"What If It Is Wrong": Effects of Power Dynamics and Trust Repair Strategy on Trust and Compliance in HRI

Ulas Berk Karli*
ukarli1@jhu.edu
Johns Hopkins University
Baltimore, MD, USA

Shiye Cao*
scao14@jhu.edu
Johns Hopkins University
Baltimore, MD, USA

Chien-Ming Huang
cmhuang@cs.jhu.edu
Johns Hopkins University
Baltimore, MD, USA

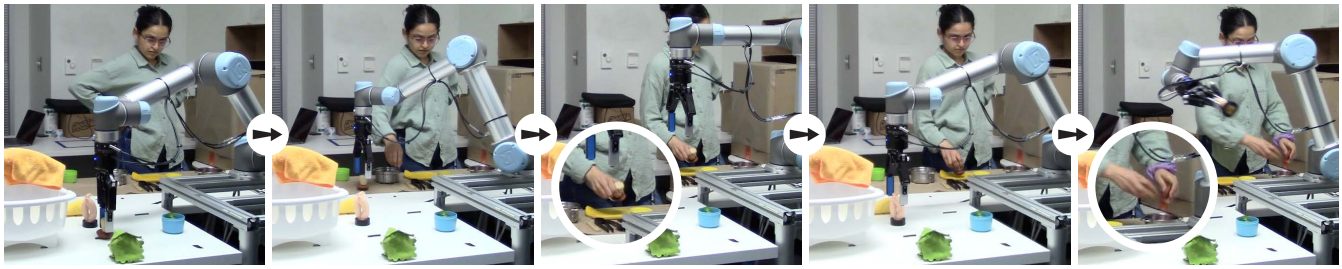

Figure 1: We study the effects of power dynamics and trust repair strategy on repairing user trust in the robot after a technical robot error. This figure illustrates an example of a user complying with the supervisor robot and cooking with the incorrect ingredient, even though they noticed the robot error.

ABSTRACT

Robotic systems designed to work alongside people are susceptible to technical and unexpected errors. Prior work has investigated a variety of strategies aimed at repairing people's trust in the robot after its erroneous operations. In this work, we explore the effect of post-error trust repair strategies (promise and explanation) on people's trust in the robot under varying power dynamics (supervisor and subordinate robot). Our results show that, regardless of the power dynamics, promise is more effective at repairing user trust than explanation. Moreover, people found a supervisor robot with verbal trust repair to be more trustworthy than a subordinate robot with verbal trust repair. Our results further reveal that people are prone to complying with the supervisor robot even if it is wrong. We discuss the ethical concerns in the use of supervisor robot and potential interventions to prevent improper compliance in users for more productive human-robot collaboration.

CCS CONCEPTS

• **Human-centered computing**; • **Computer systems organization** → **Robotics**;

KEYWORDS

trust repair, human-robot collaboration, human-robot trust, human-robot power dynamics

ACM Reference Format:
Ulas Berk Karli, Shiye Cao, and Chien-Ming Huang. 2023. "What If It Is Wrong": Effects of Power Dynamics and Trust Repair Strategy on Trust and Compliance in HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3568162.3576964

1 INTRODUCTION

Robots are envisioned to take on different roles as supervisors (*e.g.,* [2, 19, 24, 30]), peers (*e.g.,* [19]), or subordinates (*e.g.,* [19]), to enable complex human-robot collaboration. In such interactions, robot errors are inevitable due to imperfect technology (*e.g.,* uncertainty in visual perception and intent recognition in speech) and unexpected events (*e.g.,* context shifts and disturbance in the environment). These errors damage productive collaboration between humans and robots and erode people's trust in the robot [5, 39].

To repair the eroded trust, prior work has investigated a range of trust repair strategies including apology, explanation, and promise [13], and explored how factors, such as anthropomorphism [14] and human attitude [12], may modulate the effectiveness of these strategies. However, little is known about how power dynamics—supervisor robots versus subordinate robots—might shape the efficacy of trust repair strategies. Would people react differently to errors from a supervisor robot versus a subordinate robot? Would it be easier for a supervisor robot to regain trust after errors? How should erroneous robots with varying levels of authority repair its relationship with people?

*Both authors contributed equally to this research.

HRI '23, March 13–16, 2023, Stockholm, Sweden

ACM ISBN 978-1-4503-9964-7/23/03.
https://doi.org/10.1145/3568162.3576964

As a first step towards answering these questions, in this work, we conducted a between-subjects experiment contextualized in a collaborative cooking scenario (Figure 2) to study how power dynamics (supervisor robot vs. subordinate robot) and trust repair strategy (explanation vs. promise) might influence people's trust in and compliance with an erroneous robot. Our results indicate that:

- minimal modifications of robot speech content are adequate to manipulate people's perceptions of the robot's authority even when the robot is a non-anthropomorphic manipulator;
- people are willing to trust a supervisor robot that attempts to verbally mitigate its errors more than a subordinate robot;
- promise is a better trust repair strategy than explanation
- people have the tendency to comply with a supervisor robot even if it is wrong.

This paper presents empirical evidence on the interplay between power dynamics and trust repair in human-robot collaboration. Next, we review relevant prior research that motivates this work.

2 RELATED WORK

2.1 Trust in HRI

Trust has been studied in contexts such as trust for supervisors [49] and trust towards machines [27]. While there lacks a widely accepted definition of trust both within the human-human and human-automation trust literature, trust is commonly conceptualized as "a multidimensional psychological attitude involving beliefs and expectations about the trustee's trustworthiness derived from experience and interactions with the trustee in situations involving uncertainty and risk" [21, 29]. In HRI, trust is affected by cognitive (*e.g.,* how well the robot is expected to perform on the task that it was designed to do [29]) and affective (*e.g.,* empathetic robot expressions are perceived as more trustworthy [50]) features.

Trust can be gained or lost over time due to robot-related, human-related, and task and environment related factors [15, 18, 23]. Prior work characterizes trust over time with three phases [29, 35]: (1) *trust formation* occurs at the beginning of an interaction when user trust is built upon the robot's appearance, context information, and the person's prior experience with robots; (2) *trust dissolution* occurs during the interaction when users lower their trust in robot due to a trust violation, *e.g.,* robot error [5, 39]; (3) *trust restoration* describes when user trust stops decreasing after a trust violation and gets restored [5]. Thus, trust is often measured several times throughout the interaction to account for its dynamic nature. In this work, we focus on trust restoration and investigate ways to repair user trust in the robot under different power dynamics.

2.2 Trust Repair Strategies in HRI

Trust violations are inevitable in HRI; robot errors, causing trust violation, are unavoidable [8]. One robot failure is sufficient to reverse the majority of participants' attitude toward the robot or to refuse use of the robot during an emergency [5, 39]. Thus, the quality of continued HRI depends heavily on how the robot repairs lost user trust due to trust violations [43].

The effectiveness of various trust repair strategies (including apology, denial, explanation, and promise) identified from human-human interaction have been explored in HRI [9]. Studies found mixed results (from repaired trust to damaged trust) in the efficacy of apologies and denials as robot trust repair strategies [1, 7, 14, 25, 26, 28, 32, 54]. Findings on the effectiveness of explanations and promise on trust repair were also mixed, but to a lesser extent (from repaired trust to no effect) [7, 14, 25, 26, 28, 28, 32, 52]. In particular, one online study using a simulated task showed that, among the four trust repair strategies mentioned, promise was most effective at repairing integrity and benevolence aspects of trust in non-anthropomorphic robot, while explanation was most effective at repairing ability aspects of trust [14].

Studies also explored factors that affect the effectiveness of trust repair strategies. Users' attitude towards the robot moderates the efficacy of repair strategies, *i.e.,* promises were more effective at repairing trust when the user had a positive attitude towards the robot [12]. Moreover, apologies, denial, and promises were less effective at repairing integrity when given by a non-anthropomorphic robot [14]. In our study, we explore the effectiveness of explanation and promise used by a non-anthropomorphic robot for trust repair under varying human-robot power dynamics.

2.3 Power Dynamics in HRI

Status hierarchy is a part of social and organizational life. Difference in structural power (*i.e.,* supervisor vs. subordinate) changes how one perceive another person's capabilities [48], quality of their work [37], and weight of their opinion [46]. Furthermore, structural power fosters psychological power, which increases one's willingness to maintain agency [51] and assume responsibility [38]. In human-human relationships, authoritative power is enough to pressure one to comply with carrying out a destructive order [22, 31].

Similar association between power dynamic and compliance also applies in HRI. People obeyed to requests of authoritative robots, even when the tasks were tedious [16] or inappropriate for the experimental context [4]. User over-trust in automation may compromise information security [41], profitability [36, 41] and cause potentially catastrophic physical and psychological consequences for "individuals, groups of individuals, and society at large" [3, 4, 27, 40]. Moreover, when the task is more serious, urgent, or disagreeable, a more serious or authoritative robot elicits more compliance in its users [17]. Study has also explored how human-robot power dynamics (robot as supervisor, peer, and subordinate) and the level of human-likeness of the robot's appearance change the amount of responsibility people feel for the task [19]. People did not feel less responsible for the task when collaborating with the supervisor robot; however, participants felt more responsible for the task when working with a machine-like subordinate robot [19]. To the best of our knowledge, little to no work has explored how power dynamics affect the effectiveness of trust repair strategies in human-robot teams.

3 METHODS

We designed and conducted an in-person user study with *power dynamics* (subordinate and supervisor) and *trust repair strategy* (promise and explanation) as two between-subjects factors to investigate the relationship between power dynamics and trust repair strategy and their effects on people's collaborative behavior and perceptions of trust repair.

3.1 Hypotheses

We hypothesized that manipulating the robot's trust repair strategy and the power dynamic between the user and the robot will affect people's behavior and trust towards the robot. More specifically, we formulated the following hypotheses:

- **H1:** Robots that attempt verbal trust repair will be perceived as more trustworthy. This hypothesis is informed by prior work demonstrating that verbal trust recovery methods in HRI are effective at repairing trust after breaking a promise [43] and after robot errors in a simulated setting [14].
- **H2:** Promise will be a more effective trust repair strategy than explanation. Prior work showed that promise is more effective than explanation at repairing the benevolence and integrity aspects of trust in non-anthropomorphic robots in an online HRI experiment [14]. We speculate that this finding will extend to an in-person setting.
- **H3:** People will have a higher chance of complying with the supervisor robot than the subordinate one even when the robot makes a mistake. Prior work showed that people complied more to a more serious or authoritative robot [17]. Thus, we speculate that this finding continues to hold for faulty human-robot interactions.

3.2 Experimental Task

We contextualized our investigation in a cooking task as cooking robots are gaining interest in domestic and professional kitchen settings [6, 10, 47, 53]. Moreover, human-human collaboration in the kitchen typically involves clearly defined hierarchical structures and power dynamics. The head/sous chef supervises the kitchen, delegating tasks to station/junior chefs. This power dynamic will likely transfer to HRI. However, it remains unclear what role robots should assume in human-robot teams. Dexai designed their cooking robot, *Alfred*, to be "the smart sous chef in your kitchen" taking a more dominant role [10], while Sugiura et al. designed their cooking robot, *Cooky*, to take on the subordinate role (*i.e.,* transporting raw food, stirring the pot, adjusting the heat) and for the human to instruct the robot on what ingredients to use and how to adjust the heat [47]. In this study, we are interested in how user trust and reaction to robot error may change under different power dynamics.

During the task, a UR5 robot is in charge of placing the correct raw ingredient in front of the participant when the recipe calls for the ingredient. The robot is placed on a separate table across from the participants with ingredients, out of reach from the participant. The robot used a female voice generated using the Amazon Polly text-to-speech tool and was played through a speaker hidden underneath the robot. A make-believe stove, cooking utensils, a cutting board, and a pot are placed in the cooking area directly in front of the participants. A monitor showing the current recipe is placed on the side next to the robot. Figure 2 shows our experimental setup.

3.3 Manipulations

We manipulated the robot's performance (*error manipulation*), its role in the task (*power dynamics manipulation*), and what it "said" during error recovery (*trust repair strategy manipulation*) [1].

[1]Examples of our power dynamics and trust repair manipulations available at https://youtu.be/h19Bqxf0XDw

Figure 2: Overview of the experimental setup. The robot is in charge of picking up the ingredient called for in the recipe and placing it in front of the user for the user to cook with.

3.3.1 Error Manipulation. We pre-programmed the robot to provide the user with the incorrect ingredient once in each recipe at the same point with respect to the recipe. During the error, the robot delivers the incorrect ingredient to the participant: giving a mushroom instead of sausage in one recipe and zucchini instead of spinach in the other recipe.

3.3.2 Power Dynamics Manipulation. To create different power dynamics between the user and the robot, we manipulated the role of the robot in the task. The supervisor robot uses voice commands to provide step-by-step instructions for the users and takes initiative in providing users with the ingredient needed at that stage in the recipe. On the contrary, the subordinate robot does not provide any instructions to the users and waits for the user to request for the needed ingredient by saying "Please give me a ___," where the blank is the ingredient. After the participant's verbal request, the robot would pick up and place the ingredient in front of the participant. The robot specifies its role in the task to the user when it makes an self-introduction to the user at the beginning of the experiment. Specifically, the supervisor robot emphasizes that it *"will be providing you the necessary food items, as well as step by step guidance along the way*; while the subordinate robot stresses that it is *"still learning how to guide people so [it] will need supervision"* and ask users to *"use the phrase written on the paper to [their] left to request items from [the robot]. Since [the robot] has not yet learned these recipes, it will be [the user's] job to ask for the correct item"*.

3.3.3 Trust Repair Strategy Manipulation. The robot physically recovers from each error during both the control and the experimental trials by passing the correct ingredient to the user. In the experimental trial, the robot also attempts to verbally mitigate the lost trust due to its error through promise or explanation. In both strategies, the robot first verbally acknowledges the error 3 seconds after its occurrence: "Oops. I think I made a mistake". Then, depending on the manipulation, the robot either makes a promise to improve future performance (*"I will make sure to do better next time."*) or explains why the error occurred (*"My vision system sometimes has problems identifying same color food items."*).

3.4 Study Procedure

Before the study, participants were randomly assigned to one of the four experimental conditions. They filled out a personality questionnaire gauging their agreeableness using the "Big Five Inventory" after signing the consent form[20, 33]. Then, the experimenter explained the experimental setup to the participants and described the name of each ingredient. Afterwards, the experimenter went behind a divider while the participants completed two cooking trials with the robot. In the first, control trial, the robot acted according to the power dynamic manipulation but did not perform any verbal trust repair after the error occurred. Upon completion of the first trial, the participants filled out a 40-item questionnaire [42] gauging their trust in the robot. In the second, experimental trial, the robot continued to act according to the power dynamic manipulation. In addition, 3 seconds after the occurrence of the robot error, the robot verbally attempted to repair the lost trust; the trust repair strategy differed depending on the manipulation. Upon completion of the second trial, the participants again filled out the same trust questionnaire. At the end of the study, the experimenter conducted an interview with each participant to understand their experience collaborating with the robot on the cooking task. There was one open-ended question on the participants' thoughts and feelings about the robot in the two trials and if they observed any differences between the trials.

3.5 Measures

3.5.1 *Manipulation Check.*

- Robot Error Check (binary): participants were asked after each trial whether the robot made any mistakes in that trial as an attention check.
- Power Dynamics Check: participants were asked after each trial to rate the robot's perceived authority in that trial on a scale of 1–6 (1 being subordinate and 6 being supervisor).

3.5.2 *Subjective Measures.*

- Perceived Trust after Trial (range: 0–100): captures participants' trust in the robot after the control or experimental trial using the Trust Perception Scale-HRI [42]. The scale consists of 40 items presented on a 11-point scale. Perceived trust was defined to be the average across all 40 items transformed to 0% to 100%.
- Change in Perceived Trust (range: -100–100): captures the difference in participants' perceived trust in the robot between the control trial and the experimental trial to see how much their perceived trust improved or deteriorated as a result of the trust repair strategy.

3.5.3 *Behavioral Measures.*

- Improper Compliance: considers whether the participant "cooked" with the incorrect ingredient provided by the robot, *i.e.,* cut it on the cutting board or put it into the pot.
- User Reaction to Error: considers whether the participant reacted to robot error (binary); if so, whether the reaction was verbal or non-verbal (binary).
- User Reaction to Trust Repair Strategy (binary): considers whether the user had a positive social reaction during error recovery, *i.e.,* show approval or affection towards the robot.

3.6 Participants

We recruited 39 participants (10 worked with the supervisor robot using explanation as the trust repair strategy; 10 subordinate & explanation; 10 supervisor & promise; 9 subordinate & promise) through convenience sampling from the local community, using physical flyers and electronic posts to community newsletters and mailing lists. One participant, who failed the robot error check and did not mention the robot error during the post-study interview was excluded from our analysis. Of the 38 participants (17 female, 21 male), their age ranged from 18 to 60 ($M = 26.32, SD = 10.33$) and had diverse educational backgrounds (based on college major). They were some what experienced with using technology ($M = 3, SD = 1.27$, 5-point scale with 1 being expert and 5 being novice) and using robots ($M = 3.42, SD = 1.33$, 5-point scale with 1 being expert and 5 being novice). The study took roughly 30 minutes and participants were compensated at the rate of $15.00 per hour. The study was approved by our institutional review board (IRB).

4 RESULTS

In the analyses reported below, unless specified otherwise, we performed two-way analysis of variance (ANOVA) tests to examine the main effects of power dynamics and trust repair strategy and their interactions effects. Figures 4, 5, and 6, and Tables 1 and 2 summarize our results.

4.1 Modification to Robots Speech Sufficient for Power Dynamic Manipulation

As mentioned previously, we did not include data from participants who failed the robot error check in our analyses. Though social reactions to robot errors were not used as a manipulation check nor the main focus of this work, we did observe that participants exhibited a range of social reactions to robot errors (Table 1 and Figure 3); this observation is similar to prior work [44]. Below, we report our power dynamics manipulation check.

We checked whether we successfully created two power dynamics (supervisor and subordinate) through the manipulation of the recipe delivery method using a Welch's t-test assuming unequal variances. Our results revealed that participants who worked with the supervisor robot ($M = 4.38, SD = 1.24$) had significantly higher

Table 1: Reactions in response to robot error observed during both trials

Reaction		Count
Verbal (*e.g.,* "bad robot", "okay", "it's not the sausage")		14
Non-Verbal	Hand Movement (*e.g.,* scratch head, fidget)	4
	Head Movement (*e.g.,* shake, tilt)	5
	Brow Movement (*e.g.,* raise, squeeze)	3
	Mouth Movement (*e.g.,* pout, smile)	16
	Eye Movement (eye widen, fast blinking)	2
	Scrunch Face	4
	Total	34
No Reaction		17
Total		65

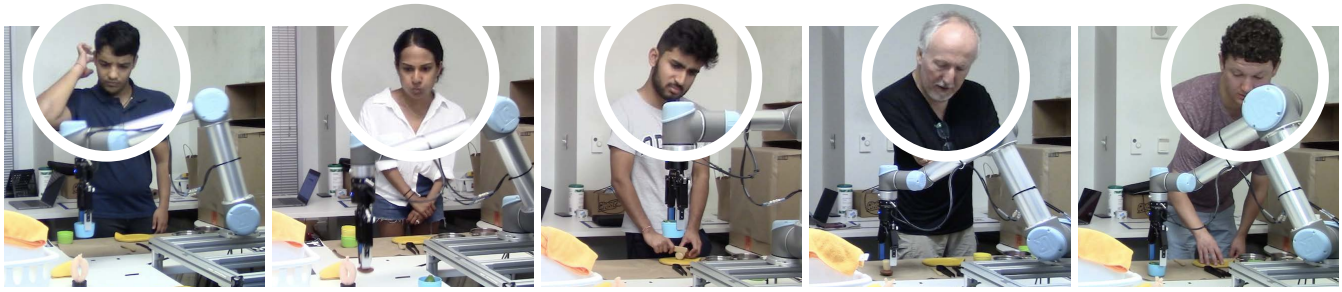

Figure 3: Examples of participants' social reactions to robot errors.

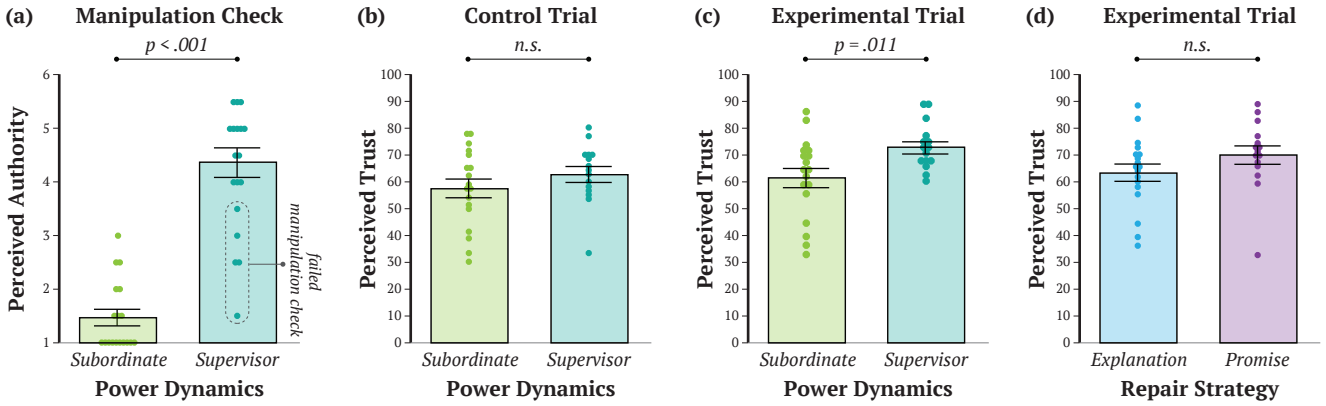

Figure 4: (a) Effect of power dynamic manipulation on participant's perceived robot authority rating, which serves as the power dynamic manipulation check. (b) Effect of power dynamics on the user's perceived trust in the robot after the control trial (without verbal trust repair). (c) Effect of power dynamics on the user's perceived trust in the robot after the experimental trial (with verbal trust repair). (d) Effect of trust repair strategy on the user's perceived trust in the robot after the experimental trial. The error bars represent standard error.

robot authority rating than participants who worked with the subordinate robot ($M = 1.47, SD = 0.65$), $t(29.32) = 9.13, p < .001$, $d = 2.94$, indicating that our power dynamics manipulation was adequate (Figure 4 a). In the rest of the analysis, we excluded the five participants who provided the incorrect robot authority rating for their power dynamic manipulation; *i.e.,* participants who gave an robot authority rating of greater than three to the subordinate robot and participants who gave an robot authority rating of less than four to the supervisor robot. In this study, all five excluded participants were originally assigned to the supervisor robot.

4.2 Power Dynamics Alone Does Not Affect User Perceived Trust in the Robot

We conducted a Welch's t-test assuming unequal variances to investigate the effect of power dynamic on user perceived trust in the robot during the control trial. The results showed no significant differences in the perceived trust ratings among participants who worked with the subordinate robot ($M = 57.72, SD = 14.67$) and participants who worked with the supervisor robot ($M = 62.99, SD = 11.40$), $t(30.88) = 1.16$, $p = .256$, $d = 0.40$ (Figure 4 b).

4.3 Higher Perceived Trust in the Supervisor Robot That Attempted Verbal Trust Repair

We conducted a two-way ANOVA test to examine the effect of power dynamics and trust repair strategy on user's trust in the robot after the experimental trial. Our results showed that participants working with the supervisor robot using verbal trust repair ($M = 72.91, SD = 8.84$) had significantly higher perceived trust than participants working with the subordinate robot using verbal trust repair ($M = 61.54, SD = 15.06$), $F(1, 32) = 7.35, p = .011$, $\eta_p^2 = 0.20$ (Figure 4 c). However, no main effect of trust repair strategy ($F(1, 32) = 2.86, p = .102$, $\eta_p^2 = 0.09$, Figure 4 d) nor interaction effect of power dynamics and trust repair strategy ($F(1, 32) = 1.95, p = .174$), $\eta_p^2 = 0.06$ were found.

4.4 Promise Is More Effective at Repairing Trust Than Explanation

We conducted a two-way ANOVA test to examine the effect of power dynamics and trust repair strategy on the change in participants' self-perceived trust in the robot between the control trial and the experimental trial. Our results showed that participants working with the supervisor robot ($M = 9.93, SD = 8.09$) had a

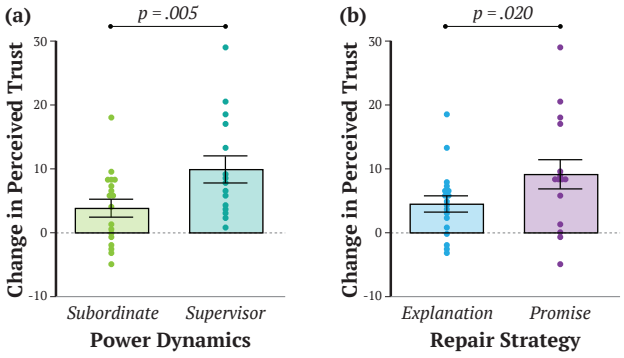

Figure 5: (a) Effect of power dynamics on the user's change in perceived trust between the control and the experimental trials. (b) Effect of trust repair strategy on the user's change in perceived trust. The error bars represent standard error.

significantly greater increase in trust from the control to the experimental trial than participants working with the subordinate robot ($M = 3.82, SD = 5.85$), $F(1, 32) = 9.25, p = .005$, $\eta_p^2 = 0.24$ (Figure 5 a). Moreover, participants who experienced the promise strategy ($M = 9.15, SD = 8.95$) had a significantly greater increase in trust from the control to the experimental trial than participants who experienced the explanation strategy ($M = 4.47, SD = 5.44$), $F(1, 32) = 6.10, p = .020$, $\eta_p^2 = 0.14$ (Figure 5 b). However, no interaction effect was found, $F(1, 32) = 1.36, p = .252$, $\eta_p^2 = 0.04$.

4.5 Users Tended to Improperly Comply with the Supervisor Robot

We conducted contingency analysis and a likelihood ratio test[2] to explore the effect of power dynamics on improper compliance behavior in participants. Participants were significantly more likely to cook with the incorrect ingredient handed to them by the supervisor robot (9 out of 30 trials = 0.30) than the subordinate robot (0 out of 35 trials = 0.00), $\chi^2(1, 64) = 15.63, p < .001$ (Figure 6 a).

4.6 Users Tended to React Verbally to Errors from the Supervisor Robot

Robot error triggered verbal reactions in 14 out of 65 [3] trials and a variety of non-verbal reactions in 33 out of 65 trials as shown in Table 1. Through contingency analysis and a likelihood ratio test, we observed no significant effect of power dynamics on whether or not the user reacted, verbally or non-verbally, to robot errors, $\chi^2(1, 64) = 2.25, p = .134$.

Among participants who did react to the robot error, results from contingency analysis and a likelihood ratio test revealed that participants who worked with the supervisor robot (12 out of 19 trials = 0.63) were significantly more likely to have verbal reaction than non-verbal reaction to robot error compared to those who worked with the subordinate robot (2 out of 28 trials = 0.07), $\chi^2(1, 46) = 17.83, p < .001$ (Figure 6 b).

Table 2: Reactions in response to robot error recovery observed during the experimental trials

Reactions	Positive		Neutral		Total
	Smile	Nod	Freeze	No Reaction	
Count	11	3	3	15	32

4.7 Promise Triggered More Positive Social Reactions Than Explanations

Out of the 32 valid participants[4], the robot's trust repair triggered people's behavioral reactions as shown in Table 2; 14 out of 32 were positive social signals (examples in Figure 7). To study the effect of power dynamics, trust repair strategy, and their interaction on whether or not the participants had a positive social reaction to the robot's mitigation attempt, we trained a binary logistic regression model. Our results based on likelihood ratio tests showed that participants were significantly more likely to have a positive reaction to the promise strategy (9 out of 15 trials = 0.60) than the explanation one (5 out of 17 trials = 0.29), $\chi^2(1, 31) = 3.97, p = .046$ (Figure 6 c). No significant main effect of power dynamic ($\chi^2(1, 31) = 2.57, p = .109$) nor an interaction effect of power dynamic and trust repair strategy ($\chi^2(1, 31) = 0.43, p = .510$) were found.

5 DISCUSSION

5.1 Effect of Trust Repair Strategy

In this study, we compared two trust repair strategies, promise and explanation, during error recovery. In general, participants provided more positive accounts of their experience working with robots with trust repair during the post-study interview. Participants reported that when working with robots without verbal trust repair, they "*felt like [they were not] heard*". In contrast, the trust repair made them feel "*warm and nice*". Additionally, **while having either of the trust repair strategies is better than no verbal trust repair, promise is more effective than explanations at repairing user trust** (Figure 5 b). This finding is consistent with our hypothesis 2 (promise will be a more effective user trust repair mechanism than explanation) and the results of a prior work's online experiment using a high-fidelity simulated human-robot interaction task (promise to be more effective than explanation in repairing benevolence and integrity, two key characteristics of trustworthiness along with ability, in non-anthropomorphic robots) [14]. Moreover, our results showed that promise triggered more positive social reactions (smiling or nodding) in users than explanations when the robot verbally acknowledged the error and either promised to "do better next time" or provided an explanation for the erroneous behavior. While participants with either strategies appreciated the robot acknowledging its error and found the robot to be funny, they described promise to be a "*very folksy human response*" that the action *made [them] feel grateful*" and the robot as

[2]We lost the data for the control trial from one participant, thus we only have 65 trials (control and experimental trials combined) from 33 participants.
[3]The robot completely blocked the camera from recording the reaction of one participants to the robot error.
[4]The robot completely blocked the camera from recording the reaction of one participants to the robot error mitigation and error recovery.

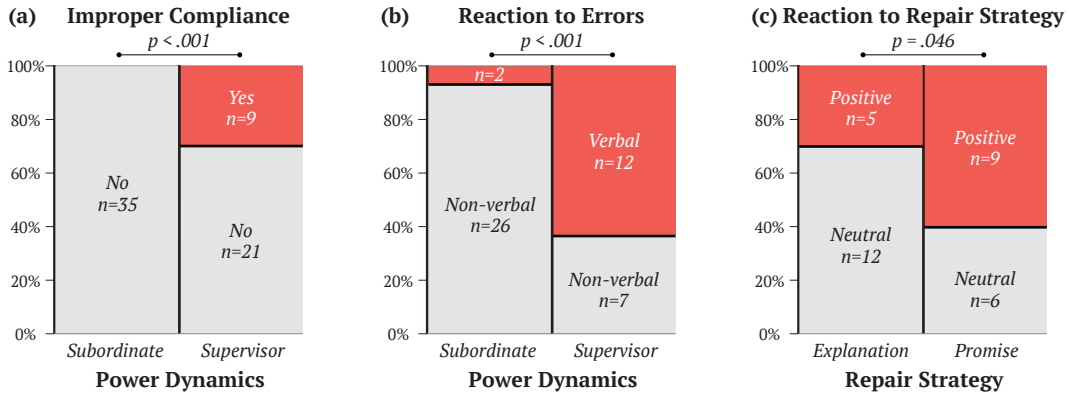

Figure 6: (a) Effect of power dynamics on the distribution of participants with improper compliance behavior. (b) Effect of power dynamics on the distribution of types of reactions users had to the robot error (verbal or non-verbal). (c) Effect of trust repair strategy on the distribution of types of reactions users had to the robot trust recovery (positive or neutral).

"*cute*". This shows the benefits to using promise as the trust repair strategy in non-anthropomorphic robots.

5.2 Generating Power Dynamics

Prior research found non-anthropomorphic robots to be perceived as less trustworthy by users initially and less likely to be accepted as team partners [11]. Moreover, interacting with machine-like robots increases the personal responsibility the user feels for the task [19]. As a result, we were initially unsure if we were able to manipulate participants' possible pre-conception of the robot's authority and establish the robot's role as a supervisor. However, **we found that modifying the robot's speech content was sufficient to manipulate people's perception of the robot's authority in the majority of the participants** (Figure 4 a). Though, we note that we failed to convey the power dynamic manipulation to five participants originally assigned to the supervisor robot.

5.3 Effect of Power Dynamics

We found that manipulating the power dynamics alone did not influence user trust in the robot (Figure 4 b). However, the addition of verbal trust repair strategy did. Under both power dynamics, participants trusted robots with verbal trust repair more than robots without verbal trust repair (Figure 5 a). This finding is consistent with our hypothesis 1 (regardless of the power dynamics, people will have higher trust for robots with verbal trust repair). Moreover, this result is consistent with prior work that found verbal trust repair to be effective in repairing lost trust in users due to robot error [14]. Furthermore, our results showed that **the addition of verbal trust repair to the robot resulted in a significantly greater increase in trust among participants with the supervisor robot than the subordinate robot** (Figure 5 a). As a result, supervisor robot with verbal trust repair for error recovery is trusted more by users than subordinate robot with verbal trust repair (Figure 4 c).

On the other hand, among participants who reacted to the robot error, they had significantly more verbal reactions (*i.e.,* saying *"Bad Robot"* and *"This is not sausage though"*) to error made by a supervisor robot than a subordinate robot. We speculate that part of the reason that we observed this result may be because the supervisor robot talked the users through the recipe, making the users feel like they could communicate with the robot through speech.

Figure 7: Examples of participants showing positive reactions to the robot's use of the *promise* strategy for trust repair.

5.4 Improper Compliance

Prior work showed that people are more compliant with robots that have a more authoritative social demeanor [17]. In this study, we found that **even when the robot was making a mistake, interacting with the supervisor robot increased the chance of compliance in participants** (Figure 6 a). This finding is consistent with our hypothesis 3 (supervisor robot will lead to greater compliance even when the robot makes a mistake). Among the nine trials where the participant cooked with the incorrect ingredient provided by the robot, four participants did not notice the robot error. These four participants did not react to the robot error and mentioned in the post-study interview that they did not notice the robot error until the robot acknowledged its own error because "the zucchini [the incorrect ingredient passed to the participants] and the spinach [the correct ingredient called for in the recipe] looked alike" and they "*just weren't thinking about it too much*". However,

five other participants who improperly complied, clearly pointed out the robot error verbally, *i.e.,* "*what if it is wrong*" (illustrated in Figure 1) or "that is a mushroom [the name of the incorrect ingredient picked up by the robot]", but then still decided to comply with the robot and proceeded to "cook" with the wrong ingredient. This observation shows that these five participants were not mindlessly complying with the robot, but rather their motivation to obey authority over-powered their agency and independent thinking. This behavior is common in human-human interaction. Cues from an authority are a powerful motivating mechanism for people to comply, even when the cue was destructive [22, 31]; this may explain why participants complied with the robot even when they noticed that the robot made an error.

We also observed that robot error acknowledgement may be an effective way to prevent improper compliance in users. Eight out of nine cases of improper compliance occurred in the control trial; the only exception was the participant mentioned above who did not notice the robot error. In the experimental trial, the robot acknowledged their mistake three seconds after the error. As a result, in all but one case, the robot made the error acknowledgement while the users were contemplating or "second-guessing" themselves on what to do with the incorrect ingredient. Robot error acknowledgement appeared to resolve users' self-doubt; no improper compliance was observed during nor after robot trust repair, suggesting the importance of error detection and acknowledgement in avoiding improper compliance in users.

5.5 Design and Ethical Implications

Our empirical findings had important implications for the design of collaborative robot systems. First, we showed the benefits of including trust repair mechanisms in robots. In particular, regardless of the role of the robot, promise effectively repaired user trust in non-anthropomorphic robots after one robot error. Not only did users perceive robots with promise as its trust repair strategy to be more trustworthy, promise also elicited more positive social reactions in the users (Figure 6 c).

However, our results showed concerning ethical implications in future use of supervisor robots, particularly those without error awareness. Prior work encouraged the use of more authoritative robots in more serious, urgent, or disagreeable task contexts—such as getting a chore done, taking medication or sticking to an exercise routine—to induce more compliance in users [17]. Yet, in this study, users, affected by the robot's authority, sometimes complied with the robot even though they noticed a robot error. This kind of improper compliance may lead to potentially catastrophic failures in critical applications, *e.g.,* search and rescue. Even in non-critical tasks, *e.g.,* cooking task used in this study, improper compliance following a robot error may cause severe consequences. For example, adding an incorrect ingredient may trigger allergic reactions in customers with dietary restrictions. Thus, measures should be taken to prevent improper compliance in HRI particularly when the robot has a more active and responsible role.

In this study, we found the robot error acknowledgement three seconds after the error appeared to be sufficient in intervening against the majority of improper compliance cases. Another potential intervention is to periodically remind the users to stay alert as robot errors are possible. Future work should explore other interventions to help prevent improper compliance in users. In summary, caution must be used in the design of robot systems that adopt a supervisor or other more authoritative role.

5.6 Limitations and Future Work

There exists a few limitations to this study that call for further exploration. Our small sample size limited our ability to study interaction effects of our independent variables. Future work should recruit more participant per condition to further investigate the potential interaction effects. In this work, we manipulated power dynamic indirectly through changing how the robot is introduced (supervisor as task expert vs. subordinate as still learning) and its role in the task (active vs. passive), which may have confounded our results. We did not provide details on the robot capabilities when introducing the robot to the user. Thus, users, unfamiliar with robots, may have presumed the seemingly "expert" supervisor robot to be always correct, leading to more improper compliance. Future work should explore compliance under other power dynamics manipulation paradigms. We observed that a prompt error acknowledgement effectively intervened improper compliance behavior in users. However, for a real-world scenario, to give a prompt error acknowledgement would require the robot to timely detect its mistake. A recent work has demonstrated the possibility of timely, automatic detection of robot errors using people's instinct social reactions to the errors [45]. Indeed, similar to prior work [44], we observed that our participants reacted to robot errors socially even if the robot is a non-anthropomorphic robot manipulator (Figure 3). Future work should investigate the integration of automatic error detection and the uses of trust repair strategies to mitigate unavoidable robot errors in complex human-robot collaboration. Future work should also explore other interventions, such as periodically reminding the users of the robot's mean time between failure [34], to prevent improper compliance.

Furthermore, as the technology powering robot capabilities continues to advance, robots are going to become more sophisticated and able to take on a variety of roles to assist and collaborate with people. In this work, we focused only on studying a clear-cut power dynamic (supervisor vs. subordinate) between the participants and the robot. Future work should explore more nuanced dynamics and how different trust repair strategies might achieve their intended outcomes under more nuanced, complex dynamics. Finally, results reported in this work are based on a short interaction session. One participant mentioned in the post-study interview that promise would only work the "first few times", which indicate that explanation or a combination of promise and explanation may potentially be more effective under repeated robot errors. Future work should investigate whether various trust repair strategies has continued positive effects over multiple interaction sessions and how power dynamics over time might additionally shape the effectiveness of trust repair strategies.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation award #2141335.

REFERENCES

[1] Yusuf Albayram, Theodore Jensen, Mohammad Maifi Hasan Khan, Md Abdullah Al Fahim, Ross Buck, and Emil Coman. 2020. Investigating the effects of (empty) promises on human-automation interaction and trust repair. In *Proceedings of the 8th International Conference on Human-Agent Interaction*. 6–14.

[2] Sean Andrist, Erin Spannan, and Bilge Mutlu. 2013. Rhetorical robots: Making robots more effective speakers using linguistic cues of expertise. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Tokyo, Japan). IEEE.

[3] Alexander M Aroyo, Jan De Bruyne, Orian Dheu, Eduard Fosch-Villaronga, Aleksei Gudkov, Holly Hoch, Steve Jones, Christoph Lutz, Henrik Sætra, Mads Solberg, et al. 2021. Overtrusting robots: Setting a research agenda to mitigate overtrust in automation. *Paladyn, Journal of Behavioral Robotics* 12, 1 (2021), 423–436.

[4] Alexander Mois Aroyo, T Kyohei, Tora Koyama, Hideyuki Takahashi, Francesco Rea, Alessandra Sciutti, Yuichiro Yoshikawa, Hiroshi Ishiguro, and Giulio Sandini. 2018. Will people morally crack under the authority of a famous wicked robot?. In *2018 27th IEEE international symposium on robot and human interactive communication (RO-MAN)*. IEEE, 35–42.

[5] Anthony L. Baker, Elizabeth K. Phillips, Daniel Ullman, and Joseph R. Keebler. 2018. Toward an Understanding of Trust Repair in Human-Robot Interaction: Current Research and Future Directions. *ACM Trans. Interact. Intell. Syst.* 8, 4, Article 30 (nov 2018), 30 pages. https://doi.org/10.1145/3181671

[6] Justin Bishop, Jaylen Burgess, Cooper Ramos, Jade B Driggs, Tom Williams, Chad C Tossell, Elizabeth Phillips, Tyler H Shaw, and Ewart J de Visser. 2020. CHAOPT: a testbed for evaluating human-autonomy team collaboration using the video game overcooked! 2. In *2020 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 1–6.

[7] David Cameron, Stevienna de Saille, Emily C Collins, Jonathan M Aitken, Hugo Cheung, Adriel Chua, Ee Jing Loh, and James Law. 2021. The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computers in human behavior* 114 (2021), 106561.

[8] Jennifer Carlson, Robin R Murphy, and Andrew Nelson. 2004. Follow-up analysis of mobile robot failures. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, Vol. 5. IEEE, 4987–4994.

[9] Ewart J De Visser, Richard Pak, and Tyler H Shaw. 2018. From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics* 61, 10 (2018), 1409–1427.

[10] DEXAI. 2022. DEXAI-Meet Alfred. https://www.dexai.com/meet-alfred. Accessed: 2022-09-24.

[11] Brian R Duffy. 2003. Anthropomorphism and the social robot. *Robotics and autonomous systems* 42, 3-4 (2003), 177–190.

[12] Connor Esterwood, Lionel Robert, et al. 2022. Having The Right Attitude: How Attitude Impacts Trust Repair in Human-Robot Interaction. (2022).

[13] Connor Esterwood, Lionel Robert, et al. 2022. A Literature Review of Trust Repair in HRI. (2022).

[14] Connor Esterwood and Lionel P. Robert. 2021. Do You Still Trust Me? Human-Robot Trust Repair Strategies. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. 183–188. https://doi.org/10.1109/RO-MAN50785.2021.9515365

[15] Diego Gambetta. 1988. Trust: Making and breaking cooperative relations. (1988).

[16] Denise Y Geiskkovitch, Derek Cormier, Stela H Seo, and James E Young. 2016. Please continue, we need more data: an exploration of obedience to robots. *Journal of Human-Robot Interaction* 5, 1 (2016), 82–99.

[17] Jeniffer Goetz, Sara Kiesler, and Aaron Powers. 2003. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.* 55–60. https://doi.org/10.1109/ROMAN.2003.1251796

[18] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie Y C Chen, Ewart J de Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors* 53, 5 (Oct. 2011), 517–527.

[19] Pamela J. Hinds, Teresa L. Roberts, and Hank Jones. 2004. Whose Job Is It Anyway? A Study of Human-Robot Interaction in a Collaborative Task. *Human–Computer Interaction* 19, 1-2 (2004), 151–181. https://doi.org/10.1080/07370024.2004.9667343 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/07370024.2004.9667343

[20] Oliver P John and Sanjay Srivastava. 1999. The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of personality: Theory and research*, L A P Pervin & O (Ed.). Guilford Press, 102–138.

[21] Gareth R Jones and Jennifer M George. 1998. The experience and evolution of trust: Implications for cooperation and teamwork. *Academy of management review* 23, 3 (1998), 531–546.

[22] Alexandros Karakostas and Daniel John Zizzo. 2016. Compliance and the power of authority. *Journal of Economic Behavior & Organization* 124 (2016), 67–80. https://doi.org/10.1016/j.jebo.2015.09.016 Taxation, Social Norms and Compliance.

[23] Zahra Rezaei Khavas. 2021. A Review on Trust in Human-Robot Interaction. (2021).

[24] Yunkyung Kim and Bilge Mutlu. 2014. How social distance shapes human–robot interaction. *International Journal of Human-Computer Studies* 72, 12 (2014), 783–795. https://doi.org/10.1016/j.ijhcs.2014.05.005

[25] Spencer C Kohn, Ali Momen, Eva Wiese, Yi-Ching Lee, and Tyler H Shaw. 2019. The consequences of purposefulness and human-likeness on trust repair attempts made by self-driving vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 222–226.

[26] Esther S Kox, José H Kerstholt, Tom F Hueting, and Peter W de Vries. 2021. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous agents and multi-agent systems* 35, 2 (2021), 1–20.

[27] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[28] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 203–210.

[29] Michael Lewis, Katia Sycara, and Phillip Walker. 2018. *The Role of Trust in Human-Robot Interaction*. Springer International Publishing, Cham, 135–159. https://doi.org/10.1007/978-3-319-64816-3_8

[30] Dan Leyzberg, Brian Scassellati, Samuel Spaulding, and Mariya Toneva. 2012. The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains.

[31] Stanley Milgram. 1963. Behavioral study of obedience. *J. Abnorm. Psychol.* 67, 4 (Oct. 1963), 371–378.

[32] Manisha Natarajan and Matthew Gombolay. 2020. Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 33–42.

[33] Warren T Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology* 66, 6 (1963), 574.

[34] Illah R Nourbakhsh, Clayton Kunz, and Thomas Willeke. 2003. The mobot museum robot installations: A five year experiment. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, Vol. 4. IEEE, 3636–3641.

[35] Linda Onnasch and Clara Laudine Hildebrandt. 2021. Impact of anthropomorphic robot design on trust and attention in industrial human-robot interaction. *ACM Transactions on Human-Robot Interaction (THRI)* 11, 1 (2021), 1–24.

[36] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[37] Jeffrey Pfeffer, Robert B Cialdini, Benjamin Hanna, and Kathleen Knopoff. 1998. Faith in supervision and the self-enhancement bias: Two psychological reasons why managers don't empower workers. *Basic and Applied Social Psychology* 20, 4 (1998), 313–321.

[38] Karlene H Roberts, Suzanne K Stout, and Jennifer J Halpern. 1994. Decision dynamics in two high reliability military organizations. *Management science* 40, 5 (1994), 614–624.

[39] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE, 101–108.

[40] Ericka Rovira, Kathleen McGarry, and Raja Parasuraman. 2007. Effects of imperfect automation on decision making in a simulated command and control task. *Human factors* 49, 1 (2007), 76–87.

[41] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1–8.

[42] Kristin E Schaefer. 2016. Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In *Robust intelligence and trust in autonomous systems*. Springer, 191–218.

[43] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. 2019. "I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 57–65. https://doi.org/10.1109/HRI.2019.8673169

[44] Maia Stiber and Chien-Ming Huang. 2020. Not all errors are created equal: Exploring human responses to robot errors with varying severity. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 97–101.

[45] Maia Stiber, Russell Taylor, and Chien-Ming Huang. 2022. Modeling Human Response to Robot Errors for Timely Error Detection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[46] Fred L Strodtbeck, Rita M James, and Charles Hawkins. 1957. Social status in jury deliberations. *American Sociological Review* 22, 6 (1957), 713–719.

[47] Yuta Sugiura, Daisuke Sakamoto, Anusha Withana, Masahiko Inami, and Takeo Igarashi. 2010. Cooking with robots: designing a household system working in open environments. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2427–2430.

[48] Janet K Swim and Lawrence J Sanna. 1996. He's skilled, she's lucky: A meta-analysis of observers' attributions for women's and men's successes and failures. *Personality and Social Psychology Bulletin* 22, 5 (1996), 507–519.

[49] Hwee H Tan and Christy S Tan. 2000. Toward the differentiation of trust in supervisor and trust in organization. *Genet. Soc. Gen. Psychol. Monogr.* 126, 2 (May 2000), 241–260.

[50] Adriana Tapus, Mataric Maja, and Brian Scassellatti. 2007. The Grand Challenges in Helping Humans Through Social Interaction. *The Grand Challenges in Socially Assistive Robotics. IEEE Robotics and Automation Magazine* 14 (2007).

[51] Leigh Plunkett Tost. 2015. When, why, and how do powerholders "feel the power"? Examining the links between structural and psychological power and reviving the connection between power and responsibility. *Research in Organizational Behavior* 35 (2015), 29–56.

[52] Ning Wang, David V Pynadath, Ericka Rovira, Michael J Barnes, and Susan G Hill. 2018. Is it my looks? or something i said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In *International Conference on Persuasive Technology*. Springer, 56–69.

[53] WX Yan, Zhuang Fu, YH Liu, YZ Zhao, XY Zhou, JH Tang, and XY Liu. 2007. A novel automatic cooking robot for Chinese dishes. *Robotica* 25, 4 (2007), 445–450.

[54] Xinyi Zhang. 2021. "Sorry, It Was My Fault": Repairing Trust in Human-Robot Interactions. (2021).