Three Strikes and You are Out!: The Impacts of Multiple Human-Robot Trust Violations and Repairs on Robot Trustworthiness

Connor Esterwood, Lionel P. Robert

Highlights

- No repair strategy fully restored a robot's trustworthiness.
- Apologies, explanations, & promises did not restore perceptions of the robot's ability.
- Apologies, explanations, & promises did not restore perceptions of the robot's integrity.
- Apologies, explanations, & promises equally restored the robot's benev-olence.
- Denials were the least effective repair strategy.

Computers in Human Behavior, *forthcoming*

Three Strikes and You are Out!: The Impacts of Multiple Human-Robot Trust Violations and Repairs on Robot Trustworthiness

Connor Esterwood[a], Lionel P. Robert[a,b]

[a]*School of Information, University of Michigan, 105 S State St Ann Arbor, 48109, MI, United States of America*
[b]*Robotics Department, University of Michigan, 2505 Hayward St Ann Arbor, 48109, MI, United States of America*

Abstract

Robots like human co-workers can make mistakes violating a human's trust in them. When mistakes happen, humans can see robots as less trustworthy which ultimately decreases their trust in them. Trust repair strategies can be employed to mitigate the negative impacts of these trust violations. Yet, it is not clear whether such strategies can fully repair trust nor how effective they are after repeated trust violations. To address these shortcomings, this study examined the impact of four distinct trust repair strategies: apologies, denials, explanations, and promises on overall trustworthiness and its sub-dimensions: ability, benevolence, and integrity after repeated trust violations. To accomplish this, a between-subjects experiment was conducted where participants worked with a robot co-worker to accomplish a task. The robot violated the participant's trust and then provided a particular repair strategy. Results indicated that after repeated trust violations, none of the repair strategies ever fully repaired trustworthiness and two of its sub-dimensions: ability and integrity. In addition, after repeated interactions, apologies, explanations, and promises appeared to function similarly to one another, while denials were consistently the least effective at repairing trustworthiness and its sub-dimensions. In sum, this paper contributes to the literature on human—robot trust repair through both of these original findings.

Keywords: Human–Robot Interaction, HRI, Trust Repair,
Trustworthiness, Error Recovery, Ability, Benevolence, Integrity

January 10, 2023

Esterwood, C. and Robert, L. P. (2023). **Three Strikes and You are Out!: The Impacts of Multiple Human-Robot Trust Violations and Repairs on Robot Trustworthiness**, *Computers in Human Behavior,* forthcoming.

1. Introduction

Organizations increasingly rely on humans and robots to work collaboratively. For collaboration to be effective, however, adequate degrees of trust are vital [1–12]. An important determinant of trust is trustworthiness. Trustworthiness is the degree to which humans believe a robot is worthy of their trust [11, 13, 14]. A robot's trustworthiness, however, can be undermined when it violates the human's trust [15–17]. When this happens, different responses (i.e. trust repair strategies) to this violation have the potential to mitigate this negative impact [15, 17–20].

Scholars examining trust repair in human–robot collaboration have made significant advances, yet several important aspects of this phenomenon remain largely unexplored. First, the literature has focused on an overall measure of trust or trustworthiness. This approach fails to consider that a particular repair strategy might be more or less effective at repairing a specific sub-component of trustworthiness [21]. Two, researchers have typically examined the effectiveness of repair strategies after only one violation [21, 22]. This is problematic because in a real-world environment repair strategies are likely to be employed more than once and after repeated use their effectiveness is likely to change [23]. As a result it is important to understand the impact of repeat repairs as this can help determine which repair strategies are likely to be successful when repeated errors are likely to occur. Finally, no research to date has directly compared the effectiveness of any trust repair strategy against a no-trust violation or error-free condition. This makes it difficult to know whether or not trust repair strategies can restore trust fully. Therefore, this precludes us from understanding whether it is possible for trust to be fully restored to a pre-violation state after violations have occurred.

To address these shortcomings, we conducted a between-subjects experiment with 240 participants. Participants were placed in a collaborative work arrangement with a robot and tasked with sorting a series of boxes in a warehouse. Over this task, participants experienced 10 interactions with the robot along with three trust violations, each followed by a trust repair. During the experiment, we measured perceptions of the robot's trustworthiness, ability, integrity, and benevolence. Results from our analysis produced two overarching findings. First, across our trust repair strategies, no strategy was able to repair trustworthiness to a pre-violation level. Second, the effectiveness of each repair strategy, with the exception of repairing benevolence, appeared to be similar.

Overall, this paper provides theoretical contributions to the literature on human-robot trust repair in the following ways. One, after repeated trust violations despite their unique theoretical basis apologies, explanations, and promises appears to be as equally ineffective with regard to repairing trustworthiness, ability, and integrity. Two, apologies, explanations, and promises appear to be equally effective with regard to repairing the trustworthiness sub-component of benevolence but not integrity or ability despite their distinct theoretical connections. Finally, contrary to the theory of misinforming, denials delivered after multiple violations and repairs do not appear to positively impact trustworthiness nor ability, integrity and benevolence.

2. Literature Review

2.1. Trust & Trustworthiness

In this paper, we define trust as the willingness of the trustor to be vulnerable to the actions of the trustee [13]. Trustworthiness precedes and largely determines the trust a trustor places in a trustee [13, 14]. Trustors principally assess the trustworthiness of the trustee over time by learning from the outcomes of their interactions with the trustee. In this way, one can consider trustworthiness as learned rather than dispositional or situational factor. Trustworthiness has been shown to be largely influential in determining a human's trust in a robot [24].

Trustworthiness can be subdivided into three distinct components: ability, integrity, and benevolence [13, 25]. Ability is the degree to which a trustee is seen as skillful or competent within a specific domain [13]. Integrity is the degree to which a trustee is seen as honest and adherent to a set of principles [24]. Benevolence is the degree to which a trustee is seen as altruistic and as acting without conflicting egocentric or profit-based motives [13]. While some work on ability, benevolence, and integrity has been conducted in the HRI literature [24], less is known about how trust violations and repairs influence these components of trustworthiness.

2.2. Trust Repair

Trust repair can be defined as action taken by a trustor to help restore trust in them after they have committed a violation of trust [26]. These actions can take many forms but are frequently actualized through short-term verbal repairs such as apologies, denials, explanations, and/or promises [21, 27, 28]. Each of these approaches –i.e. trust repair strategies– can be embodied by related but distinct overarching theoretical frameworks. This paper identifies four such frameworks which can be used to organize the existing literature on HRI trust repair and their corresponding trust repair strategies. These four overarching theoretical frameworks are: 1) the *Theory of Forgiveness*, 2) the *Theory of Forgetting*, 3) the *Theory of Informing*, and 4) the *Theory of Misinforming*. In the following subsections, we introduce each of these theoretical basics of trust repair and discuss how they relate to the different trust repair strategies of apologies, denials, explanations, and promises.

2.2.1. Theory of Forgiveness

Seeking forgiveness is at the heart of one theoretical framework of trust repair. Forgiveness is a complex process through which relationships can be restored after trust violations occur [27]. At a conceptual level, forgiveness lacks a universally accepted definition but many authors agree that forgiveness is not pardoning, condoning, justifying, excusing, self-denial, or forgetting [27, 29]. Instead, one way to understand forgiveness is to consider it as a process through which a trustor "relinquishing anger, resentment, and the desire to seek revenge against someone who has caused harm" [30, Pg.251]. This often involves the trustor extending a benevolent orientation towards the trustee [31] and consciously moving away "from negative thoughts, feelings, and behaviors toward the transgressor (i.e. trustee) to more positive thoughts, feelings, and behaviors" [32, Pg.307]. One of the main results of this process is a restoration in good-will and the re-establishment of positive perceptions – such as benevolence – which, in turn, can lead to increased trustworthiness and by extension trust.

The "Theory of Forgiving" is often represented in the form of apologies. Apologies are expressions of regret or remorse [33]. In the human–human literature one way that apologies repair trust is by appealing to the emotions and affect of the trustor [27]. By doing so, apologies convey that the trustee is remorseful and feels bad about their actions. This may also promote forgiveness [27, 34, 35] as apologies signal that the trustee is emotionally invested in the relationship at hand. Furthermore, as apologies are primarily based on the social and emotional elements of a relationship they may signal that the trustee genuinely cares about the trustor. This may improve the trustor's perceptions of the trustee's benevolence and in-turn increase the likelihood of the trustor reciprocating with benevolent action of their own in the form of forgiveness.

Generally, the HRI literature has found apologies to be effective at repairing trust in some cases but ineffective in others. In particular, two studies have shown that apologies were effective at repairing trust between humans and robots [36, 37], three have found them to be ineffective [38–40] and one found that they were actually damaging to trust [41]. In the case of those who found that apologies repaired trust, [36] examined trust repair where the robot was a self-driving vehicle conducting a driving task and found that not only did the robot effectively repair trust via apologies but that apologies actually outperformed explanations. Similarly, [37] also found that apologies were effective and outperformed explanations but did so using four different types of robots (Pepper, Nao, Kuri, and Sawyer) which each conducted an information assistance task.

Conflicting with the above studies, three additional studies examining trust repair [38–40] each found that apologies had no significant impact on trust after violations occurred and one found that apologies actually damaged trust [41]. For those indicating non-significant results, two [38, 40] looked at the same type of robot and task as [36] while the remaining study [39] did not, instead using both a human-like (Snackbot) and machine-like (HERB) robot [39] each of which performed a service task (snack delivery). For the study indicating that apologies actually damaged trust [41], this study utilized a small ground-based machine-like robot (ROBO-GUIDE) that conducted an information assistance task. Ultimately the impact and general efficacy of apologies appear mixed across the literature indicating that moderators and/or additional contextual factors may be at play. Table 1 summarizes these findings.

2.2.2. Theory of Forgetting

Forgetting or forgetfulness forms the basis of another theoretical framework of trust repair. Forgetting or forgetfulness can be understood as the willingness of a trustor to overlook past negative acts in favor of future –and potentially positive– acts. Unlike, forgiveness which is past-oriented forgetfulness is future-oriented. This can repair trust by attempting to redirect the trustee's attention to the future behavior of the trustor encouraging the trustee to give the benefit of the doubt moving forward [42]. One key to the success of forgetfulness is the belief that past poor performance can be ignored and replaced with positive future performance expectations. To this end, statements that redirect the trustee's attention toward future performance can be especially useful in promoting forgetfulness.

The "Theory of Forgetting" can be represented by promises which are statements that attempt to shift the focus from past to future behaviors. Promises are statements that convey positive future performance [43]. Promises repair trust by seeking to explicitly set expectations for the future. As a result promises promote forgetfulness by shifting perceptions towards the future by attempting to guarantee future positive performance.

Four studies have directly assessed the efficacy of promises in repairing trust between humans and robots. In two of these cases, promises appeared to be effective [44, 45] and in two they did not [46, 47]. For the two finding that promises appeared to be effective at repairing trust, one examined a small machine-like ground robot engaging in a service task [45]. Similarly, the other study finding promises were effective also used a small machine-like ground robot but this robot was engaged in an information assistance task, and promises were only found to be effective when given immediately after violations occurred instead of given after time had passed [44]. For the studies that found non-significant results, the first [46] examined a dog-like and machine-like ground robot engaged in an information assistance task while the other [47] examined a self-driving vehicle engaged in a driving-related task. Once again, none of these studies considered trustworthiness as opposed to trust and further research is warranted given the diverging results present both for this and all previously mentioned repair trust repair strategies. Table 1 summarizes these findings.

2.2.3. Theory of Informing

Informing is at the core of another theoretical framework of trust repair. Informing can be understood as the process through which one conveys factual accounts of something in the world [48]. Informing can take many forms but the ultimate goal of informing is to facilitate accurate and true accounts of events [49, 50]. Informing can impact trust as the way that one interprets and makes sense of trust violations is directed by the information available [51]. Furthermore, informing can promote transparency which can increase trust and indeed act as a form of trust repair [52].

The "Theory of Informing" can be represented by explanations. Explanations are statements that seek to convey clear and direct reasoning behind why a violation of trust occurred [53]. Explanations repair trust by attempting to establish a shared account of what transpired [27, 34]. By establishing this shared account, explanations help a trustor make more informed judgments about the cause of the trust violation and if the trust violation is likely to re-occur [28, 54]. Additionally, explanations as a representation of informing can directly impact perceptions of transparency [55, 56]. This can ultimately lead to higher attributions of integrity on behalf of the trustee therefore making the trustee seem more trustworthy and less likely to be engaging in *"cheap talk"* or in-genuine behavior overall increasing trust [28, 57].

Within the HRI literature, explanations have been effective at repairing trust in some cases but not in others. For example, three studies found that explanations effectively repaired trust. The first of these [37] found that explanations were effective when given by each of four different robots after they provided inaccurate or incorrect advice. The second study [58] found that explanations were effective but that the type of explanation examined played a significant role in determine how effective they were. Specifically, [58] highlighted how explanations containing more details on why the violation occurred were more effective than less informative explanations or simple accounts. Finally, one additional study [41] found that explanations were effective at decreasing perceptions of a robot's deceitfulness but not its performance or integrity after failing at an information assistance task.

Aside from the aforementioned studies, seven additional studies found that explanations had no significant impact on trust repair [38–40, 47, 59–61]. These studies ranged in the type of robot used with 3 using self-driving vehicles [38, 40, 47] conducting driving-related tasks, 3 using small machine-like robots [59–61] conducting information assistance tasks, and one using multiple different types of robots [39] performing a service task. Taken together these non-significant and significant results indicate mixed results but largely lean towards explanations as ineffective. This is surprising given the general view that explanations are useful in promoting trust [53] but, more work is needed. Table 1 summarizes these findings.

2.2.4. Theory of Misinforming

Misinforming forms the basis of another theoretical framework of trust repair. Misinforming is the process of providing information that is "factually false," or inconsistent with the best available evidence [48]. Often the goal of misinforming is to incorrectly re-frame or bias perceptions of an event in such a way as to benefit an individual or organization [62]. By doing so, misinforming can impact trust as it acts as a direct method by which one can influence how another comprehends interactions and events [48, 51, 62]. In this way, misinforming is largely the opposite to informing but largely impacts trust in the same way.

The “Theory of misinforming” can be represented by denials. Denials are trust repair strategies that seek to redirect blame or reject culpability for a violation of trust [17]. Denials repair trust by seeking to establish the complete innocence of the trustee [28]. To accomplish this denials seek to fully shift blame (I.E. attribution as to the cause of a trust violation) away from the trustee and onto some other entity [27]. In doing so denials are “purely attributional in nature” [28, Pg.11]. From this perspective, we can consider denials as relying on the misattribution of a trust violation via misinforming (i.e. lying to) the trustor. Denials are most effective when there is a high degree of faith in the trustee and/or a high degree of doubt in the exact nature or causes of the trust violation [27].

HRI studies examining denials have generally found mixed results, with some showing that denials are effective at repairing trust [36] and others showing they are not [38, 47]. For the study that found that denials were effective at repairing trust [36] results indicated that denials had a positive impact on trust but less so than apologies. For those showing that denials were ineffective both showed non-significant results. Interestingly, all three of these studies examined the same type of robot and task, namely, self-driving vehicles conducting driving tasks. As a result, it is not clear if other types of robots and tasks would persist in producing mixed results or what differences or moderators may have influenced these findings. Table 1 summarizes these findings.

	Repairs Trust	**Does Not Repair Trust**	**Damages Trust**
Apology	[36, 37]	[38–40]	[41]
Denials	[36]	[38]	[47]
Explanation	[37, 41, 58]	[38–40, 47, 59–61]	–
Promise	[44, 45]	[47, 63]	–

Table 1: Table summarizing the efficacy of different repair strategies in the HRI trust repair literature.

3. Hypotheses

Prior literature has not sought to directly theorize or empirically examine how particular trust repair strategies may relate to different sub-dimensions of trustworthiness. In response, this paper approaches these gaps by presenting several related hypotheses and theoretical relationships. To do this, we developed an overarching theoretical research model. This model is summarized in figure 1 and draws from the prior literature on trust repair linking the trust repair strategies to different sub-dimensions of trustworthiness and trust repair frameworks.

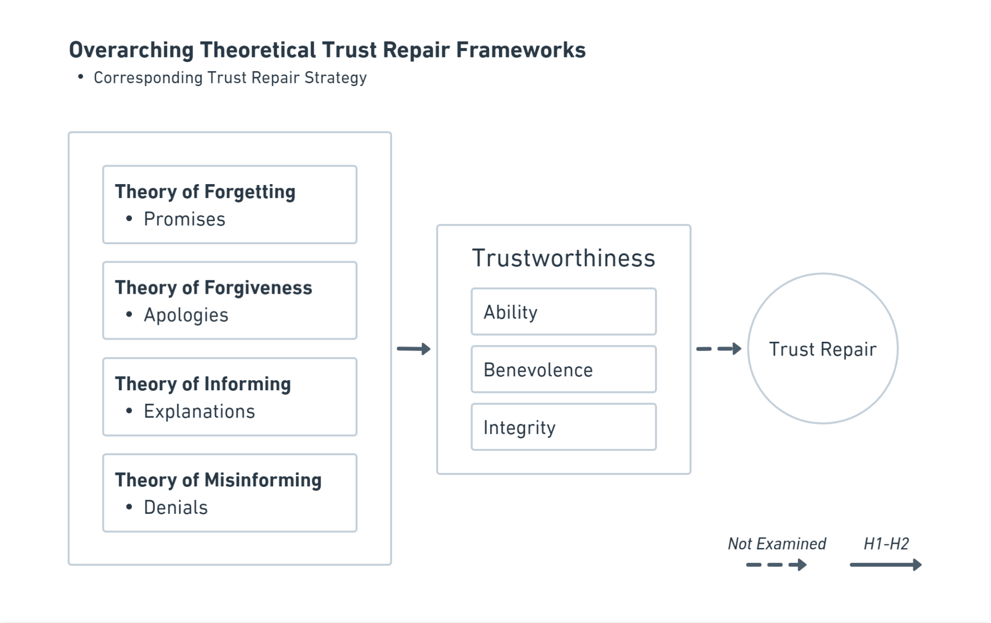

Figure 1: Theoretical research model proposed for this paper.

First and foremost, we embedded repeated trust violations into our theorizing for several reasons. Theoretically, we expect that differences between each repair strategy are likely to be more profound after repeated trust violations rather than after a single trust violation. The theoretical logic which underlies each theory of trust repair is best articulated and examined through repeated rather than one-time trust violations. This is because the theoretical mechanisms that underlie their differences are likely to become more salient after repeated trust violations [23]. Practically, there are likely to be smaller differences between any two trust repair strategies or between any trust repair strategy and the perfect condition after just one trust violation. In fact, one mistake is likely to be overlooked if no other mistakes occur there afterward. This may explain many of the mixed results associated with studies focusing on one trust violation (See: [21]). Contrary, we believe any differences in trust repair strategies after repeated trust violations are likely to be accurate, meaningful, and lasting. Taking this into account as well as the different elements of our theoretical model, we, therefore, propose the following hypothesis:

> ***Hypothesis 1:*** *Human-robot trust repair strategies should fully restore trustworthiness and its various sub-dimensions: ability, benevolence, and integrity after repeated trust violations.*

In addition to this overarching hypothesis, this paper also examines the elements of our theoretical research model in more detail. In particular, we follow up the above hypothesis with a series of more specific hypotheses. These are presented in the following sections of this paper.

3.1. Apologies & Repairing Trustworthiness

We propose that apologies, as a trust repair strategy, should have a stronger impact on the trustworthiness sub-dimension of benevolence than other trust repair strategies after repeated trust violations. This is because apologies work through relatively affective mechanisms [27]. Specifically, apologies convey a degree of emotional investment in a particular relationship [64, 65]. As a result, an apology may lead the trustee to believe that the trustor genuinely cares about them and is acting without an egocentric motive. These perceptions are generally associated with benevolence as both the belief that someone or something cares about you and is acting without an egocentric motive are core components of this concept. Therefore over-repeated trust violations apologies are likely to improve perceptions of benevolence over and above the existing trust repair strategies:

> ***Hypothesis 2a:*** *Apologies will be more effective at repairing benevolence than other repair strategies after repeated trust violations.*

3.2. Explanations & Repairing Trustworthiness

Explanations, as a trust repair strategy, should have a stronger impact on the trustworthiness sub-dimension of integrity than other trust repair strategies after repeated trust violations. This is because explanations do not seek to re-frame or influence perceptions of events but instead seek to provide transparency and accountability [66]. This may improve perception of the trustee as honest and morally consistent therefore bolstering perceptions of integrity. As a result, over-repeated trust violations explanations are likely to improve perceptions of integrity over and above the existing trust repair strategies leading us to our next hypothesis:

> ***Hypothesis 2b:*** *Explanations will be more effective at repairing integrity than other repair strategies after repeated trust violations.*

3.3. Promises & Repairing Trustworthiness

As a repair strategy, promises are likely to have a stronger impact on the trustworthiness sub-dimension of ability than other trust repair strategies after repeated trust violations. This is because promises repair trust through directly conveying that not only can a trustee do better moving forward but that the trustee will do so [43]. To convey that one will do better is to signal that their performance or ability to perform the task will improve. As a result, over-repeated trust violations promises are likely to improve perceptions of ability over and above the existing trust repair strategies leading us to the following hypothesis:

> ***Hypothesis 2c:*** *Promises will be more effective at repairing ability than other repair strategies after repeated trust violations.*

3.4. Promises & Repairing Trustworthiness

Finally, for denials, it is likely that this trust repair strategy will be less effective across all three trustworthiness sub-dimensions. This is because denials rely primarily on the degree of doubt present in the trustor's perception of events [67]. Given that the teaming task used in this study involves direct interactions and observations, we expect that this level of doubt will be relatively low. Therefore, denials are likely to be generally ineffective or even potentially damaging. As a result, over-repeated trust violations denials are likely to be the least effective trust repair strategy leading us to our final hypothesis:

Hypothesis 2d: *Denials will likely be the least effective at repairing ability, integrity, and benevolence than the other repair strategies after repeated trust violations.*

With these hypotheses in place, we now transition to a discussion of our methodology. In particular, we introduce the task, apparatus, experimental design, and variables examined in this paper. Furthermore, we also discuss our procedure and provide a breakdown of the sample collected and the broad demographic characteristics inherent within it.

4. Methodology

4.1. Task

The task used in this study was designed to mirror existing collaborations between humans and robots. For example, Amazon (a U.S. retailer) has begun testing a robot that "takes totes off of a robotic shelf and uses a robotic arm to deliver it [sic] to employees, so they can remain in a more comfortable, stable, and ergonomically friendly position" [68]. Using this and similar real-world examples as reference points, we assigned participants roles as "checkers" and robots the role of "picker." In these roles, the robot would pick boxes from a nearby queue and present them to the participant. Participants would then either approve or reject the boxes based on whether the serial number on the box matched the serial number presented on a nearby monitor. If participants approved the box, the robot would place the box on a nearby conveyor belt where it would be transported to another part of the warehouse environment. If participants rejected the box, the robot would return the box to the queue. Over the course of the study, 10 boxes were processed, with the robot picking the wrong box at three evenly distributed time points (box 3, box 6, and box 9). This produced a reliability rate of 70% based on [69]. The inclusion of three errors rather than just one single error was based on the assumption that imperfect robots are likely to make mistakes more than once over repeated interactions.

4.2. Apparatus

Participants engaged as members of the human–robot team via an interactive virtual environment developed in Unreal Engine 4. We chose to use a virtual environment over real-world robots for two reasons. First, our study took place during the coronavirus disease 2019 (COVID-19) pandemic, making in-person research studies difficult. Second, virtual representations of robots have been used by other studies on this topic and found to be adequately robust [38, 39, 44, 70]. This environment was modeled to appear as a realistic warehouse environment. Within this environment, participants were positioned behind a table containing two displays and three buttons. The displays showed the team's current score, how fast boxes were being processed, and the serial number participants needed to assess the box presented by their robotic teammate. Each team's score increased by 1 point every time a correct box was placed on the conveyor belt and was reduced by 1 point every time an incorrect box was placed on the conveyor belt. In cases where the robot picked the wrong box and participants flagged this as an error, an indicator appeared on-screen showing this box was incorrect but points were neither given nor deducted from the team's score. The team's score was used only to encourage engagement in the simulation and did not impact participants' compensation. Figure 2 shows the environment and workstation as presented to the participant, and Figure 3 illustrates how scores were calculated during the experiment.

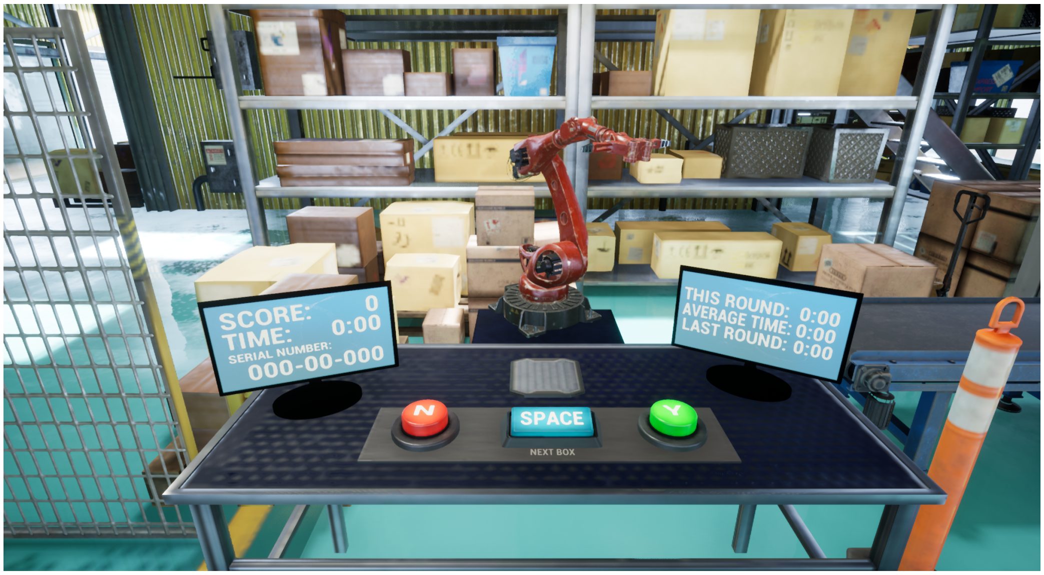

Figure 2: Point of view of the participant in the virtual environment.

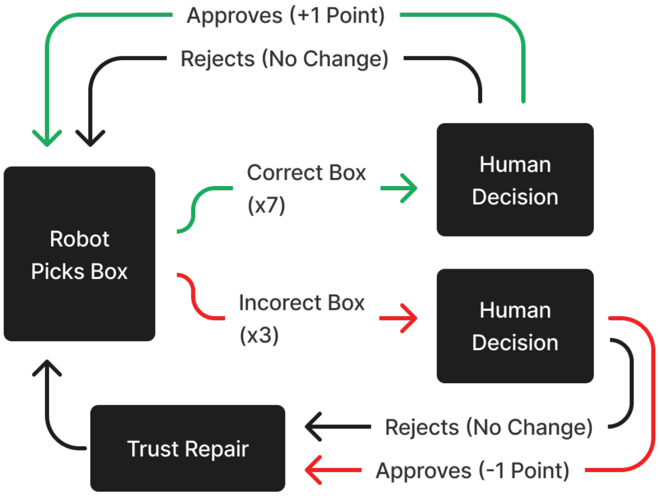

Figure 3: Flow diagram illustrates possible outcomes and scores based on the boxes the robot picks and the decisions the participant makes.

4.3. Experimental Design

To examine our hypotheses, we used a between-subjects design with four repair conditions and two control conditions. The control conditions took the form of the robot staying silent after making an error (no repair) or the robot performing perfectly at its task and committing no errors during the experiment (perfect performance). The repair conditions used in this study were apologies, denials, explanations, or promises. These were deployed after each error condition (box 3, box 6, and box 9). For the apologies condition, the robot stated, *"I'm sorry I got the wrong box that time."* For denials, the robot stated, *"I picked the correct box that time so something else must have gone wrong."* For explanations, the robot stated, *"I see, that was the wrong serial number."* For the promises condition, the robot stated, *"I'll do better next time and get the right box."* Each of these responses was designed to represent only one type of repair strategy and to avoid unintentionally combining two or more strategies. These repairs were communicated to participants via both audio and text subtitles during the experiment. Notably, the robot only temporarily changed its behavior after the repairs were delivered, retrieving correct boxes two additional times until the next error occurred.

4.4. Independent & Control Variables

The experimental condition that participants were assigned to was the independent or treatment variable in this study. As per the, between-subjects design, participants only encountered one of these conditions throughout their participation in this study.

4.5. Dependent Variables

In this study, we focused on trustworthiness overall and the three different trustworthiness sub-dimensions that comprise it, namely, ability, integrity, and benevolence. To accomplish this we developed a new trustworthiness questionnaire based on previous work where subjects rated their agreement with a series of statements on a 1-7 (agree/disagree) Likert scale [71–73]. This questionnaire was deployed at the end of the study for each subject (i.e. after the 10th box). The individual items used along with the details of the factor analyses for their validation can be found in Appendix A.

4.6. Procedure

Participants were presented with the opportunity to participate in this study via Amazon Mechanical Turk (M-Turk). Upon accepting the "HIT" (task) on M-Turk, they were provided a link to the virtual environment used in this study. They were then familiarized with the environment and interface via a brief tutorial. This tutorial featured the virtual environment, a generic mannequin, and guided text overlays that indicated the functions of all buttons and interactive components of this study. Participants were then presented with one correct box and one incorrect box by the mannequin and were guided by the tutorial on how to process these boxes appropriately.

After completing this tutorial, participants were given a pre-test questionnaire containing demographic information and a link to their assigned experimental condition. Participants were only afforded one condition and no repeat participants were permitted. Throughout the study, participants were presented with several attention-check questions to ensure the integrity of the data collected. Attention-check questions are questions embedded in the questionnaire that ask for a specific response and therefore flag any participants who select the wrong answer. These questions help to ensure the integrity of data because only participants who read each question are able to discern their presence and answer them correctly, indicating that sufficient thoughtfulness and attention was paid during the questionnaire. If participants failed any of these questions their data were excluded from analysis, participation was immediately ended, and no payment was given. Upon completion of their assigned experimental condition participants were then presented with a post-test questionnaire containing our trustworthiness measure. After this questionnaire's completion subjects were given an exit code, paid, and dismissed.

4.7. Participants

Two hundred forty participants were recruited for this study (40 per experimental condition). Across all conditions, participants' ages ranged 22–78 years, with a mean age of 40. In addition, 64% of participants identified as male while 36% identified as female. Participants were compensated at a minimum rate of $15/hr, with the study's duration lasting 15–25 min. This research complied with the American Psychological Association Code of Ethics and was approved by the institutional review board at the BLINDED FOR REVIEW. Informed consent was gathered upon participants' acceptance of the HIT.

5. Results

To investigate these hypotheses, we first validated our measure of trustworthiness via factor analysis and reliability testing. After the validation of this instrument was confirmed we then used a series of non-parametric Kruskal-Wallis rank sum tests followed by post hoc Dunn's tests of multiple comparisons with a Benjamini-Hochberg correction to control for multiple hypothesis testing. We selected these methods over others because data in this study were non-normally distributed. The first of these tests examined our manipulation of trustworthiness by comparing differences in trustworthiness between the perfect performance condition and the no-repair condition. The second used three separate Kruskal-Wallis tests followed by post hoc examinations to determine participants' ratings of ability, benevolence, and integrity across repair conditions. The analysis code and data used are available upon request and supplementary tables of means, medians, and standard deviations are visible in Appendix B.

5.1. Measurement Validation

All items meet or exceed the benchmark criteria of ≥ 0.7 for construct reliability [74]. Item reliabilities include $\alpha = 0.76$ for ability, $\alpha = 0.95$ for integrity, $\alpha = 0.92$ for benevolence and $\alpha = 0.92$ for trustworthiness.

Discriminant and convergent validity were first assessed by conducting a thorough exploratory factor analysis (EFA). All items loaded at 0.7 or above on their corresponding construct with cross-loading less than 0.4 or higher with the exception of 'Ability 2' (See: Appendix A for item text). To determine if we should keep this item for construct face validity we conducted a confirmatory factor analysis (CFA) using `lavaan` in R studio [75]. A CFA is similar to EFA, with the additional restriction of imposing a structure on the data [76]. The prominent fit index used to measure fit or misfit is the comparative fit index (CFI) [77]. Values of the CFI range from 0, indicating no fit, to 1, indicating a perfect fit. CFI values ≥ 0.95 are considered to be an indication of a good fit [76]. The CFI value for our measurement model was 0.97, indicating a sufficient fit.

5.2. Manipulation Check

To test our manipulation of trustworthiness we compared trustworthiness overall between a no repair and a perfect performance condition. In the no repair condition, the robot made errors on the 3rd, 6th, and 9th boxes but did not implement a trust repair strategy. In perfect performance condition, the robot made no errors and by extension delivered no repairs. By comparing these two conditions on the basis of trustworthiness we sought to determine if the trust violations present in our study's design were effective (i.e. decreased trustworthiness). Using a Kruskal-Wallis test comparing trustworthiness in the no repair condition to trustworthiness in the perfect performance condition, we observed a significant difference between these two conditions ($p < 0.001$, $\chi^2 = 19.2$, $\eta^2 = 0.23$) such that trustworthiness was significantly lower in the no repair condition than in the perfect performance condition. Figure 4 illustrates this relationship. Overall, these results show that when a robot makes a mistake, trustworthiness decreases. Furthermore, this decrease was statistically significant. This leads us to therefore conclude that our manipulations of trustworthiness in this study was effective and functioned as intended.

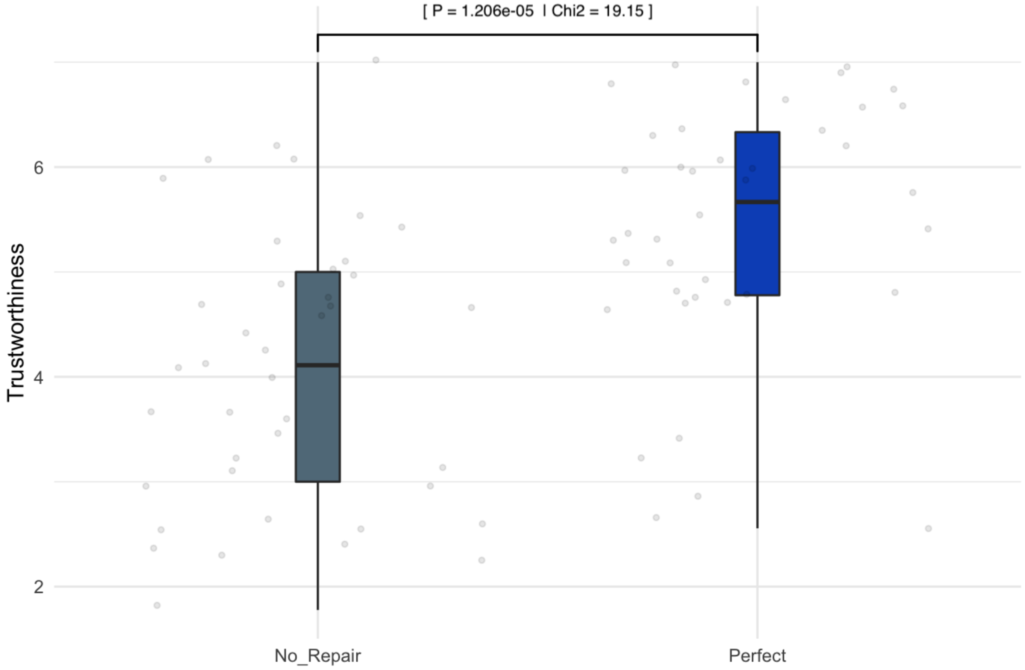

Figure 4: Graph comparing trustworthiness for subjects in the no repair condition and trustworthiness for subjects in the perfect performance conditions with the results of a Kruskal-Wallis test annotated at the top of the figure.

5.3. Trust Repair and Trustworthiness

To examine the effect of different trust repair strategies on trustworthiness, we conducted a Kruskal-Wallis test followed by a post hoc Dunn's test of multiple comparisons. Results of the Kruskal-Wallis test showed significant differences among repair strategies overall ($p = 1.07e - 10$, $\chi^2 = 55.41$, $\eta^2 = 0.23$) allowing for post-hoc analysis. Results of this post-hoc analysis indicated significant differences between the perfect condition and all other conditions as well as the denial conditions and all other conditions. These results are summarized in Table 2 and a visual presentation of means and standard deviations is visible in Figure 5.

In general, these findings failed to support H1 that "human-robot trust repair strategies should fully restore trustworthiness and its various subdimensions: ability, benevolence, and integrity after repeated trust violations". Specifically, no trust repair strategy was significantly more effective

Group 1	Group 2	Mean Rank 1	Mean Rank 2	Mean Rank Diff	p.adj	p.adj.signif
No Repair	Perfect	116.14	183.46	67.33	<0.01	***
No Repair	Apology	116.14	115.84	-0.3	1.00	ns
No Repair	Denial	116.14	69.18	-46.96	0.02	*
No Repair	Promise	116.14	123.00	6.86	1.00	ns
No Repair	Explanation	116.14	115.39	-0.75	1.00	ns
Perfect	Apology	183.46	115.84	-67.62	<0.01	***
Perfect	Denial	183.46	69.18	-114.29	<0.01	***
Perfect	Promise	183.46	123.00	-60.46	<0.01	**
Perfect	Explanation	183.46	115.39	-68.08	<0.01	***
Apology	Denial	115.84	69.18	-46.66	0.02	*
Apology	Promise	115.84	123.00	7.16	1.00	ns
Apology	Explanation	115.84	115.39	-0.45	1.00	ns
Denial	Promise	69.18	123.00	53.83	0.01	**
Denial	Explanation	69.18	115.39	46.21	0.02	*
Promise	Explanation	123.00	115.39	-7.61	1.00	ns

NOTE: Mean Rank Diff = Mean Rank 2 - Mean Rank 1

****(p <0.0001) **(p <0.001) *(p <0.05) n.s.(p 0.05)*

Table 2: Results of a post-hoc Dunn's test comparing trustworthiness by repair condition.

at repairing trust than the perfect performance condition suggesting that no strategy was capable of fully repairing trust. In addition to these findings, our analysis also indicates that denials were the least effective strategy in repairing perceptions of trustworthiness when compared to apologies, explanations, and promises. Furthermore, denials also produced lower trustworthiness than the no-repair condition. Together this may indicate that denials were not only ineffective but might be less effective than not deploying any repair strategy and remaining silent.

5.3.1. Trust Repair and Ability

To examine the effect of different trust repair strategies on ability, we again conducted a Kruskal-Wallis test followed by a post hoc Dunn's test of multiple comparisons. Results of this Kruskal-Wallis test showed that trust repair conditions had a statistically significant impact on perceptions of the robot's ability ($p = 9.681e - 12$, $\chi^2 = 60.48$, $\eta^2 = 0.32$) permitting for further examination via a post hoc investigation of these effects using a Dunn's test of multiple comparisons at the $p \leq 0.05$ level. This post-hoc test revealed significant differences between the perfect performance condition and apologies, denials, explanations, promises, and the no repair condition. Additionally, this test also revealed significant differences between denials and apologies, promises, and explanations. Table 3 summarizes these findings and a visual presentation of means and standard deviations is visible in Figure

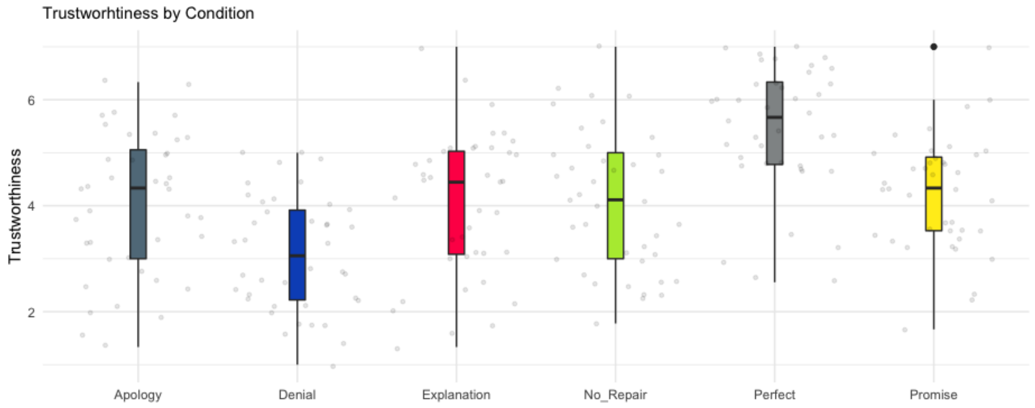

Figure 5: Box-plot showing trustworthiness by repair condition where denials can be seen as the least effective strategy and no repairs met or surpassed perfect performance.

6.

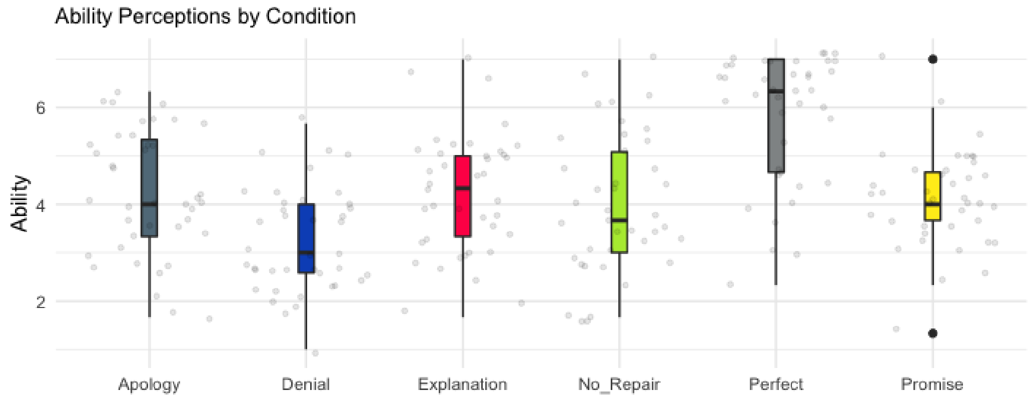

Figure 6: Box-plot showing ability by repair condition where no condition matched perfect performance and denials appeared to be the least effective repair strategy overall.

In general, these findings only partially support *H2c* that promises will be more effective at repairing ability than other repair strategies after repeated trust violations. This is the case as significant differences were observed between denials and promises but not between promises and apologies or explanations. Further examination of these results, however, provide additional insights. In particular, the perfect performance condition produced

Group 1	**Group 2**	**Mean Rank 1**	**Mean Rank 2**	**Mean Rank Diff**	**P-Val**	**Sig**
No Repair	Perfect	106.29	188.05	81.76	<0.01	***
No Repair	Apology	106.29	120.1	13.81	1.00	ns
No Repair	Denial	106.29	70.83	-35.46	0.15	ns
No Repair	Promise	106.29	116.44	10.15	1.00	ns
No Repair	Explanation	106.29	121.3	15.01	1.00	ns
Perfect	Apology	188.05	120.1	-67.95	<0.01	***
Perfect	Denial	188.05	70.83	-117.23	<0.01	***
Perfect	Promise	188.05	116.44	-71.61	<0.01	***
Perfect	Explanation	188.05	121.3	-66.75	<0.01	***
Apology	Denial	120.1	70.83	-49.27	0.01	*
Apology	Promise	120.1	116.44	-3.66	1.00	ns
Apology	Explanation	120.1	121.3	1.2	1.00	ns
Denial	Promise	70.83	116.44	45.61	0.03	*
Denial	Explanation	70.83	121.3	50.47	0.01	*
Promise	Explanation	116.44	121.3	4.86	1.00	ns

NOTE: Mean Rank Diff = Mean Rank 2 - Mean Rank 1

****(p <0.0001) **(p <0.001) *(p <0.05) n.s.(p 0.05)*

Table 3: Results of a post-hoc Dunn's test comparing ability by repair condition.

significantly higher perceptions of the robot's ability than apologies, denials, explanations, promises, and the no-repair condition. This seems to indicate that no strategy was fully effective in repairing perceptions of the robot's ability. Furthermore, denials also appeared to be significantly less effective at repairing perceptions of ability than apologies, explanations, and promises. This provides partial support for H2d that denials will likely be the least effective at repairing ability, integrity, and benevolence than the other repair strategies after repeated trust violations.

5.3.2. Trust Repair and Integrity

To examine the effect of different trust repair strategies on integrity, we once more conducted a Kruskal-Wallis test followed by a post hoc Dunn's test of multiple comparisons. Results of this Kruskal-Wallis test showed that trust repair conditions had a statistically significant impact on perceptions of the robot's integrity ($p = 6.589e - 13$, $\chi^2 = 66.11$, $\eta^2 = 0.29$). A post hoc investigation of these effects using a Dunn's test of multiple comparisons at the $p \leq 0.05$ level revealed that the perfect performance condition produced significantly higher perceptions of integrity than apologies, denials, explanations, promises, and the no-repair condition. Furthermore, denials appeared significantly different from explanations, promises, and the no repair condition. These results are summarized in Table 4 while a visual representation

Group 1	Group 2	Mean Rank 1	Mean Rank 2	Mean Rank Diff	P-Val	Sig
No Repair	Perfect	122.64	192.24	69.6	<0.01	***
No Repair	Apology	122.64	104.48	-18.16	1.00	ns
No Repair	Denial	122.64	70.71	-51.92	0.01	**
No Repair	Promise	122.64	113.85	-8.79	1.00	ns
No Repair	Explanation	122.64	119.09	-3.55	1.00	ns
Perfect	Apology	192.24	104.48	-87.76	<0.01	***
Perfect	Denial	192.24	70.71	-121.53	<0.01	***
Perfect	Promise	192.24	113.85	-78.39	<0.01	***
Perfect	Explanation	192.24	119.09	-73.15	<0.01	***
Apology	Denial	104.48	70.71	-33.76	0.21	ns
Apology	Promise	104.48	113.85	9.38	1.00	ns
Apology	Explanation	104.48	119.09	14.61	1.00	ns
Denial	Promise	70.71	113.85	43.14	0.04	*
Denial	Explanation	70.71	119.09	48.38	0.02	*
Promise	Explanation	113.85	119.09	5.24	1.00	ns

NOTE: Mean Rank Diff = Mean Rank 2 - Mean Rank 1

****(p <0.0001) **(p <0.001) *(p <0.05) n.s.(p 0.05)*

Table 4: Results of a post-hoc Dunn's test comparing integrity by repair condition.

of means and standard deviations is visible in Figure 7.

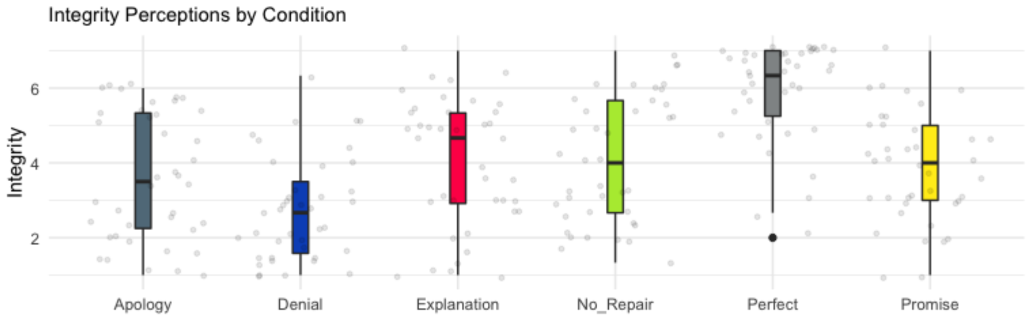

Figure 7: Box-plot showing integrity by repair condition where denials appear to be least effective, apologies, explanations, and promises appeared to function similarly, and no repairs are as effective as perfect performance.

In general, these findings only partially support H2c that explanations will be more effective at repairing integrity than other repair strategies after repeated trust violations. This is the case as even though a significant difference between promises and denials was present no significant difference between explanations and apologies and promises was observed. Further examination of these results also indicate that not only do denials appear to be the least effective strategy at repairing perceptions of integrity but that perfect performance was significantly different from all other conditions. The former of these findings provides support for H2d while the latter supports the notion that no strategy was fully effective in repairing perceptions of robot integrity.

5.3.3. Trust Repair and Benevolence

To examine the effect of different trust repair strategies on benevolence, we conducted a Kruskal-Wallis test followed by a post hoc Dunn's test of multiple comparisons. Results of this Kruskal-Wallis test showed that trust repair conditions had a statistically significant impact on perceptions of the robot's benevolence ($p = 0.003$, $\chi^2 = 18.34$, $\eta^2 = 0.008$). A post hoc investigation of these effects using a Dunn's test of multiple comparisons at the $p \leq 0.05$ level revealed that the perfect performance condition was significantly different than denials. In addition, denials were significantly different than promises. These results are summarized in Table 5 while a visual representation of means and standard deviations is visible in Figure 8.

Group 1	**Group 2**	**Mean Rank 1**	**Mean Rank 2**	**Mean Rank Diff**	**p.adj**	**p.adj.signif**
No_Repair	Perfect	120.2125	139.4875	19.28	1.00	ns
No_Repair	Apology	120.2125	125.9125	5.7	1.00	ns
No_Repair	Denial	120.2125	86.325	-33.89	0.32	ns
No_Repair	Promise	120.2125	142.65	22.44	1.00	ns
No_Repair	Explanation	120.2125	108.4125	-11.8	1.00	ns
Perfect	Apology	139.4875	125.9125	-13.58	1.00	ns
Perfect	Denial	139.4875	86.325	-53.16	0.01	**
Perfect	Promise	139.4875	142.65	3.16	1.00	ns
Perfect	Explanation	139.4875	108.4125	-31.08	0.45	ns
Apology	Denial	125.9125	86.325	-39.59	0.14	ns
Apology	Promise	125.9125	142.65	16.74	1.00	ns
Apology	Explanation	125.9125	108.4125	-17.5	1.00	ns
Denial	Promise	86.325	142.65	56.33	<0.01	**
Denial	Explanation	86.325	108.4125	22.09	1.00	ns
Promise	Explanation	142.65	108.4125	-34.24	0.32	ns

NOTE: Mean Rank Diff = Mean Rank 2 - Mean Rank 1

****(p <0.0001) **(p <0.001) *(p <0.05) n.s.(p 0.05)*

Table 5: Results of a post-hoc Dunn's test comparing benevolence by repair condition.

In general, these findings provide no support H2a that apologies will be more effective at repairing benevolence than other repair strategies after repeated trust violations. This is because apologies did not have a significantly different effect on benevolence than explanations, promises, or denials.

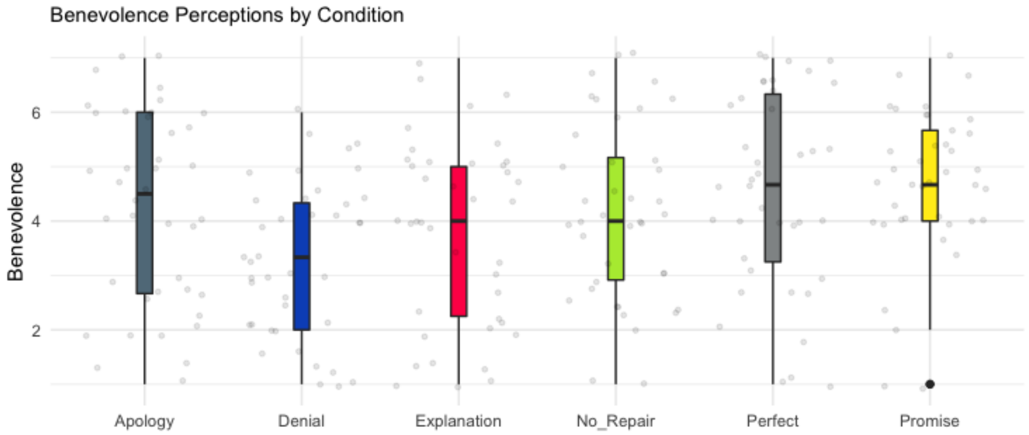

Figure 8: Box-plot showing benevolence by repair condition where denials differ from promises, apologies, explanations, and promises appear to act similarly and perfect performance did not appear to differ from any other condition examined.

In addition to this finding, our analysis also provides partial for H2d as denials appeared to be less effective than the perfect performance condition, promises, and apologies. Notably, however, denials did not significantly differ from the no repair condition and the explanation condition. This diverges from result associated with ability and integrity.

5.4. Summary of Findings

Results of our analysis generally supported *H2d* but only partially supported our other hypotheses. In particular, the impacts of apologies, explanations, and promises failed to differ significantly from one another for ability, integrity, or benevolence. Some support for these hypotheses was present for ability and integrity where significant differences were observed between denials and other repair conditions but, this failed to be the case for benevolence. Notably, several unexpected results emerged from this analysis as well. These results can be can be organized into two overarching findings. First, the repair strategies we examined were unable to repair trustworthiness to a pre-violation level leading us to partially reject *H1*. This was because trustworthiness was significantly lower across all repair strategies than it was in the perfect (i.e. no error) condition. Second, across ability, integrity, and benevolence, we observed similar trends for ability and integrity as for trustworthiness overall. In particular, ratings of ability and integrity never returned to a pre-violation level. This was the case regardless of the repair strategy used. For perceptions of benevolence, however, few differences emerged between conditions with denials and promises as well as denials and the perfect performance condition being the only exceptions.

6. Discussion

The goal of this paper was to examine whether certain trust repair strategies are effective in repairing trustworthiness and whether these repair strategies have different impacts on different sub-dimensions of trustworthiness including ability, benevolence, and integrity. Next, we detail the study's contributions to theory and the literature at large followed by a brief discussion of its limitations and opportunities for future work.

Overall, this study has several theoretical contributions to the HRI literature. First, our results indicate that after repeated trust violations apologies, explanations, and promises appear to be equally ineffective with regards to repairing trustworthiness, ability, and integrity. This result was unexpected given the theoretical differences in how trust repairs function as outlined in our theoretical research model. In particular, we expected to see promises outperform other repairs for perceptions of ability through the theory of forgetting and to see explanations outperform other repairs in repairing integrity via the theory of informing. Results, however indicated that this was not the case. One explanation for this might relate to possible interactions and the impact of repeated violations and repairs. Specifically, it could be that repair strategies function through the hypothesized theoretical frameworks at the first time point but that the unique impacts of these frameworks on ability and integrity wash out over repeated interactions. To examine this, future researchers might wish to more directly compare apologies, explanations, and promises at each time point to determine what impact repeat interactions might have on the efficacy of these repairs.

Second, we found that not only are apologies effective at repairing perceptions of benevolence but that promises and explanations were as well. From a theoretical perspective, this seems to support the idea that the theory of forgetting and informing can possibly repair trust via benevolence as much as the theory of forgiving. This also supports the notion that benevolence may be more repairable than ability and integrity. This conflicts, however, with findings from Alarcon et al. (2020), who found that integrity (process) and benevolence (purpose) were both capable of being fully restored [16].

An explanation for this difference, however, may relate to the fact that the Alarcon et al., (2020) did not consider the impact of multiple trust violations and repairs nor did they examine the same types of repair strategies used in this paper. Therefore these differences may relate to the unique impact of different repair strategies and/or the impact of repeated violations and repairs. That being said, it may be that a robot's response regardless of what it is (i.e. apologies, promises, and explanations) promotes benevolence. Regardless, benevolence has been largely under-examined, and our study's results indicate that more research on this topic is needed.

Third, our study showed that denials were consistently the least effective method of repairing overall trustworthiness as well as perceptions of ability, integrity, and benevolence. From a theoretical perspective, the results seem to indicate that if misinforming is to be effective as a trust repair strategy it will not be due to increases in ability, integrity, and benevolence at least not after repeated trust violations. There is evidence that denials can be effective as a repair strategy, although not necessarily as effective as other repair strategies [36]. One explanation for this might be the reliance of denials on successfully misattributing blame for a trust violation [27, 67, 78]. In particular, the degree of doubt individuals have in their own interpretation of events may have been lower after repeated violations. As a result, this strategy might be effective once but any positive impacts observed at that first-time point might be lost over repeated interactions due to the true source of the error becoming more obvious over repeated interactions. Future researchers might therefore wish to examine denial's impacts across interactions to verify whether such a trend is visible.

In addition to the above, our results also indicate that trustworthiness was never fully repaired to a pre-violation state. This finding is important because it sheds light on other findings within the HRI trust repair literature. In particular, studies finding that trust repairs had positive impacts on human trust often concluded that these repair strategies were ultimately effective [36–38, 41, 44, 46]. One difference between our study and these earlier studies, however, may relate to our use of repeated trust violations where the robot only temporarily changed its behavior and continued to make mistakes. To this end, more research is needed.

Comparing these results to those in human–human trust repair, our study empirically verified that trust is never fully repaired by apologies, denials, explanations, or promises. One set of literature on human–human trust repair, although not empirically examined, generally supports this perspective

[27, 79–81]. An alternative set of the literature, however, suggests that full human trust repair is possible but depends largely on the type of trust violation [43]. The results of our study add to this discourse and highlight the important role of repeated violations and repairs. In particular, our study's results indicate that after three violations and repairs trust cannot be fully restored thus supporting the adage "three strikes and you're out". In doing so it presents a possible limit that may exist regarding when trust can be fully restored. Future research, however, may wish to explore and compare the effectiveness of one v.s. two v.s. three violations and repairs. In doing so such research can expand upon our findings and establish if, when, and how many repeated violations and repairs can be used before they lose their ability to repair trust.

6.1. Limitations and Future Work

The study has several limitations that offer opportunities for future research. First, this study relied on a virtual environment. We chose to use a virtual environment as a result of the COVID-19 pandemic, which made recruiting and running in-person studies difficult. Virtual environments, however, are not uncommon in the HRI field and several studies have used them in the past to great effect [38, 39, 44, 70]. Prior HRI research has shown that humans in virtual environments behave quite similarly to humans interacting with robots in the real world [82]. Regardless, we might have gotten a stronger response with the use of physical robots. Future studies could be conducted to replicate our findings with physical robots.

Second, this study looked at the effects of each trust repair strategy separately and didn't compare possible permutations of these strategies. This allowed us to examine the selected trust repair strategies in greater detail but prevented a broader examination of the interactions that might have occurred when combining these strategies. In addition, by only examining one permutation of these strategies we also were able to focus the scope of the study but in doing so acknowledge that certain variants of these strategies were unexamined. Future researchers might therefore wish to take the repair strategies used in this paper and either combine them to examine which combinations are more or less effective or examine variants of these strategies. For example, future work may wish to compare the effects of apologies in combination with promises or compare how different types of explanations conveying more or less information may have impacted trust. Furthermore, future studies may also wish to examine other forms of trust repair such as nonverbal and long-term trust repair in HRI.

Third, this study used only one specific type of explanation. In particular, the explanation used by the robot conveyed that the robot understood why an error occurred but did not provide additional detail on what caused the robot to make the error. Although this is consistent with previous literature [21, 23, 42], recent work has highlighted that this particular repair strategy – referred to as acknowledgement – is but one of many types of possible explanation [58]. In particular, [58] highlights how other types of explanations may be more effective, even when they over multiple trust violations. As a result, future research should examine the effectiveness of other types of explanations in addition to simple acknowledgements and fully leverage the different types of explanations highlighted in [58] moving forward.

Fourth, this study focused exclusively on human–robot interaction. In doing so, it did not implement a human—human condition. As a result, direct comparisons with the human–human trust repair literature is difficult. Therefore future studies may wish to build upon our results by incorporating a human teammate into the study design. Fourth, this study focused on a specific type of task and scenario. In particular, we did not examine trust repair across different robot and task types. We also didn't permanently change the robot's behaviors in response to its failures. Future work is therefore needed to consider how various types of robots (anthropomorphic vs. machine-like), tasks, behaviors, and settings might impact the relationship between trust repair and trustworthiness in HRI especially given findings from other research highlighting how some of these may be influential to trust resilience [83] and trust repair [22].

Finally, this study's design measured trustworthiness prior to the first box and after the last box rather than after each box in the study. The logic behind this decision was to minimize the amount of disturbance to the subjects by presenting them with additional questionnaires during their interactions with the robot and task. While we feel that this is justified and increases the external validity of our design, future work may wish to measure trustworthiness at each time point. In doing so, such studies can examine how trustworthiness changes at a higher level of fidelity. Specifically, this future work could directly examine the impact of repeated violations and repairs at each time point rather than the cumulative effect overall. Furthermore, such work may also wish to combine this approach with the one used in this study to determine if these effects differ in their impact on ability, benevolence, and integrity.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Connor Esterwood: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project Administration

Lionel P. Robert: Conceptualization, Methodology, Validation, Resources, Data Curation, Writing - Review & Editing, Project Administration, Supervision, Funding Acquisition.

Declaration of Competing Interest

Referring to the manuscript, “Fool Me Three Times: Human–Robot Trust Repair & Trustworthiness Over Multiple Violations & Repairs” hereby we would like to declare that there is no conflict of interest among authors and any other party directly or indirectly.

Acknowledgements

The authors are extremely thankful to the Editor-in-Chief, the Associate Editor, and anonymous reviewers for their suggestions and feedback during the review process. We would also like to thank the Emerging Technology Group at the University of Michigan’s James and Anne Duderstadt Center. In particular we wish to thank Sara Eskandari and Stephanie O’Malley for their help in the development of our experimental platform.

Appendix A. Trustworthiness Questionnaire Items

The trustworthiness questionnaire used in this research was adapted from [71–73]. The individual items as well as the trustworthiness sub-dimensions associated are listed in table A1 below. To validate this measure we conducted an exploratory factor analysis with a varimax rotation via the `psych` package in R [84]. Results of this analysis indicated general support for most item's loadings with values of ≥ 0.7. One exception, however, was that of the second ability item ('Ability 2'). This lead us to conduct a followup confirmatory factor analysis using `lavaan` to assess how impactful this item may be on the overall measure [75]. Results of this followup analysis indicated that our measurement model was generally supported overall due to a comparative fit index ($CFI = 0.97$) and Tucker-Lewis index ($TLI = 0.95$) of ≥ 0.9. In addition, this measure of trustworthiness was found to be generally reliable as well with an overall reliability of $\alpha = 0.92$ and individual reliabilities of $\alpha = 0.76$ for ability, $\alpha = 0.95$ for integrity, and $\alpha = 0.92$ for benevolence. As a result, we opted to utilize this measure and consider it to be relatively robust.

Question	**Coding**	**Sub-Dimension**	**Source(s)**
The robot I worked with failed me.	Reverse Coded	Ability 1	[71]
The robot I worked with communicated clearly.		Ability 2	[72, 73]
The robot I worked with did not perform well.	Reverse Coded	Ability 3	[72, 73]
The robot I worked with was dependable.		Integrity 1	[71]
The robot I worked with had my confidence.		Integrity 2	[85]
The robot I worked with could be counted on to do its job.		Integrity 3	[85]
The robot I worked with cares about helping me do a good job.		Benevolence 1	New Items
The robot I worked with wants to help me do a good job.		Benevolence 2	New Items
The robot I worked with cares about doing a good job.		Benevolence 3	New Items

Table A1: Trustworthiness items used in this study.

Appendix B. Supplementary Tables

	Apology, N = 40	**Denial, N = 40**	**Explanation, N = 40**	**No_Repair, N = 40**	**Perfect, N = 40**	**Promise, N = 40**
Trustworthiness	4.03,4.33 (1.35)	3.10,3.06 (1.09)	4.02,4.44 (1.36)	4.08,4.11 (1.32)	5.45,5.67 (1.21)	4.22,4.33 (1.07)
Ability	4.22,4.00 (1.28)	3.25,3.00 (1.06)	4.25,4.33 (1.25)	3.98,3.67 (1.43)	5.83,6.33 (1.34)	4.13,4.00 (1.00)
Benevolence	4.25,4.50 (1.76)	3.31,3.33 (1.41)	3.80,4.00 (1.70)	4.17,4.00 (1.65)	4.59,4.67 (1.83)	4.67,4.67 (1.38)
Integrity	3.62,3.50 (1.72)	2.74,2.67 (1.37)	4.01,4.67 (1.75)	4.09,4.00 (1.67)	5.93,6.33 (1.33)	3.87,4.00 (1.45)

Mean, Median (SD)

Table B1: Table containing statistical summaries (Mean, Median, and Standard Deviations) for each measure by condition.

References

[1] V. Pitardi, B. Bartikowski, V.-S. Osburg, V. Yoganathan, Effects of gender congruity in human-robot service interactions: The moderating role of masculinity, International Journal of Information Management (2022) 102489.

[2] N. Savela, A. Oksanen, M. Pellert, D. Garcia, Emotional reactions to robot colleagues in a role-playing experiment, International Journal of Information Management 60 (2021) 102361.

[3] N. Sinha, P. Singh, M. Gupta, P. Singh, Robotics at workplace: An integrated twitter analytics–sem based approach for behavioral intention to accept, International Journal of Information Management 55 (2020) 102210.

[4] C. Esterwood, 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 1 (2021) 9–9.

[5] C. Esterwood, L. P. Robert, Personality in healthcare human robot interaction (H-HRI) a literature review and brief critique, in: Proceedings of the 8th International Conference on Human-Agent Interaction, 2020, pp. 87–95.

[6] C. Esterwood, K. Essenmacher, H. Yang, F. Zeng, 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.

[7] S. You, J.-H. Kim, S. Lee, V. Kamat, L. P. Robert Jr, Enhancing perceived safety in human–robot collaborative construction using immersive virtual environments, Automation in Construction 96 (2018) 161–170.

[8] C. Esterwood, L. P. Robert, Human robot team design, in: Proceedings of the 8th International Conference on Human-Agent Interaction, 2020, pp. 251–253.

[9] L. P. Robert Jr, A. R. Dennis, M. K. Ahuja, Differences are different: Examining the effects of communication media on the impacts of racial and gender diversity in decision-making teams, Information Systems Research 29 (2018) 525–545.

[10] A. Rossi, K. Dautenhahn, K. L. Koay, M. L. Walters, Human perceptions of the severity of domestic robot errors, in: A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, H. He (Eds.), Social Robotics, Springer International Publishing, Cham, 2017, pp. 647–656.

[11] S. Ye, G. Neville, M. Schrum, M. Gombolay, S. Chernova, 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), IEEE, 2019, pp. 1–7.

[12] H. Azevedo-Sa, X. J. Yang, L. P. Robert, D. M. Tilbury, A unified bi-directional model for natural and artificial trust in human–robot collaboration, IEEE Robotics and Automation Letters 6 (2021) 5913–5920.

[13] R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, Academy of management review 20 (1995) 709–734.

[14] L. Robert, S. You, Are you satisfied yet? shared leadership, trust and individual satisfaction in virtual teams, in: Proceedings of the iConference, 2013.

[15] S. S. Sebo, P. Krishnamurthi, 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), IEEE, 2019, pp. 57–65.

[16] G. M. Alarcon, A. M. Gibson, 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), IEEE, 2020, pp. 1–6.

[17] A. L. Baker, E. K. Phillips, D. Ullman, 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) 8 (2018) 1–30.

[18] M. Salem, G. Lakatos, F. Amirabdollahian, K. Dautenhahn, 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, 2015, pp. 1–8.

[19] D. B. Quinn, Exploring the Efficacy of Social Trust Repair in Human-Automation Interactions, Master's thesis, Clemson University, 2018.

[20] E. J. De Visser, R. Pak, T. H. Shaw, From ‘automation’to ‘autonomy’: the importance of trust repair in human–machine interaction, Ergonomics 61 (2018) 1409–1427.

[21] C. Esterwood, L. P. Robert, A literature review of trust repair in hri, in: Proceedings of 31th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN 2022), IEEE, 2022.

[22] C. Esterwood, 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.

[23] C. Esterwood, L. Robert, et al., Having the right attitude: How attitude impacts trust repair in human-robot interaction (2022).

[24] W. Kim, N. Kim, J. B. Lyons, C. S. Nam, Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach, Applied ergonomics 85 (2020) 103056.

[25] L. P. Robert, A. R. Denis, Y.-T. C. Hung, Individual swift trust and knowledge-based trust in face-to-face and virtual team members, Journal of Management Information Systems 26 (2009) 241–279.

[26] R. M. Kramer, R. J. Lewicki, Repairing and enhancing trust: Approaches to reducing organizational trust deficits, Academy of Management annals 4 (2010) 245–277.

[27] R. J. Lewicki, C. Brinsfield, Trust repair, Annual Review of Organizational Psychology and Organizational Behavior 4 (2017) 287–313.

[28] K. Sharma, F. D. Schoorman, G. A. Ballinger, How can it be made right again? a review of trust repair research, Journal of Management 0 (2022) 01492063221089897.

[29] M. H. Butler, S. K. Dahlin, S. T. Fife, “languaging” factors affecting clients’acceptance of forgiveness intervention in marital therapy, Journal of Marital and Family Therapy 28 (2002) 285–298.

[30] R. J. Bies, L. J. Barclay, T. M. Tripp, K. Aquino, A systems perspective on forgiveness in organizations, The Academy of Management Annals 10 (2016) 245–318.

[31] M. H. Butler, L. G. Hall, J. B. Yorgason, The paradoxical relation of the expression of offense to forgiving: A survey of therapists' conceptualizations, The American Journal of Family Therapy 41 (2013) 415–436.

[32] G. R. Maio, G. Thomas, F. D. Fincham, K. B. Carnelley, Unraveling the role of forgiveness in family relationships., Journal of personality and social psychology 94 (2008) 307.

[33] V. R. Waldron, Encyclopedia of human relationships, in: H. T. Reis, S. Sprecher (Eds.), Apologies, volume 3 of *1*, 1 ed., Sage Publishing Inc., Thousand Oaks, CA, 2009, pp. 98–100.

[34] R. J. Lewicki, B. Polin, R. B. Lount Jr, An exploration of the structure of effective apologies, Negotiation and Conflict Management Research 9 (2016) 177–196.

[35] M. E. McCullough, E. L. Worthington Jr, K. C. Rachal, Interpersonal forgiving in close relationships., Journal of personality and social psychology 73 (1997) 321.

[36] S. C. Kohn, A. Momen, E. Wiese, Y.-C. Lee, 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 63 (2019) 222–226.

[37] M. Natarajan, 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.

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

[39] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, 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, pp. 203–210.

[40] J. Xu, A. Howard, Evaluating the impact of emotional apology on human-robot trust, in: 2022 31st IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), IEEE, 2022.

[41] D. Cameron, S. de Saille, E. C. Collins, J. M. Aitken, H. Cheung, A. Chua, E. J. Loh, J. Law, The effect of social-cognitive recovery strategies on likability, capability and trust in social robots, Computers in Human Behavior 114 (2021) 106561–106561.

[42] B. Bozic, V. G. Kuppelwieser, Customer trust recovery: An alternative explanation, Journal of Retailing and Consumer Services 49 (2019) 208–218.

[43] M. E. Schweitzer, J. C. Hershey, E. T. Bradlow, Promises and lies: Restoring violated trust, Organizational behavior and human decision processes 101 (2006) 1–19.

[44] P. Robinette, A. M. Howard, A. R. Wagner, Timing is key for robot trust repair, in: International conference on social robotics, Springer, 2015, pp. 574–583.

[45] S. Reig, E. J. Carter, T. Fong, J. Forlizzi, A. Steinfeld, Flailing, hailing, prevailing, ACM/IEEE, 2021, pp. 58–167.

[46] N. Wang, D. V. Pynadath, S. G. Hill, Trust calibration within a human-robot team: Comparing automatically generated explanations, in: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2016, pp. 109–116.

[47] Y. Feng, H. Tan, Comfort or promise? investigating the effect of trust repair strategies of intelligent vehicle system on trust and intention to use from a perspective of social cognition, in: International Conference on Human-Computer Interaction, Springer, 2022, pp. 154–166.

[48] K. James, Q. Paul, J. Jennifer, S. David, R. Robert, Misinformation and the currency of democratic citizenship, Journal of Politics 62 (2000) 790–816.

[49] F. I. Dretske, Knowledge and the Flow of Information, MIT Press, 1981.

[50] S. O. Søe, A unified account of information, misinformation, and disinformation, Synthese 198 (2021) 5929–5949.

[51] L. Vornik, S. Sharman, M. Garry, The power of the spoken word: Sociolinguistic cues influence the misinformation effect, Memory 11 (2003) 101–109.

[52] T. Kähkönen, K. Blomqvist, N. Gillespie, M. Vanhala, Employee trust repair: A systematic review of 20 years of empirical research and future research directions, Journal of Business Research 130 (2021) 98–109.

[53] N. Du, J. Haspiel, Q. Zhang, D. Tilbury, A. K. Pradhan, X. J. Yang, 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 104 (2019) 428–442.

[54] E. C. Tomlinson, A. M. Carnes, When promises are broken in a recruitment context: The role of dissonance attributions and constraints in repairing behavioural integrity, Journal of Occupational and Organizational Psychology 88 (2015) 415–435.

[55] A. Ezenyilimba, M. Wong, A. Hehr, M. Demir, A. Wolff, E. Chiou, N. Cooke, Impact of transparency and explanations on trust and situation awareness in human–robot teams, Journal of Cognitive Engineering and Decision Making (2022) 15553434221136358.

[56] B. R. Rawlins, Measuring the relationship between organizational transparency and employee trust (2008).

[57] W. P. Bottom, K. Gibson, S. E. Daniels, J. K. Murnighan, When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation, Organization Science 13 (2002) 497–513.

[58] J. B. Lyons, I. aldin Hamdan, T. Q. Vo, Explanations and trust: What happens to trust when a robot partner does something unexpected?, Computers in Human Behavior 138 (2023) 107473.

[59] E. S. Kox, J. H. Kerstholt, T. F. Hueting, P. W. De Vries, Trust repair in human-agent teams: the effectiveness of explanations and expressing regret, Autonomous Agents and Multi-Agent Systems 35 (2021).

[60] K. Hald, K. Weitz, E. André, M. Rehm, “an error occurred!”-trust repair with virtual robot using levels of mistake explanation, in: Proceedings of the 9th International Conference on Human-Agent Interaction, 2021, pp. 218–226.

[61] A. Thomsen, NoRobot’s perfect: trust repair in the face of agent error. How do individual factors influence trust development in human-agent teams?, B.S. thesis, University of Twente, 2022.

[62] S. Lewandowsky, U. K. Ecker, J. Cook, Beyond misinformation: Understanding and coping with the “post-truth” era, Journal of applied research in memory and cognition 6 (2017) 353–369.

[63] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, 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.

[64] E. C. Tomlinson, R. C. Mayer, The role of causal attribution dimensions in trust repair, Academy of Management Review 34 (2009) 85–104.

[65] E. C. Tomlinson, B. R. Dineen, R. J. Lewicki, The road to reconciliation: Antecedents of victim willingness to reconcile following a broken promise, Journal of management 30 (2004) 165–187.

[66] X. Zhang, “Sorry, It Was My Fault”: Repairing Trust in Human-Robot Interactions, Master’s thesis, University of Oklahoma, 2021.

[67] P. H. Kim, D. L. Ferrin, C. D. Cooper, 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 89 (2004) 104.

[68] Amazon.com, New technologies to improve amazon employee safety, 2021. URL: `https://tinyurl.com/2nknch48`.

[69] J. R. Rein, A. J. Masalonis, J. Messina, 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, volume 57, SAGE Publications Sage CA: Los Angeles, CA, 2013, pp. 280–284.

[70] M. Nayyar, 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.

[71] D. H. Mcknight, M. Carter, J. B. Thatcher, P. F. Clay, Trust in a specific technology: An investigation of its components and measures, ACM Transactions on management information systems (TMIS) 2 (2011) 1–25.

[72] J. Bernotat, F. Eyssel, J. Sachse, Shape it–the influence of robot body shape on gender perception in robots, in: International Conference on Social Robotics, Springer, 2017, pp. 75–84.

[73] J. Bernotat, F. Eyssel, J. Sachse, The (fe)male robot: how robot body shape impacts first impressions and trust towards robots, International Journal of Social Robotics (2019) 1–13.

[74] C. Fornell, D. F. Larcker, Structural equation models with unobservable variables and measurement error: Algebra and statistics, 1981.

[75] Y. Rosseel, lavaan: An r package for structural equation modeling, Journal of statistical software 48 (2012) 1–36.

[76] L.-t. Hu, P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural equation modeling: a multidisciplinary journal 6 (1999) 1–55.

[77] G. J. Medsker, L. J. Williams, P. J. Holahan, A review of current practices for evaluating causal models in organizational behavior and human resources management research, Journal of management 20 (1994) 439–464.

[78] P. H. Kim, K. T. Dirks, C. D. Cooper, 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 99 (2006) 49–65.

[79] R. J. Lewicki, B. B. Bunker, et al., Developing and maintaining trust in work relationships, Trust in organizations: Frontiers of theory and research 114 (1996) 139.

[80] R. J. Lewicki, C. Wiethoff, Trust, trust development, and trust repair, The handbook of conflict resolution: Theory and practice 1 (2000) 86–107.

[81] T. Haesevoets, Can broken trust be repaired?: a social and neuropsychological perspective, Ph.D. thesis, Ghent University, 2017.

[82] A. Heydarian, J. P. Carneiro, D. Gerber, B. Becerik-Gerber, T. Hayes, W. Wood, Immersive virtual environments versus physical built environments: A benchmarking study for building design and user-built environment explorations, Automation in Construction 54 (2015) 116–126.

[83] E. J. De Visser, S. S. Monfort, R. McKendrick, M. A. Smith, P. E. McKnight, F. Krueger, R. Parasuraman, Almost human: Anthropomorphism increases trust resilience in cognitive agents., Journal of Experimental Psychology: Applied 22 (2016) 331.

[84] W. Revelle, psych: procedures for psychological, psychometric, and personality research (r package version 1.9. 12), Evanston, IL: Northwestern University (2019).

[85] S. K. Ranganathan, V. Madupu, S. Sen, J. R. Brooks, Affective and cognitive antecedents of customer loyalty towards e-mail service providers, Journal of Services Marketing (2013).