Having the Right Attitude: How Attitude Impacts Trust Repair in Human–Robot Interaction

Connor Esterwood
School of Information, University of Michigan
Ann Arbor, MI, USA
cte@umich.edu

Lionel P. Robert Jr
School of Information, University of Michigan
Robotics Institute, University of Michigan
Ann Arbor, MI, USA

***Abstract*—Robot co-workers, like human co-workers, make mistakes that undermine trust. Yet, trust is just as important in promoting human–robot collaboration as it is in promoting human–human collaboration. In addition, individuals can significantly differ in their attitudes toward robots, which can also impact or hinder their trust in robots. To better understand how individual attitude can influence trust repair strategies, we propose a theoretical model that draws from the theory of cognitive dissonance. To empirically verify this model, we conducted a between-subjects experiment with 100 participants assigned to one of four repair strategies (apologies, denials, explanations, or promises) over three trust violations. Individual attitudes did moderate the efficacy of repair strategies and this effect differed over successive trust violations. Specifically, repair strategies were most effective relative to individual attitude during the second of the three trust violations, and promises were the trust repair strategy most impacted by an individual's attitude.**

***Index Terms*—Human-Robot Interaction, Trust Repair, Attitude**

I. Introduction

The importance of trust in human–robot collaboration has spurred much interest in the study of trust repair. Human–robot trust repair can be defined as the approaches taken to amend the loss of trust between the human and the robot [1]. Humans are increasingly being expected to work with and actively collaborate with robot co-workers in new work arrangements [2]–[5]. Trust can be defined 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" [6, Pg.712]. However, robot co-workers, like human co-workers, make mistakes that can undermine trust. This has led to the emergence of research on human–robot trust repair.

Despite this, much remains to be learned about the conditions under which trust repair strategies are or are not effective [7], [8]. More specifically, research has tended to ignore individual differences that might significantly alter the effectiveness of trust repair strategies. Yet, we know from prior literature that individuals can significantly differ in their attitudes toward robots, which can also promote or hinder their trust in robots [9], [10]. This implies that individual attitudes toward robots might make such repairs harder or easier. Therefore, it is critical to examine whether such differences do indeed alter the effectiveness of trust repair strategies.

To address this, we conducted a between-subjects experiment with 100 participants. The study randomly assigned participants to one of four trust repair strategies to determine whether their attitude toward working with robots influences the effectiveness of trust repair strategies. Overall, this paper con tributes to the human–robot interaction (HRI) trust repair literature in the following ways. One, it extends our theoretical understanding of trust repair strategies by demonstrating how individual attitude can significantly and meaningfully influence which trust repair strategies are effective for a particular individual. Two, it examined the effectiveness of trust repair strategies over multiple human–robot interactions; results showed that the effectiveness of trust repair strategies varies greatly over successive interactions. In these ways, our study extended the traditional static models of trust repair by bringing in a dynamic perspective.

Esterwood, C. and Robert, L. P. (2022) **Having The Right Attitude: How Attitude Impacts Trust Repair in Human-Robot Interaction**, *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (**HRI 2022**), March 7-10, 2022.

II. Background

A. Trust Repair

Robots, like humans, inevitably make mistakes; when this occurs trust repair is vital to minimizing the loss of trust [1], [5], [8], [11]–[17]. Trust repair refers to the efforts undertaken by a trustee to restore trust following an actual or perceived trust violation [5], [8], [18]–[20]. These efforts can rely on one of several strategies: apologies, denials, explanations, or promises [5], [11], [21]–[24]. Apologies are a communication and/or expression of remorse or regret [5], [8], [25]–[27]; denials are a rejection of culpability and often a redirection of blame [5], [11], [21]–[24]; explanations are clear and direct reasoning behind why a violation of trust occurred [1], [5], [8], [28]; and promises are statements conveying the intention to correct future behavior [8], [16], [22].

B. Trust Repair in HRI

In this section we review the HRI trust repair literature. Specifically, we examine the HRI literature as it relates to the efficacy of apologies, denials, explanations, and promises in repairing trust. In doing so we establish what the current state of the art in relation to trust repair in human–robot interaction.

The findings on efficacy of apologies in HRI have been relatively mixed. In particular, three studies found that apologies repaired trust [29]–[31], three found that they did not [32]–[34], and one found that apologies actually damaged trust [35]. One study that found that apologies repaired trust, Natarajan et al. [30], found that when robots provided bad advice to their human counterpart, apologies were as effective as explanations in repairing trust. Similarly, Kohn et al. [31] examined apologies after robots engaged in inappropriate and unsafe behaviors and found that apologies repaired trust after such violations and even did so more effectively than denials. Finally, Albayram et al. [29] found that apologies were effective but did not compare them to other trust repair strategies.

One study that found that apologies were not effective at repairing trust, Lee et al. [32], examined apologies after a robot retrieved an incorrect item; their results indicated no significant difference between trust in a no-repair strategy condition and an apology condition, indicating that the robot's apologies had no effect. In addition, Kox et al. [33] examined trust after a robot apologized for providing bad advice, finding similarly non-significant results where trust after apologies was not significantly different from trust after a no-repair strategy condition. Finally, Kohn et al. [34] considered apologies given after a robot performed poorly on a navigation task. Once again, results provided no evidence to support that apologies impacted trust more or less than if no repair strategy was deployed. Overall this is consistent with Lee et al. [32] and Kox et al. [33] but conflicts with other studies examining the subject.

In addition to the six aforementioned studies, Cameron et al. [35] found that apologies actually made thing worse. In particular, they found that apologies were not only ineffective but also decreased trust. Within this study, however, trust was subdivided into three latent constructs. These constructs were performance, integrity, and deceitfulness. Results identified that apologies impacted these constructs deferentially, with apologies damaging performance but having no effect on integrity or deceitfulness. Overall for apologies, the current literature in human–robot interaction appears mixed.

Studies examining denials in HRI have also found mixed results overall. In particular, one study found that denials are effective [31], one found that they are not [34], and a third found that their effects depend on the type of trust examined [36]. The last of these studies [36] compared competence-based and integrity-based trust. Findings indicated that denials do not effectively repair competence-based trust and actually have a negative impact on trust for integrity-based trust. Therefore, it might be that denials are not only ineffective at repairing trust but also make matters worse.

Explanations' impact on trust in the HRI literature, like apologies and denials, was also mixed, though by a lesser degree. In particular, one study [30] examined explanations and found that they are effective in repairing trust, while three studies [32]–[34] found that explanations fail to repair trust. Finding mixed results, Cameron et al. [35] examined trust by dividing it into three components, namely ability (performance), integrity, and deceit. Results of this study found that explanations are not effective in repairing the ability and integrity components of trust, but the authors observed that explanations significantly reduce the deceit component of trust. This might indicate that explanations are effective repair strategies to the degree that they decrease deceit. Once again, more research is needed and inconsistencies are present across this literature.

Promises, unlike apologies, denials, and explanations, are relatively under-examined in the HRI trust repair literature. In particular, two studies reported the direct impact of promises on trust [1], [37]. Specifically, the first of these [37] examined promises that were either provided separately or given jointly with an explanation. Results showed that in cases where promises were given independent of explanations they were not effective, but when promises were coupled with explanations, they were capable of repairing trust. The second of these studies [1] examined the impact of promises when given after a robot made an ability-based error by presenting a human co-worker with an incorrect box. This study's results indicated that promises are effective in repairing humans' perceptions of a robot's trustworthiness but only in the case of benevolence and not for ability or integrity. Given that only two studies have examined the direct effects of promises, more work needs to be conducted and study replications are needed.

Based on this review of the literature, the field of HRI appears far from consensus regarding the efficacy of apologies, denials, explanations, and promises. Given the mixed results regarding each of these repair strategies and a general lack of studies examining promises, we therefore examined a possible moderator that could explain these inconsistent results. In particular, we examined individuals' attitudes toward working with robots. In the subsequent section we provide a brief background on how individual differences have been shown to impact trust and then discuss how positive attitudes toward working with robots might impact trust repair via cognitive dissonance.

C. Individual Differences and Trust

Although no studies have examined the impact that attitudes toward working with robots have on trust repair, studies have examined other types of individual differences and their impact on trust. In particular, studies examining automation such as recommendation and collision avoidance systems have found that individuals differ in their expectations and beliefs of a system's reliability and competence and that these expectations might impact trust [38], [39]. For example, Lyons et al. [38] found that as participants' expectations rose, so did their trust in a collision avoidance system. Similarly, Pop et al. [39] observed that participants with higher expectations of an automated system placed more trust in that system.

In addition, authors have examined how personality traits and cultural backgrounds might impact expectations and trust in automation [40]–[44]. For example, Merritt and Ilgen [43] found that not only can trust propensity predict trust but users'

extroversion can as well. Along similar lines, Szalma and Taylor [42] considered individual differences and how they impact perceptions of automation. In particular, they found that an individual's personality traits of agreeableness and extroversion impacted the stress operators experienced when faced with unreliable automation and that neuroticism influenced the likelihood of an individual to accept the recommendation of said system. Finally, Chien et al. [41] considered cultural differences and found that trust in automated systems differs based on the cultural attributes of an individual. Ultimately, these studies provide support for the idea that individual differences have a role to play when considering trust in automation. This was further highlighted in a recent meta-analysis on personality and robot acceptance [45]. Nonetheless, we know little with regard to whether such individual differences alter the effectiveness of trust repair strategies. Answering this question might help us determine which repair strategy might be more or less appropriate for a specific individual.

D. Individual Attitudes Toward Working with Robots

Theories of trust have openly acknowledged the importance of individual attitudes [46]. Attitudes represent favorable or unfavorable feelings toward a particular person, place, thing, event, or action [47]. Attitudes can influence an individual's thoughts and behaviors, which explains why they are prominent in many social–psychological theories [47], [48]. Generally, the more favorable an attitude, the more likely an individual is to engage and enjoy engaging with a person, place, thing, event or action. Attitudes toward robots can differ significantly among individuals and have been used to explain why people trust robots [10], [49].

In this paper, we propose that positive attitudes toward robots will moderate the effectiveness of trust repair strategies. Attitudes toward robots drive individual expectations about robots [50], [51]. Those who have positive attitudes about working with robots are likely to have positive beliefs about future actions with robots [50], [51]. Therefore, individuals with a more positive attitude toward robots are likely to have positive expectations about robots, while those with less positive attitudes toward robots are likely to have much lower expectations.

III. Theoretical HRI Trust Repair Model

In this paper, we assert that prior attitudes about robots influence the effectiveness of trust repair strategies and that cognitive dissonance can explain why. Cognitive dissonance theory is used to explain what happens when an individual is presented with new information that seems counter to their initial belief [52], [53]. Cognitive dissonance occurs when individuals have to cognitively reconcile their initial belief with this new information that seems to challenge or contradict what they already thought was true [52]–[54]. Cognitive dissonance leads to stress, anxiety, and feelings of discomfort [52], [53], [55]. As a result, individuals exert effort to resolve this contradiction and reach a level of internal psychological consistency.

A particular repair strategy is likely to reduce or exacerbate the effort needed to resolve this cognitive dissonance and reduce discomfort. Cognitive dissonance is likely to emerge when an individual has an initial positive attitude toward working with a robot and that robot violates their trust. The robot's trust violation creates a discrepancy between expectations and actual experiences, triggering cognitive dissonance. Once cognitive dissonance emerges, the individual has a strong drive to reduce this dissonance because of the general discomfort that this dissonance creates [56]. Cognitive dissonance can be reduced by either changing one's attitudes or minimizing the degree of contradiction between one's initial beliefs and experiences [54], [55].

Promises and explanations are likely to reaffirm an individual's positive attitude toward working with a robot. This is because promises to do better are assurances that an individual's initial belief was actually correct, while explanations provide a rational reason or justification for why they were not correct this particular time. This, in turn, allows individuals to hold onto their initial belief without the need to cognitively reconcile their initial positive belief with their actual experience with the robot. Apologies and denials further exacerbate the dissonance between what was expected and what was experienced. Apologies are a clear and explicit admission of guilt while denials further call into question the robot's integrity in addition to its ability, both of which should further reduce trust in the robot. This pushes humans to reassess their attitudes as the contradiction between expectations and observations is explicitly presented and reinforced.

Although reduction of cognitive dissonance can be achieved through either reducing the degree of inconsistency or changing one's beliefs, the method one adopts depends on how effortful a given method is [53], [56]. In the context of our study, it is likely that humans who possess higher pre-existing positive attitudes toward robots might find it easier to reconcile their initial positive belief with their actual experience with the robot when given a promise or an explanation. On the contrary, they would find it much more effortful to minimize the discrepancy between their pre-existing attitudes and experiences when presented with a denial or an apology that requires them to actively reconcile these differences.

> ***Hypothesis 1:*** *Human–robot trust repair strategies that seek to reduce dissonance via reducing the discrepancy between pre-existing attitudes and experiences (i.e. promises and explanations) are more effective for individuals who have more positive attitudes toward robots versus trust repair strategies that do not (i.e. apologies and denials).*

If a robot can make a mistake once, there is no reason to believe it cannot make another mistake. Unfortunately, little research has investigated how robust a particular repair strategy is over multiple violations [57]. In such cases, it is not clear how effective any particular trust repair strategy is after the first violation of trust. Existing research in the domain of human–human trust repair, however, provides some

insight into the trust repairs and trust violations over time. In particular, repeated errors decrease the efficacy of all trust repair strategies [21], [58]. This is the case because more frequent and continuous trust violations begin to be seen as normal behavior [59]. When this occurs it is also likely that the degree of positive attitude toward a robot does not matter because all individuals, both high and low in positive attitude, would eventually have to reconcile their initial beliefs with their actual experience.

> ***Hypothesis 2:*** *Human–robot trust repair strategies are less effective over repeated trust violations for those with more positive and less positive attitudes toward robots.*

Next, we describe the method used to investigate these hypotheses. In particular, we detail the task used in the study and the experimental apparatus we developed. Further, we also outline our experimental design, the variables examined, procedure, and participants.

IV. Methodology

A. Task

The task utilized in this study required collaboration between a participant and a robot. Both the participant and the robot shared the goal of processing a series of boxes in a warehouse. The robot was assigned the role of "picker" while the participant was the "checker," The robot acting as the picker moved boxes from a pile in a warehouse to the checker and, if approved, the robot would move the box to a nearby conveyor belt to be loaded and shipped. The participant was the checker and inspected boxes presented by the robot to make sure they were correctly selected. Correctly selected boxes were those that possessed a serial number that matched the serial number provided to the checker.

The task consisted of reviewing 10 boxes, with the experiment concluding after the 10th box. The robot made three errors in the form of picking a box with an incorrect serial number and presenting it to the participants. Each error took place at three specific time points —at box 3, 6, and 9 (see figure 1) —to give the robot in the study a reliability rate of 70%. This rate was consistent across conditions and was selected based on [60], which found that automation only increases performance with a reliability rate greater than 67%.

B. Experimental Apparatus

Participants performed the tasks in an interactive virtual environment developed in the UnReal Engine 4. The UnReal Engine was selected because of its graphics, asset options, and flexible deployment capabilities. The virtual environment was developed to resemble a realistic warehouse environment. As seen in figure 2, participants were positioned in the environment where two monitors and three buttons were made immediately visible. The monitors displayed the correct serial number, the time it took to approve or reject a box, and the participant's total score based on points gained for loading a correct box (+1) or lost for loading an incorrect box (-1). No points were given or deducted from participants' scores when they rejected a box presented by the robot. Points were visible at all times to encourage completion and attention.

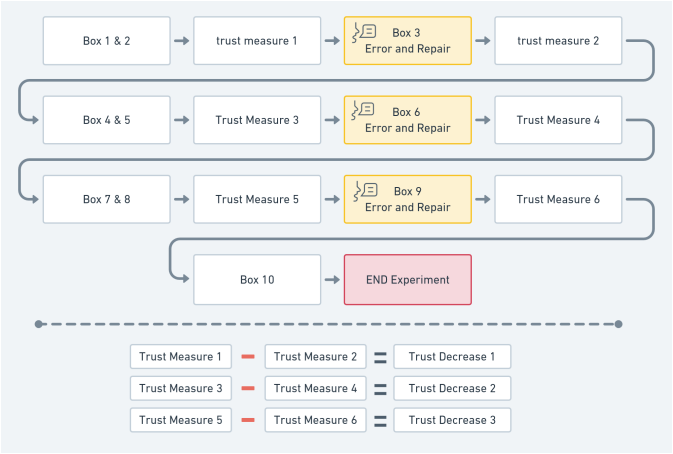

Fig. 1. Visual representation when our trust measure was deployed and how trust decrease was calculated.

C. Experimental Design

To examine our hypotheses, this study used a between- subjects experimental design with four repair strategy conditions and one no-repair condition. The repair strategy conditions examined apologies, denials, explanations, and promises. In the apology condition, the robot stated, *"I'm sorry I got the wrong box that time."* In the denial condition, the robot stated, *"I picked the correct box that time so something else must have gone wrong."* In the explanation condition, the robot stated, *"I see, that was the wrong serial number."* In the promise condition, the robot stated, *"I'll do better next time and get the right box."* Because less-than-perfect reliability is likely to result in multiple mistakes and not just one, we included errors at three points. As a result, messages were communicated three times in total (once after each violation of trust). These messages were communicated in both audio and on-screen text. Our control condition was the no-repair condition, where the robot remained silent throughout the experiment. A visual representation of our study's design is visible in figure 1.

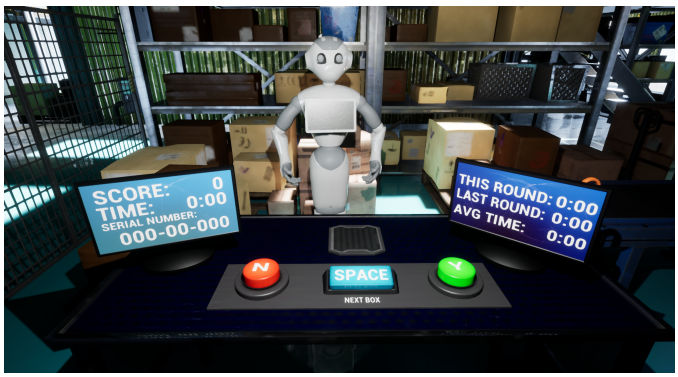

Fig. 2. Subject's perspective of virtual environment

Attitude Towards Work Robots (AWOR)	
—-	*I am someone who would:*
Q1	Enjoy working with a robot.
Q2	Be happy to receive work from a robot.
Q3	Find it fun to give work to a robot to perform
Q4	Like to collaborate with a robot to accomplish my work.
Q5	Find it fun to work with a robot.
Q6	Prefer to work with a robot.

TABLE I
ATTITUDE TOWARDS WORKING WITH ROBOTS (AWOR)

D. Variables

1) Independent and Control Variables: The independent variables used in this study were participants' positive attitude toward working with robots (AWOR) and repair condition. We measured AWOR via a scale based on [10]. This scale comprised six questions and was deployed as part of the study's pre-test procedure. The individual items for the AWOR scale are presented in table I. Repair strategies varied by the assigned condition and each participant was assigned to only one repair condition throughout the experiment. For the control variable, we measured trust propensity using six items taken from [61]. The individual items used for trust propensity are presented in appendix I.

2) Dependent Variable: The dependent variable examined in this study was participant's trust decrease. We calculated this decrease by subtracting trust before a violation from trust after a violation and its repair. To accomplish this, we relied on a three-item trust scale based on [46] in Appendix I. This scale was deployed at six time points and was accompanied by attention-check questions. Figure 1 illustrates when this measure was deployed and how trust decrease was calculated.

Because decreases in trust were the variable of interest rather than change in general, we took all positive difference values (trust decrease 1/2/3) and converted them to 0. In doing so, samples with a 0 indicated no decrease in trust and samples with negative values indicated a decrease in trust. To verify whether this approach was valid, we tested all samples that were converted to 0 to determine whether a significant difference existed between these responses prior to the violation versus after a violation with repair. Results of this test showed no significant differences between the pre- and post- violation trust scores, indicating that the conversion of the samples to 0 was valid.

E. Procedure

Recruitment took place via Amazon Mechanical Turk, where participants were presented with a human information task, or HIT. Upon acceptance of this HIT participants were first screened to determine whether they had participated in any prior conditions and were then presented with a link to a training scenario where they were familiarized with the virtual environment and the interface. The training scenario demonstrated the box task by giving participants one correct box and one incorrect box accompanied with dialogue. The dialogue communicated what button to press when the box was correct and what button to press when it was incorrect and the consequences of each action for the score.

After this training scenario was complete, a pre-test survey was deployed containing a general demographic questionnaire, trust propensity, and AWOR measures. After completing this pre-test survey, participants were assigned a scenario and proceeded to progress through the 10-box picking and checking task. As visible in figure 1, participants were given incorrect boxes at boxes 3, 6, and 9. Prior to each of these boxes (i.e. after boxes 2, 5, and 8), participants were presented with our trust measure. This measure was once more deployed after boxes 3, 6, and 9.

After the subjects had completed all 10 box tasks, they were asked to enter their worker identification (ID) for payment, which concluded their participation in the experiment. Throughout this process we implemented quality and attention -check questions. These took the form of randomly placed questions requesting a specific response from participants. If participants provided incorrect responses to these questions, their participation was immediately terminated and their data were excluded from our analysis.

F. Participants

For this study, we recruited a total of 100 participants (20 per condition) via Amazon Mechanical Turk. Participants were not allowed to participate more than once in this experiment. Across all conditions, ages ranged between 22 and 71 with a mean age of 38. Participants were compensated at a minimum rate of $15/hr, with the study's duration lasting 15–25 minutes. This research complied with the American Psychological Association Code of Ethics and was approved by the institutional review board at the University of Michigan. Informed consents were gathered upon participants' acceptance of the HIT.

V. MANIPULATION CHECK

A major assumption in our study's design is that participants noticed the robot's trust violations and these violations impacted the participants' trust in the robot. To verify that trust violations actually led to trust decreases, we compared trust decreases between the no-error condition and the error-with-no-repair (no-repair) condition. In the no-error condition, the robot had perfect performance and always returned the correct box, whereas in the no-repair condition the robot made the same errors as in the treatment conditions but offered no repair. When comparing trust decrease between these conditions, results of a Welch two-sample t-test after the first error suggested that there was a statistically significant difference between the no-error and no-repair conditions (difference = –1.08, 95% confidence interval [CI] [0.55, 1.62], $t(22.81) = 4.19$, $p < 0.001$; $d = 1.76$, 95% CI [0.78, 2.71]). In addition, for trust decrease after the second error,results suggested that the there was a statistically significant difference between the

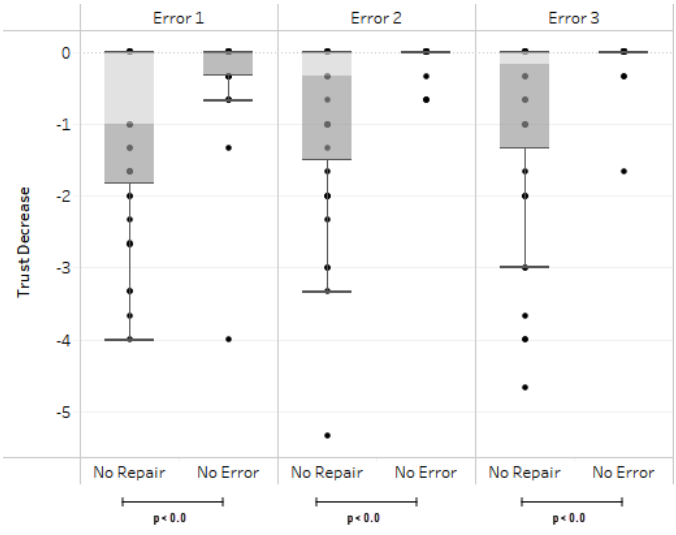

Fig. 3. Manipulation check showing differences in trust decrease between no repair and perfect performance conditions after all three errors.

no-error and the no-repair conditions (difference = –0.93, 95% CI [0.37, 1.50], t(20.59) = 3.42, p = 0.003; d = 1.51, 95% CI [0.52, 2.47]). Finally, for trust decrease after the third error, results once again suggested that the there was a statistically significant difference between the no-error and the no-repair conditions (difference = –1.00, 95% CI [0.37, 1.63], t(22.07) = 3.29, p = 0.003; d = 1.40, 95% CI [0.46, 2.32]). Notably, we also observed slight decreases in trust in the no-error condition for all three errors. This decrease, however, was non-significant in all cases. Taken together, these results show that we were successful at manipulating trust. Figure 3 summarizes these results.

VI. Results

To examine the hypotheses, we initially analyzed the data using a repeated measures analysis of covariance (ANCOVA), then followed up with three separate one-way ANCOVAs (one per trust violation). All the analyses included main effects , with trust propensity as a control variable and repair condition and AWOR as main effects. The interaction effects analysis included the main effect variables and an interaction term involving repair condition and AWOR.

A. Repeated Measures ANCOVA Results

We conducted a repeated measures ANCOVA to examine the overall trust decreases associated with the same participant over the three errors (error 1, error 2 and error 3). Results indicated that the main effect associated with when the error occurred was non-significant (Wilks Λ = 0.99, F = 0.64, p = 0.53). We also conducted additional investigations of interaction effects. In particular, we examined whether the influence of AWOR on the effectiveness of a particular repair strategy differed over the three errors.

Results suggested that the interaction between when the error occurred and AWOR (Wilks Λ = 0.98, F = 0.87, p = 0.42) as well as between when the error occurred and trust propensity (Wilks Λ = 0.99, F = 0.04, p = 0.96) were also non-significant. For the interaction between when the error occurred and the repair condition, however, a significant effect emerged (Wilks Λ = 0.79, F = 2.79, p = 0.01). Furthermore, when investigating the three-way interaction effect of when the error occurred, repair condition, and AWOR, we observed a significant effect, as well (Wilks Λ = 0.81, F = 2.49, p = 0.02). This final result indicates that the impacts of repair and AWOR on trust decrease differs by error. To examine this potential relationship in more detail, we subsequently conducted three follow-up one-way ANCOVAs, detailed next. These results and all subsequent results are presented in table II.

B. Trust Decrease - Error 1

Results examining trust decrease after the first error (power of α = 0.95) suggested that the main effect of repair condition was not statistically significant (F(4, 93) = 2.32, p = 0.063; η_p^2 = 0.09, 90% CI [0.00, 0.17]). In addition, the main effect of AWOR was also not statistically significant (F(1, 93) = 0.41, p = 0.522; η_p^2 = 0.004, 90% CI [0.00, 0.05]). Trust propensity, however, was statistically significant (F(1, 93) = 7.29, p = 0.008; η_p^2 = 0.07, 90% CI [0.01, 0.17]). Results examining the interaction between repair condition and AWOR showed that this effect was not statistically significant (F(4, 89) = 0.57, p = 0.686; η_p^2 = 0.02, 90% CI [0.00, 0.06]). These results are summarized in table II while the means for main effects are available in Appendix II figure 1 and an accompanying interaction plot is presented in Appendix II figure 4.

C. Trust Decrease - Error 2

1) Main Effects: Results examining the main effects of trust decrease after the second error (power of α = 0.99) suggested that the main effect of repair condition was not statistically significant (F(4, 93) = 1.24, p = 0.30; η_p^2 = 0.05, 90% CI [0.00, 0.11]). In addition, the main effect of AWOR was not statistically significant (F(1, 93) = 0.33, p = 0.568; $\eta_p^2 <$ 0.01, 90% CI [0.00, 0.05]). Trust propensity, however, was statistically significant (F(1, 93) = 11.13, p $<$ 0.01; η_p^2 = 0.11, 90% CI [0.03, 0.21]). These results are summarized in table II and means are available in Appendix II figure 2.

2) Interaction Effects: Results examining the interaction between repair condition and AWOR were statistically significant (F(4, 89) = 2.81, p = 0.030; η_p^2 = 0.11, 90% CI [0.01, 0.20]). These results are summarized in table II.

Based on this significant interaction effect, we conducted a post hoc power analysis. The results of this analysis indicated a high level of statistical power associated with this finding ($\alpha = 0.996$). Given these results, we then moved to an examination of slopes. In doing so, we tested whether the slopes were significantly different from zero, conducted a pairwise comparison between slopes, and produced an interaction plot to visually examine these relationships. To test whether slopes were significantly different from zero we conducted a simple slopes test via the reghelper package in R [62]. The results of this test found that the slope of promises was significant (p =

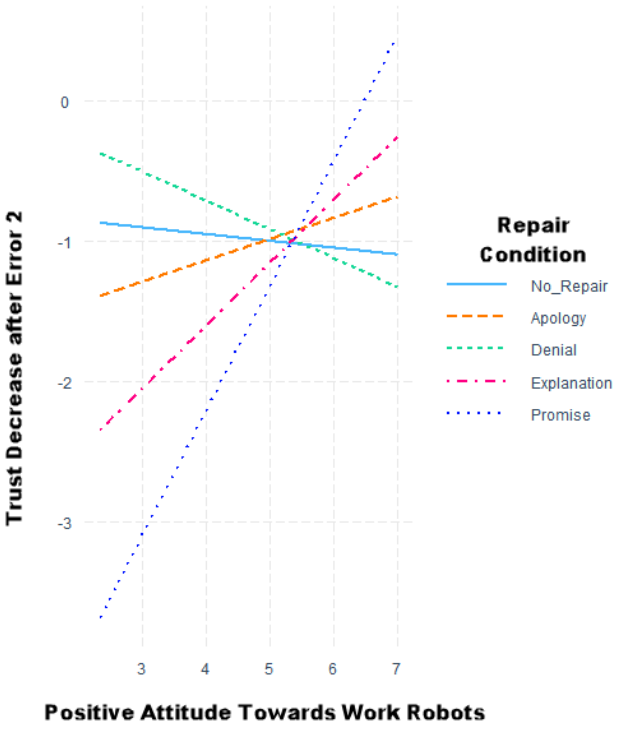

Fig. 4. Interaction plot showing the interaction effect between positive attitude and repair condition after a robot's second error.

0.001). To examine differences between slopes, we conducted a pairwise comparison of slopes via the emmeans package in R [63]. Results indicated significant differences between promises and the no-repair conditions (p = 0.049) and between promises and denials (p = 0.023). No other comparisons were significant.

Two significant trends emerged when we investigated these slopes. First, an investigation of slopes indicated that the positive impact of promises (i.e. less decrease in trust after an error) was more prominent when participants possessed higher AWOR, whereas when participants possessed less AWOR this strategy was not only less effective but possibly the least effective repair strategy examined. Second, when comparing denials to promises, we found significant differences between these slopes. Where promises appeared to increase in efficacy when AWOR increased, denials appeared to decrease when AWOR increased. A similar trend was evident when we examined the no-repair condition (see figure 4).

D. Trust Decrease - Error 3

Results examining the main effects of trust change decrease after the third error and repair (power of α = 0.73) suggested that the main effect of the repair condition is not statistically significant (F(4, 93) = 0.46, p = 0.766; η_p^2 = 0.02, 90% CI [0.00, 0.04]). In addition, the main effect of AWOR was not statistically significant (F(1, 93) = 0.95, p = 0.333; η_p^2 = 0.01, 90% CI [0.00, 0.07]). The main effect of trust propensity,however, was statistically significant (F(1, 93) = 7.38, p = 0.008; η_p^2 = 0.07, 90% CI [0.01, 0.17]). Finally, the interaction between repair condition and AWOR was not statistically significant (F(4, 89) = 0.11, p = 0.977; η_p^2 = 5.11e-03, 90% CI [0.00, 0.00]). These results are summarized in table II; the means for main effects are visually presented in Appendix II figure 3, and an accompanying interaction plot is available in Appendix II figure 5.

VII. Summary of Findings

The goal of this paper was to examine whether individual attitudes moderated the effectiveness of human–robot trust repair and to investigate whether this impact varies over multiple interactions. To that end, this study's results can be organized into two overarching contributions to the literature. One, individual attitude was found to moderate the effectiveness of human–robot trust repair; two, this effect differed over successive robot trust violations. Next, we discuss the implications of these findings for research and practice.

VIII. Discussion

Overall, the goal of this paper was to identify and examine individual AWOR as an important contingency variable in understanding HRI trust repair efficacy. Our results show that the efficacy of human–robot trust repair strategy can vary by an individual's AWOR and that this effect itself changes over the course of repeated trust violations. Next, we discuss specific contributions to the literature.

First, this paper offers a theoretical rationale for and provides the first empirical test of whether individual attitudes toward robots can influence the effectiveness of repair strategy. Cognitive dissonance helps to explain why attitudes toward robots are important to the HRI trust repair literature. For example, promises were most effective at repairing trust when positive attitudes were high and were least effective when positive attitudes were low. This might imply that the effects of promises are directly tied to an individual's pre-existing attitude toward robots. Promises reaffirmed an individual's positive attitude by providing assurances that the attitude was correct, thereby reducing cognitive dissonance. Notably, for those with higher positive attitudes, this dissonance might be stronger because the gap between expectations and experiences is wider than for those with lower positive attitude toward robots. This could create a stronger desire to reduce dissonance and encourage humans to more readily believe a robot's promises because these promises offer an easy way to reduce dissonance and psychological discomfort.

Second, this paper helps to explain the mixed results associated with prior literature and helps us determine which repair strategy might be more or less appropriate for a specific individual. One reason for the mixed results is that the efficacy of repair strategies differs significantly by individual. For example, once we consider the impact of individual differences on the studies examining apologies that found [29]–[31], negative [35] and non-significant [32]–[34] effects on trust repair we might discover that apologies are always positive for specific individuals, always negative for other individuals and non-significant for yet another set of individuals. For

	Repeated Measures ANCOVA		One-Way ANCOVAs					
	Trust Decrease		Trust Decrease Error 1		Trust Decrease Error 2		Trust Decrease Error 3	
	F	Pr>(F)	F	Pr>(F)	F	Pr>(F)	F	Pr>(F)
Repair Condition^{+}	2.68	0.01	2.32	0.07	1.24	0.26	0.46	0.78
Positive Attitude^{+}	0.87	0.42	0.41	0.52	0.33	0.55	0.95	0.34
Trust Propensity^{+}	0.04	0.96	7.29	0.01	11.1	0	7.38	0.01
Repair x Attitude^{+}	2.49	0.01	0.57	0.69	2.81	0.03	0.11	0.98
Error Time	0.64	0.53	–					

$^{+}$*By error time for repeated measures analysis.*

TABLE II
RESULTS OF REPEATED MEASURES ANCOVA AND SUBSEQUENT ONE-WAY ANCOVAS FOR TRUST DECREASE AFTER ERROR 1, 2, AND 3.

example, Wang et al. [37] found that promises alone had no effect on trust, while Esterwood and Robert [1] observed that promises were effective at promoting specific elements of trustworthiness and not others. If these studies had taken into account the influence of individual differences, they might have found promises to be just as effective or ineffective.

Third, our results indicated that the influence of individual differences was not uniform across repair strategies. For example, the influence of AWOR on promises was significantly different from that of denials. In particular, while promises were more effective when AWOR was high, denials were more effective when AWOR was low. Our results also suggest that when AWOR is low, denials are likely to outperform apologies. This conflicts with a growing consensus in the literature that apologies outperform denials [1], [24], [31], [34], [36]. These studies, however, have not taken into account individual differences (e.g., AWOR). Therefore, future studies could re-examine the effectiveness of apologies versus denial while accounting for the influence of individual differences. This finding also adds to the literature on individual differences and automation trust [38]–[44], [64] by highlighting the unique characteristics of this specific individual difference.

Four, this paper contributes to the literature through its examination of repeated trust violations. Initially, we expected trust repair strategies would be more effective after the first trust error and then become less effective after the second and third errors. Findings, however, indicated that neither the trust repair strategy nor AWOR was significant for the first error, although trust repair strategy was nearly significant at p = 0.07. Regardless, these results might mean that for the first error humans are still calibrating their trust in the robot, for the second error humans have decided on how much they trust the robot, and for the third error humans are again re-calibrating their trust. As a result, it could be that only when humans have decided on how much they trust a robot can trust repair strategies be influential. This suggests that existing studies that only examined one error might not be seeing the full picture. In particular, there might be an incorrect assumption that either robots only make one error or that the efficacy of repair strategies stays consistent. Given our results, this assumption seems unlikely, and therefore it might be that strategies appearing non-significant in the existing literature (e.g., [32]–[34], [37]) are actually effective when examined after a second trust violation.

IX. LIMITATIONS AND FUTURE RESEARCH

This study has several limitations. First, it relied on virtual representation of physical robots. Although this approach offers greater flexibility, it remains possible that virtual rep- resentations have weakened participants' degree of engagement and immersion. Future research could be done to replicate our findings with physical robots in a real-world setting. Second, our study relied on measurements of trust at six evenly spaced points throughout the study. The reason for this was to avoid breaking participants' engagement with the robot and task. Future studies could adopt less obtrusive measures of trust so as to measure trust at every interaction. Finally, this study did not directly measure cognitive dissonance; therefore, future studies might wish to use a scale to validate the relationships identified in this paper. Finally, future research could examine other measures of individual differences from the automation literature, such as personality, culture, perfect automation schema and others [38]–[43], [64].

ACKNOWLEDGMENTS

We would like to thank the Emerging Technology Group at the University of Michigan's James and Anne Duderstadt Center. Particularly, we wish to thank Sara Eskandari and Stephanie O'Malley for the development of our experimental platform.

REFERENCES

[1] C. Esterwood and L. P. Robert, "Do you still trust me? human-robot trust repair strategies," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 183–188.

[2] N. Savela, M. Kaakinen, N. Ellonen, and A. Oksanen, "Sharing a work team with robots: The negative effect of robot co-workers on in-group identification with the work team," *Computers in human behavior*, vol. 115, p. 106585, 2021.

[3] T. Haidegger, M. Barreto, P. Gonçalves, M. K. Habib, S. K. V. Ragavan, H. Li, A. Vaccarella, R. Perrone, and E. Prestes, "Applied ontologies and standards for service robots," *Robotics and Autonomous Systems*, vol. 61, no. 11, pp. 1215–1223, 2013.

[4] C. Esterwood and L. Robert, "Robots and COVID-19: Re-imagining human-robot collaborative work in terms of reducing risks to essential workers," *ROBONOMICS: The Journal of the Automated Economy*, vol. 1, pp. 9–9, 2021.

[5] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[6] 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.

[7] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler, "Toward an understanding of trust repair in human-robot interaction: Current research and future directions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 4, pp. 1–30, 2018.

[8] E. J. De Visser, R. Pak, and T. H. Shaw, "From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction," *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, 2018.

[9] T. Gompei and H. Umemuro, "Factors and development of cognitive and affective trust on social robots," in *International Conference on Social Robotics*. Springer, 2018, pp. 45–54.

[10] L. P. Robert, "A measurement of attitude toward working with robots (awro): A compare and contrast study of AWRO with negative attitude toward robots (nars)," in *Human-Computer Interaction. Interaction Techniques and Novel Applications*, M. Kurosu, Ed. Cham: Springer International Publishing, 2021, pp. 288–299.

[11] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, ""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)*, 2019, pp. 57–65.

[12] G. M. Alarcon, A. M. Gibson, and S. A. Jessup, "Trust repair in performance, process, and purpose factors of human-robot ttust," in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 2020, pp. 1–6.

[13] S. Engelhardt and E. Hansson, "A comparison of three robot recovery strategies to minimize the negative impact of failure in social HRI," 2017.

[14] T. Jensen, Y. Albayram, M. M. H. Khan, M. A. A. Fahim, R. Buck, and E. Coman, "The apple does fall far from the tree: User separation of a system from its developers in human-automation trust repair," in *Proceedings of the 2019 on Designing Interactive Systems Conference*, ser. DIS '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1071–1082.

[15] R. Luo, C. Huang, Y. Peng, B. Song, and R. Liu, "Repairing human trust by promptly correcting robot mistakes with an attention transfer model," *arXiv preprint arXiv:2103.08025*, 2021.

[16] 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.

[17] M. Nayyar and A. R. Wagner, "When should a robot apologize? understanding how timing affects human-robot trust repair," in *International conference on social robotics*. Springer, 2018, pp. 265–274.

[18] A. Costa, D. Ferrin, and C. Fulmer, "Trust at work," *The Sage handbook of industrial, work & organizational psychology*, pp. 435–467, 2018.

[19] K. T. Dirks and D. P. Skarlicki, "The relationship between being perceived as trustworthy by coworkers and individual performance," *Journal of Management*, vol. 35, no. 1, pp. 136–157, 2009.

[20] R. M. Kramer and R. J. Lewicki, "Repairing and enhancing trust: Approaches to reducing organizational trust deficits," *Academy of Management annals*, vol. 4, no. 1, pp. 245–277, 2010.

[21] R. J. Lewicki and C. Brinsfield, "Trust repair," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 4, pp. 287–313, 2017.

[22] 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.

[23] L. Dai and Y. Wu, "Trust maintenance and trust repair," *Psychology*, vol. 06, no. 06, p. 767–772, 2015.

[24] D. B. Quinn, "Exploring the efficacy of social trust repair in human-automation interactions," Master's thesis, Clemson University, 5 2018.

[25] V. R. Waldron, "Encyclopedia of human relationships," in *Apologies*, 1st ed., ser. 1, H. T. Reis and S. Sprecher, Eds. Thousand Oaks, CA: Sage Publishing Inc., 2009, vol. 3, ch. Apologies, pp. 98–100.

[26] 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.

[27] 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.

[28] N. Du, J. Haspiel, Q. Zhang, D. Tilbury, A. K. Pradhan, X. J. Yang, and L. P. Robert Jr, "Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload," *Transportation research part C: Emerging technologies*, vol. 104, pp. 428–442, 2019.

[29] Y. Albayram, T. Jensen, M. M. H. Khan, M. A. A. Fahim, R. Buck, and E. Coman, "Investigating the effects of (empty) promises on human-automation interaction and trust repair," in *Proceedings of the 8th International Conference on Human-Agent Interaction*, 2020, pp. 6–14.

[30] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 33–42.

[31] S. C. Kohn, A. Momen, E. Wiese, Y.-C. Lee, and T. H. Shaw, "The consequences of purposefulness and human-likeness on trust repair attempts made by self-driving vehicles," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 63, no. 1, pp. 222–226, 2019.

[32] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," in *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Institute of Electrical and Electronics Engineers (IEEE), 2010, Conference Proceedings, pp. 203–210. [Online]. Available: https://ieeexplore.ieee.org/document/5453195/

[33] E. S. Kox, J. H. Kerstholt, T. F. Hueting, and P. W. De Vries, "Trust repair in human-agent teams: The effectiveness of explanations and expressing regret," *Autonomous Agents and Multi-Agent Systems*, vol. 35, no. 2, 2021.

[34] S. C. Kohn, D. Quinn, R. Pak, E. J. De Visser, and T. H. Shaw, "Trust repair strategies with self-driving vehicles: An exploratory study," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 62. Human Factors and Ergonomics Society Inc., 2018, Conference Proceedings, pp. 1108–1112.

[35] D. Cameron, S. de Saille, E. C. Collins, J. M. Aitken, H. Cheung, A. Chua, E. J. Loh, and J. Law, "The effect of social-cognitive recovery strategies on likability, capability and trust in social robots," *Computers in Human Behavior*, vol. 114, pp. 106 561–106 561, 2021.

[36] X. Zhang, ""sorry, it was my fault": Repairing trust in human-robot interactions," Master's thesis, University of Oklahoma, 5 2021.

[37] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill, *Is It My Looks? Or Something I Said? The Impact of Explanations, Embodiment, and Expectations on Trust and Performance in Human-Robot Teams*. Springer International Publishing, 2018, pp. 56–69.

[38] J. B. Lyons and S. Y. Guznov, "Individual differences in human–machine trust: A multi-study look at the perfect automation schema," *Theoretical Issues in Ergonomics Science*, vol. 20, no. 4, pp. 440–458, 2019.

[39] V. L. Pop, A. Shrewsbury, and F. T. Durso, "Individual differences in the calibration of trust in automation," *Human factors*, vol. 57, no. 4, pp. 545–556, 2015.

[40] S.-Y. Chien, Z. Semnani-Azad, M. Lewis, and K. Sycara, "Towards the development of an inter-cultural scale to measure trust in automation," in *International conference on cross-cultural design*. Springer, 2014, pp. 35–46.

[41] S.-Y. Chien, M. Lewis, K. Sycara, J.-S. Liu, and A. Kumru, "Influence of cultural factors in dynamic trust in automation," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp. 002 884–002 889.

[42] J. L. Szalma and G. S. Taylor, "Individual differences in response to automation: The five factor model of personality." *Journal of experimental psychology: Applied*, vol. 17, no. 2, p. 71, 2011.

[43] S. M. Merritt and D. R. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *Human factors*, vol. 50, no. 2, pp. 194–210, 2008.

[44] C. Esterwood, X. J. Yang, and L. P. Robert, "Barriers to AV bus acceptance: A national survey and research agenda," *International Journal of Human–Computer Interaction*, pp. 1–13, 2021.

[45] C. Esterwood, K. Essenmacher, H. Yang, F. Zeng, and L. P. Robert, "A meta-analysis of human personality and robot acceptance in human-robot interaction," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–18.

[46] L. P. Robert, A. R. Dennis, and Y.-T. C. Hung, "Individual swift trust and knowledge-based trust in face-to-face and virtual team members," *Journal of management information systems*, vol. 26, no. 2, pp. 241–279, 2009.

[47] H. Azevedo-Sa, X. J. Yang, L. P. Robert, and D. M. Tilbury, "A unified bi-directional model for natural and artificial trust in human–robot collaboration," *IEEE robotics and automation letters*, vol. 6, no. 3, pp. 5913–5920, 2021.

[48] S. Ye, G. Neville, M. Schrum, M. Gombolay, S. Chernova, and A. Howard, "Human trust after robot mistakes: Study of the effects of different forms of robot communication," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2019, pp. 1–7.

[49] E. R. Smith, S. Sherrin, M. R. Fraune, and S. Šabanović, "Positive emotions, more than anxiety or other negative emotions, predict willingness to interact with robots," *Personality and Social Psychology Bulletin*, vol. 46, no. 8, pp. 1270–1283, 2020.

[50] M. M. De Graaf and S. B. Allouch, "Exploring influencing variables for the acceptance of social robots," *Robotics and autonomous systems*, vol. 61, no. 12, pp. 1476–1486, 2013.

[51] L. Sinnema and M. Alimardani, "The attitude of elderly and young adults towards a humanoid robot as a facilitator for social interaction," in *International Conference on Social Robotics*. Springer, 2019, pp. 24–33.

[52] L. Festinger, *A theory of cognitive dissonance*. Stanford University Press, 1957, vol. 2.

[53] A. McGrath, "Dealing with dissonance: A review of cognitive dissonance reduction," *Social and Personality Psychology Compass*, vol. 11, no. 12, p. e12362, 2017.

[54] L. E. Sullivan, *The SAGE glossary of the social and behavioral sciences*. Sage, 2009.

[55] P. A. Van Lange, A. W. Kruglanski, and E. T. Higgins, *Handbook of theories of social psychology: Volume two*. SAGE publications, 2011, vol. 2.

[56] D. T. Levin, C. Harriott, N. A. Paul, T. Zhang, and J. A. Adams, "Cognitive dissonance as a measure of reactions to human-robot interaction," *J. Hum.-Robot Interact.*, vol. 2, no. 3, p. 3–17, Sep. 2013.

[57] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T. M. Powers, C. Dixon, and M. L. Tielman, "Taxonomy of trust-relevant failures and mitigation strategies," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 3–12.

[58] 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.

[59] E. C. Tomlinson and R. C. Mryer, "The role of causal attribution dimensions in trust repair," *Academy of management review*, vol. 34, no. 1, pp. 85–104, 2009.

[60] J. R. Rein, A. J. Masalonis, J. Messina, and B. Willems, "Meta-analysis of the effect of imperfect alert automation on system performance," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Los Angeles, CA: SAGE Publications, 2013, pp. 280–284.

[61] S. A. Jessup, T. R. Schneider, G. M. Alarcon, T. J. Ryan, and A. Capiola, "The measurement of the propensity to trust automation," in *International Conference on Human-Computer Interaction*. Springer, 2019, pp. 476–489.

[62] J. Hughes, "Reghelper: Helper functions for regression analysis; r package version 1.0.2," 2021.

[63] R. Lenth, P. Buerkner, M. Herve, J. Love, H. Riebl, and H. Singmann, "Emmeans: Estimated marginal means, aka least-squares means. 2018; r package version 1.6.3," 2021.

[64] S. M. Merritt, J. L. Unnerstall, D. Lee, and K. Huber, "Measuring individual differences in the perfect automation schema," *Human factors*, vol. 57, no. 5, pp. 740–753, 2015.