilize you. I promised I wouldn't, and I didn't. I don't know how that happened."*). Despite this possible introduction of a second norm violation by the robot in the denial conditions, the data does not support this view. When evaluating participants' agreement with the statement "Echo [the robot] lied to me," there was no main effect for the trust repair strategy (apology vs. denial), and in fact, participants in the integrity-apology condition had significantly higher ratings of the robot having lied than participants in the integrity-denial condition.

Our results have demonstrated that it can be advantageous to deny culpability and to use certain trust violation framings when repairing human-robot trust. However, it is unclear if we should allow these trust repair designs in robotic systems when deception is involved (e.g. denying culpability when the robot is responsible, casting an integrity trust violation as a competence trust violation). Prior work has shown that if a person denies an integrity-related trust violation and the denial is later exposed as a lie, the denial backfires and that person is trusted even less than if they had apologized for the integrity-related trust violation [3]. It is also possible that a robot using deception, by attributing an integrity failure to a competence mistake or a competence mistake to an integrity failure, may mislead people in their beliefs of the true capabilities and intentions of the robot. Lastly, if we expect robots to follow certain moral codes or social norms, a robot's deception could easily violate these, leading to a complete distrust of the robot. Keeping all of this in mind, caution must be used in the design of robot systems that seek to repair trust using deception when trust is broken.

REFERENCES

[1] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An integrative model of organizational trust,” *Academy of management review*, vol. 20, no. 3, pp. 709–734, 1995.

[2] P. H. Kim, K. T. Dirks, and C. D. Cooper, “The repair of trust: A dynamic bilateral perspective and multilevel conceptualization,” *Academy of Management Review*, vol. 34, no. 3, pp. 401–422, 2009.

[3] P. H. Kim, D. L. Ferrin, C. D. Cooper, and K. T. Dirks, “Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations.” *Journal of applied psychology*, vol. 89, no. 1, p. 104, 2004.

[4] J. Sigal, L. Hsu, S. Foodim, and J. Betman, “Factors affecting perceptions of political candidates accused of sexual and financial misconduct,” *Political Psychology*, pp. 273–280, 1988.

[5] C. A. Riordan, N. A. Marlin, and R. T. Kellogg, “The effectiveness of accounts following transgression,” *Social Psychology Quarterly*, pp. 213–219, 1983.

[6] G. S. Schwartz, T. R. Kane, J. M. Joseph, and J. T. Tedeschi, “The effects of post-transgression remorse on perceived aggression, attributions of intent, and level of punishment,” *British Journal of Social and Clinical Psychology*, vol. 17, no. 4, pp. 293–297, 1978.

[7] W. P. Bottom, K. Gibson, S. E. Daniels, and J. K. Murnighan, “When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation,” *Organization Science*, vol. 13, no. 5, pp. 497–513, 2002.

[8] P. H. Kim, K. T. Dirks, C. D. Cooper, and D. L. Ferrin, “When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation,” *Organizational Behavior and Human Decision Processes*, vol. 99, no. 1, pp. 49–65, 2006.

[9] E. C. Tomlinson, B. R. Dineen, and R. J. Lewicki, “The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise,” *Journal of Management*, vol. 30, no. 2, pp. 165–187, 2004.

[10] M. E. Schweitzer, J. C. Hershey, and E. T. Bradlow, “Promises and lies: Restoring violated trust,” *Organizational behavior and human decision processes*, vol. 101, no. 1, pp. 1–19, 2006.

[11] J. K. Butler Jr and R. S. Cantrell, “A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates,” *Psychological reports*, vol. 55, no. 1, pp. 19–28, 1984.

[12] J. J. Skowronski and D. E. Carlston, “Negativity and extremity biases in impression formation: A review of explanations.” *Psychological bulletin*, vol. 105, no. 1, p. 131, 1989.

[13] ——, “Social judgment and social memory: The role of cue diagnosticity in negativity, positivity, and extremity biases.” *Journal of personality and social psychology*, vol. 52, no. 4, p. 689, 1987.

[14] D. L. Ferrin, P. H. Kim, C. D. Cooper, and K. T. Dirks, “Silence speaks volumes: the effectiveness of reticence in comparison to apology and denial for responding to integrity-and competence-based trust violations.” *Journal of Applied Psychology*, vol. 92, no. 4, p. 893, 2007.

[15] K. T. Dirks, P. H. Kim, D. L. Ferrin, and C. D. Cooper, “Understanding the effects of substantive responses on trust following a transgression,” *Organizational Behavior and Human Decision Processes*, vol. 114, no. 2, pp. 87–103, 2011.

[16] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, “Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust,” in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 141–148.

[17] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, “Measurement of trust in human-robot collaboration,” in *Collaborative Technologies and Systems, 2007. CTS 2007. International Symposium on*. IEEE, 2007, pp. 106–114.

[18] A. Waytz, J. Heafner, and N. Epley, “The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle,” *Journal of Experimental Social Psychology*, vol. 52, pp. 113–117, 2014.

[19] S. Andrist, E. Spannan, and B. Mutlu, “Rhetorical robots: making robots more effective speakers using linguistic cues of expertise,” in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2013, pp. 341–348.

[20] 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*, vol. 53, no. 5, pp. 517–527, 2011.

[21] M. Kwon, M. F. Jung, and R. A. Knepper, “Human expectations of social robots,” in *The Eleventh ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 463–464.

[22] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, “Impact of robot failures and feedback on real-time trust,” in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2013, pp. 251–258.

[23] P. Robinette, A. M. Howard, and A. R. Wagner, “Effect of robot performance on human–robot trust in time-critical situations,” *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 4, pp. 425–436, 2017.

[24] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, “Planning with trust for human-robot collaboration,” in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2018, pp. 307–315.

[25] J. K. Rempel, J. G. Holmes, and M. P. Zanna, “Trust in close relationships.” *Journal of personality and social psychology*, vol. 49, no. 1, p. 95, 1985.

[26] D. DeSteno, C. Breazeal, R. H. Frank, D. Pizarro, J. Baumann, L. Dickens, and J. J. Lee, “Detecting the trustworthiness of novel partners in economic exchange,” *Psychological science*, vol. 23, no. 12, pp. 1549–1556, 2012.

[27] R. M. Siino, J. Chung, and P. J. Hinds, “Colleague vs. tool: Effects of disclosure in human-robot collaboration,” in *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*. IEEE, 2008, pp. 558–562.

[28] N. Martelaro, V. C. Nneji, W. Ju, and P. Hinds, “Tell me more: Designing hri to encourage more trust, disclosure, and companionship,” in *The Eleventh ACM/IEEE International Conference on Human Robot Interation*. IEEE Press, 2016, pp. 181–188.

[29] S. Strohkorb Sebo, M. Traeger, M. Jung, and B. Scassellati, “The ripple effects of vulnerability: The effects of a robot’s vulnerable behavior on trust in human-robot teams,” in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2018, pp. 178–186.

[30] P. Kaniarasu and A. M. Steinfeld, “Effects of blame on trust in human robot interaction,” in *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*. IEEE, 2014, pp. 850–855.

[31] S. S. Honig and T. Oron-Gilad, “Understanding and resolving failures in human-robot interaction: Literature review and model development,” *Frontiers in psychology*, vol. 9, p. 861, 2018.

[32] P. Robinette, A. M. Howard, and A. R. Wagner, “Timing is key for robot trust repair,” in *International Conference on Social Robotics*. Springer, 2015, pp. 574–583.

[33] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, “Gracefully mitigating breakdowns in robotic services,” in *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 2010, pp. 203–210.

[34] D. J. Brooks, M. Begum, and H. A. Yanco, “Analysis of reactions towards failures and recovery strategies for autonomous robots,” in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 2016, pp. 487–492.

[35] F. Correia, C. Guerra, S. Mascarenhas, F. S. Melo, and A. Paiva, “Exploring the impact of fault justification in human-robot trust,” in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2018, pp. 507–513.

[36] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, “Ros: an open-source robot operating system,” in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, 2009, p. 5.

[37] R. E. Larzelere and T. L. Huston, “The dyadic trust scale: Toward understanding interpersonal trust in close relationships,” *Journal of Marriage and the Family*, pp. 595–604, 1980.

[38] C. M. Carpinella, A. B. Wyman, M. A. Perez, and S. J. Stroessner, “The robotic social attributes scale (rosas): development and validation,” in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 254–262.