Robots That Use Physical Repair Strategies After Repeated Errors to Mitigate Trust Decline in Human-Robot Interaction: A Repeated Measures Experiment

Sophie Lane[1], Connor Esterwood[2], Dana Kulić[1] and Nicole Robinson[1]

Abstract

Robots are inherently imperfect, and collaborating with an error-prone robotic teammate can deteriorate perceptions of trust and the willingness of users to continue working with the robot. Evidence-based trust repair strategies can be implemented into a robot's design to mitigate the decline of trust in human-robot relationships following errors. It is not yet clear what trust repair strategies are most effective. To address this shortcoming, this study investigates two novel trust repair strategies: offered and automatic physical repair. A between-subjects repeated measures study was performed to determine the extent to which each type of physical trust repair was successful in restoring participants' perceptions of trust. The results indicated that, where the no-repair condition experienced a significant decrease in trust score, only the automatic repair was consistently successful in bypassing the trust decline. Detailed analysis showed that participants from the offered repair condition did not view the robot as providing the appropriate information, meaning that the offer itself may have confused them. Participants' response rate to the Multi-Dimensional Measure of Trust also revealed that users were less willing to associate moral terms with robotic teammates, though this hesitancy may reduce over time. These results contribute to research on human-robot trust repair by uncovering that physical repair is effective when it is automatic, but not when it is offered. This finding will help to further elucidate what repair strategies work to mitigate trust decline and thus help inform robot design.

I. Introduction

From AI assistants to robots [1], [2], autonomous technologies are expected to become increasingly widespread [3], [4], [5]; they are currently being used in medical settings [6], aged care [7], agriculture [8], and manufacturing [1], [2]. Given the growing prevalence of robots, it is crucial to explore ways in which robots can be designed to promote successful and beneficial interactions with humans.

The degree to which people trust robots impacts their willingness to work with them and is key to establishing collaborative human-robot teams [9]. However, trust is not a constant; it increases when robots perform well and can rapidly diminish when robots inevitably make mistakes [9]. It is necessary to investigate how trust violations and trust repairs influence perceptions of trust and people's willingness to continue working with an imperfect robotic teammate so that mitigating strategies may be developed.

Current trust repair research has centred on the effects that four distinct verbal repair strategies (apologies, explanations, promises, and denial) each have on the three sub-dimensions of trust (competence, integrity, and benevolence) [10], [11], [12]. Existing research has also identified various antecedents of trust in human-robot relationships, such as transparency and predictability [13], [14], [15], responsibility [14], [16], [17], and emotion [13], [15], [18], [19]. The current literature, however, presents mixed results on how verbal trust repairs affects these sub-dimensions of trust. Thus, their effectiveness in reducing trust decline remains unclear.

Physical trust repair, in which the robot corrects the error that initiates a trust violation, is an alternative strategy that may mitigate the decline in trust. It differs from verbal strategies by not only acknowledging that a mistake has been made, but evidencing its ability to correct it. It may function similarly to promises by encouraging people to look beyond past errors [10]. However, to further the effect of promises, a physical repair immediately shows a concrete ability to rectify erroneous behaviour. Therefore, physical repair may be a more beneficial trust repair strategy than its verbal counterparts.

This study aims to investigate the extent to which two types of physical trust repair can mitigate trust decline following a robot's trust violation. In particular, we investigated the differences between an automatic physical repair and an offered physical repair in restoring trust perceptions. This contributes to the literature by examining the performance of two novel trust repair strategies that may be more effective than the verbal strategies previously studied.

II. Literature Review

A. *Trust and Trustworthiness*

Trust can be defined as ”the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability” [20, p. 51]. Trust is vital to the success of human-robot relationships and users' willingness to continue working with robots despite unexpected behaviours.

[1] Sophie Lane, Dana Kulić and Nicole Robinson are with the Faculty of Engineering, Monash University, Melbourne, VIC 3800, Australia sophie.lane@monash.edu, dana.kulic@monash.edu, nicole.robinson@monash.edu

[2] Connor Esterwood is with the School of Information, University of Michigan, Ann Arbor, MI 48109-1285, USA cte@umich.edu

D. Kulić is supported by the ARC Future Fellowship (FT200100761).

A related concept to trust is trustworthiness, which ultimately precedes trust [20]. Trustworthiness can be divided into three sub-dimensions: performance, process, and purpose (also commonly referred to as ability, integrity, and benevolence) [20]. In summary, performance describes what the agent does, process describes how the agent does it, and purpose describes why [20]. By deconstructing trust into these sub-dimensions, trust research has been better able to measure the influence of trust violation and trust repair on the overall trust of intelligent agents. This framework has become widely adopted and used almost universally in trust and trust repair research in the HRI field [10], [16], [17], [21], [22], [23], [24].

Several antecedents of trust have been identified, such as responsibility, emotion, and transparency. Users are reportedly more eager to blame robots that display higher levels of autonomy [14], [25], and that the attribution of responsibility changes to the robot following a trust violation [14].

Users tend to like robots with empathy better [13], which in turn strengthens trust [13], [18], [24]. Expression of regret may be crucial to repairing trust [18], [19], whilst participants also prefer emotional expressiveness despite decreased efficiency [15].

Transparency and predictability are notable predictors of trust that can mitigate frustration and dissatisfaction [13], [14], [15], even when users feel a lack of control caused by the robot's high level of autonomy. Facial expressions and communicating intent have been found to help improve perceptions of transparency [15], [26].

In order to explore human-robot trust, studies tend to use one of three forms of trust exercise: a collaborative task [10], [11], [12], [14], [15], [16], [19], [23], [27], a cooperative game [17], [18], [21], or a non-cooperative game [22], [24]. Understanding the sub-dimensions of trust, as well as its antecedents, is useful in informing how trust is formed and maintained throughout human-robot relationships, and thereby helps improve robot trust repair strategies.

B. Trust Repair in HRI

Robots employ trust repair strategies to mitigate trust decline following erroneous or unexpected behaviours. Research on a variety of strategies is necessary to understand how they each perform and to thereby inform interactive robot design.

Research into trust repair includes a trust violation committed at some point during the interaction. This could be a simple error [10], [11], [15], [16], [18], [23], [27], a non-cooperative action [22], or even a betrayal of a promise [24]. Trust violations are categorised into three types drawn from the three sub-dimensions: competence-based, integrity-based or benevolence-based. Although integrity-based and benevolence-based violations are under-researched [9], competence-based violations are the most likely candidate for trust repair research as performance-related characteristics have been found to be the main predictor of trust [24], [28], [29].

Many studies of trust repair explore some or a combination of four verbal trust repair strategies: apologies, explanations, promises, and denial. Each of these repairs is theorised to influence trust through different sub-dimensions of trustworthiness [10]. In particular, [10] propose the theories of forgiving, forgetting, informing, and misinforming, and they tie these theories to apologies, promises, explanations, and denial, respectively. Empirical support for these mechanisms and indeed for the efficacy of these repairs in general appears mixed [9].

For example, [10] noted that apologies, explanations, and promises improved perceptions of benevolence but no other sub-dimension, whilst [11], [16] and [17] found that certain explanations also improved perceptions of integrity. [10] found denials did not improve trust across any sub-dimension, though [24] found that they could be effective after integrity-based violations. [18] explored apologies and explanations in combination, finding that apologies can be effective on their own, but explanations only serve to support apologies and are not effective in their own right. [15] and [19] found that apologies and explanations were effective in restoring trust, with [24] confirming that apologies work best after competence-based violations. [21] and [27] found that promises are effective trust repair strategies, with [21] elaborating that realistic promises work better in the long run than optimistic ones.

[12] reported that individual differences and prior attitudes had a significant effect on the success of trust repair strategies, which could explain the mixed effects of these verbal strategies. Studies have combined trust repair strategies [9], or noted the variability of explanations and apologies [15], [17], [19], making such results incomparable to those of other studies [30].

As a result, it remains unclear which trust repair strategies are effective for which type of trust violation. It could be beneficial to branch out from widely studied verbal repair strategies to investigate other ways in which trust decline can be mitigated. In particular, few studies have considered how physical trust repair can affect human-robot trust. Additionally, to date, no studies have examined how the administration of a physical repair, offered or automatic, impacts the effectiveness of this novel repair strategy.

C. Physical Trust Repair in HRI

1) What is Physical Trust Repair?:

A physical repair is when the robot corrects its own error by displaying the expected behaviour after the violation. A physical repair changes depending on the type of trust violation. Where the error is behavioural, simply returning to expected behaviour could be seen as a physical repair, as done by [22]. Where a trust violation consists of a more physical error, such as in this study, a physical repair can consist of physically rectifying the issue the robot has caused.

Physical trust repair is closely related to the framework of forgetting and may function similarly to promises. Physical trust repair goes beyond a promise, however, by immediately acting on the intention to improve behaviour. Promises require users to believe in the robot's ability to return to expected behaviour, whereas a physical repair promptly evidences this intention *and* capability. Therefore, a physical trust repair strategy may be more effective than a verbal promise.

2) Why Can Physical Trust Repair Work?: Physical repair can be implemented in a variety of ways. By automatically returning to expected behaviour, as in [22], perceptions of capability and reliability may increase. However, our study also investigated another type of physical repair: one that is *offered* to participants. Through an offer the user has direct control of the situation, which can in turn improve perceptions of transparency, both of which have been linked to higher levels of trust [13], [14], [15].

[22] is the only study that implements a type of physical trust repair strategy. After committing a behavioural error in the fourth round of a non-cooperative game, the trust repair consisted of simply returning to expected behaviour in the fifth round. They found that perceptions of ability were repaired, although integrity and benevolence were not [22]. This could be explained by the fact that trust violation in this experimental design could be seen as a malicious action [22], thus affecting long-term perceptions of integrity and benevolence. More research is required to gather whether physical repair could be successful in restoring trust when the robot commits a seemingly honest mistake.

Therefore, this study will investigate two different types of physical trust repair (offered and automatic), where the trust violation consists of an "honest" mistake made by the robot. These trust repair conditions will be compared to a no-repair condition in which the robot asks the participant to fix the error.

III. Method

We designed a collaborative human-robot interaction task, where the human and the robot collaborated in six block-building tasks. During some of the tasks, the robot would make an error, creating a trust violation. The robot would then initiate a repair strategy. We investigated how trust was affected, using a mixed methods approach. Quantitative data was used as an objective measure of each participant's attitude and perception of trust towards the robot. The qualitative methods were designed to gain insight into the reasoning behind the survey responses of the participants.

A. Hypotheses

The following hypotheses were posited:

H1. Physical repair (both offered and automatic) will outperform the no-repair condition in participants' perceptions of trust towards the robot.

As found by [22] and much of the current literature, it is expected that the two repair conditions will each restore some level of trust across at least one of the sub-dimensions.

H2. An automatic physical trust repair will outperform an offered physical trust repair in participants' perceptions of trust towards the robot.

Whilst an offered repair is expected to increase perceptions of transparency, performance-based qualities are more closely associated with trust [28], [24], [29]. Therefore, it is theorised that an automatic repair will perform better than an offered repair due to its efficiency and immediate impression of capability.

B. Procedure

Participants provided basic demographic information before receiving a short safety induction and demonstration of the robot. Participants were then asked to complete the initial Trust Perception Scale (TPS) and Multi-Dimensional Measure of Trust (MDMT) [31], [32], as well as a baseline mid-study survey.

Participants were informed that there were six collaborative building tasks to complete, during which they and the robot would take turns placing blocks. Audio played from the robot to indicate who should go first, when the robot was about to place a block, and when the robot expected the participant to place a block.

For each building task, the participants received an image of the correct colour order of the blocks. They were told that each build task was scored out of one hundred, and that they would lose ten points for each mistake made during the task. They were made aware that they could win a small, 3D printed figurine if they made no more than one mistake across all six tasks, motivating them to succeed.

At one point in both the third and fifth tasks, the robot placed the incorrect coloured block. This was evident to the participants as it did not match the given image. The robot behaviour participants experienced following the error was aligned to their assigned condition group. All participants would fail to score enough points to earn the prize; however, they would receive it at the end of the experimental session nonetheless.

After completing each task, they were asked to complete another mid-study survey. Following the sixth task, participants were asked to complete the mid-study survey once more, as well as a final TPS and MDMT [31], [32]. Each participant

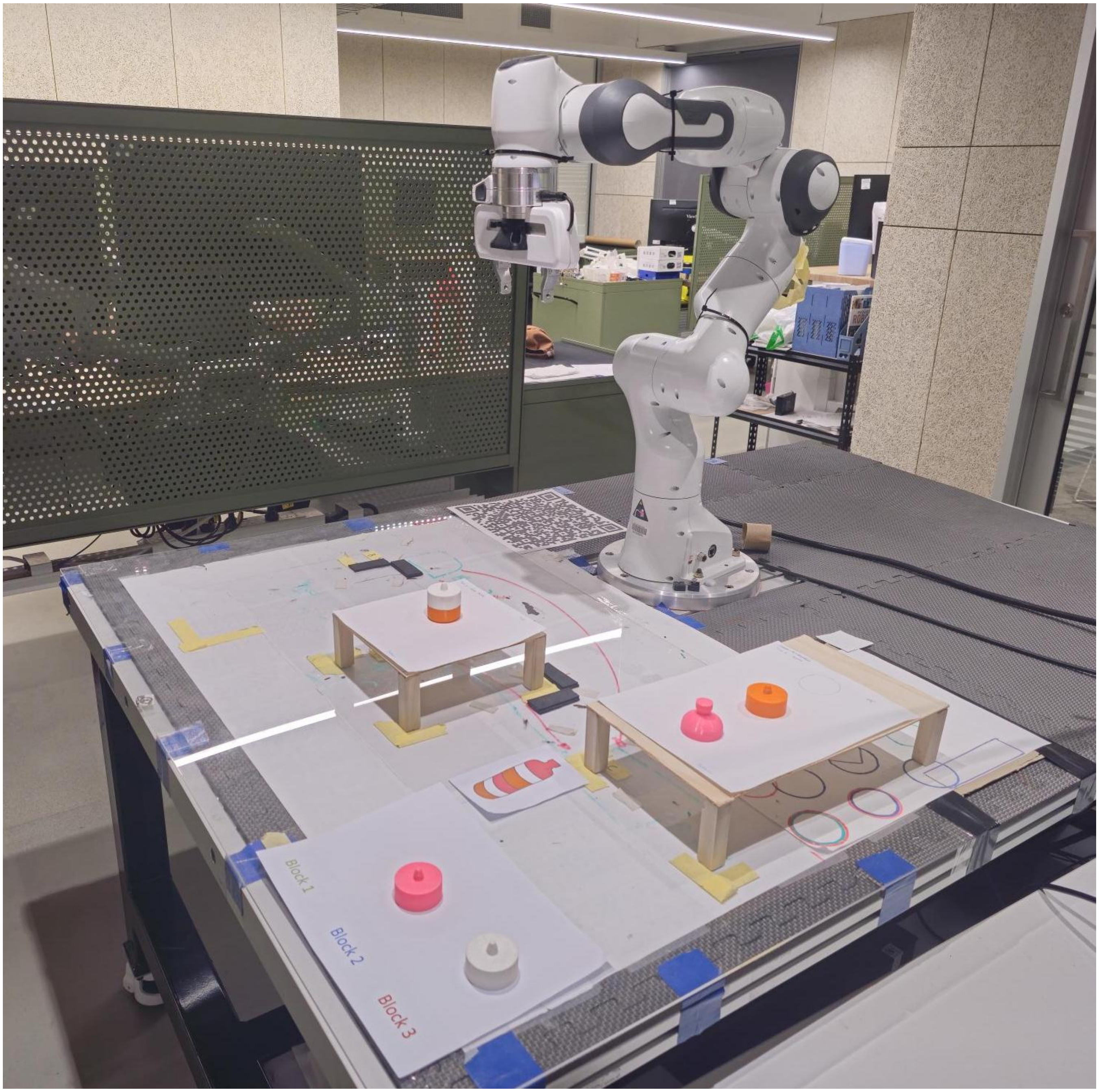

Fig. 1. Experimental set up halfway through the completion of the first task

was then interviewed using a semi-structured format, designed to gain more insight into their perception of the robot. Fig. 1 shows the experimental set up of the first task; the robot, six building blocks and final build image are all visible.

C. Experimental Design

This study was a two-way repeated measures design to investigate whether physical repair had an influence on trust recovery. In particular, participants were assigned to the no-repair, offered physical repair, or automatic physical repair conditions.

In the no-repair condition, the robot apologised for making an error and asked the participant to fix the mistake. This condition was used as a baseline indicator of participants' trust levels with the robot after seeing its erroneous behaviour without physical repair.

In the offered physical repair condition, the robot apologised and offered to fix the mistake itself. If the participant responded yes, the robot physically corrected the error. If the participant responded no, the robot asked the participant to fix the mistake.

In the automatic repair condition, the robot apologised and immediately informed the participant that it would fix the mistake, automatically rectifying the error.

D. Analytical Methods

1) Measures: During the study, three surveys were used to measure trust. Two of these surveys served as pre-post measures, while the other was administered throughout the experimental session to gain a more detailed look at the trust levels of participants over time.

Pre-Post Surveys

The TPS and MDMT were performed before and after completing all building tasks, serving as pre-post trust measures [31], [32]. TPS is a reduced 14-item pre-post scale designed to measure trust specifically in human-robot interactions [31]. MDMT is a 16-item survey that assesses whether participants believe the robot is reliable, capable, ethical, and / or sincere [32]. These four dimensions are categorised into two broader types of trust: Capacity Trust (reliable and capable), and Moral Trust (ethical and sincere) [32].

Participants were given a demonstration of the robot before completing the pre-interaction surveys, as required [31], [32]. A comparison of the responses from the pre-interaction and post-interaction survey should provide evidence on whether the participants' experience of the robot's erroneous behaviour has influenced their perception of trust. Furthermore, the comparison between conditions determines whether a physical trust repair reduces participants' decline in trust following robot errors.

Mid-Study Survey

The mid-study survey was completed at the beginning of the experimental session, as well as after each collaborative building task. This survey required participants to rate their attitudes toward the robot on a scale of one to ten, covering perceptions of reliability, capability, trust, their willingness to work with the robot, and their perception of the robot as a teammate. Completing this survey after each building task should provide a more granular look at how their levels of trust are changing over time. Comparing responses across conditions will give insight into whether offered or automatic repair might help mitigate the extent to which trust declines.

2) Participants: A total of 47 participants were recruited from the Monash University Clayton Campus, aged between 19 and 25. Data from 13 of these participants had to be excluded due to robot problems. Of the remaining 34 participants, 10 were female and 24 were male. Because of the novelty of this research, there were no prior studies that could be used to inform sample sizes.

Participants were 18 years or older to ensure they could provide informed consent and were required to have sufficient English skills to understand safety and task instructions. The participant group comprised a mixture of robot experience levels: 14 (6 female) had no experience, 14 (2 female) had a bit of experience, and 6 (2 female) had a lot of experience. Consent was registered via the intake form before subjects could schedule participation.

3) Statistical Analysis: A two-way between subjects ANOVA was used to analyse the change in scores from the pre-post surveys. A post-hoc analysis in the form of two pairwise t-tests with bonferroni adjustment were completed: a comparison between condition, within time and another comparison between time, within condition. These pairwise comparisons should reveal which condition(s) experienced a difference in trust levels over time, and if there were any significant differences in trust levels between conditions at the completion of the experiment.

To analyse results from the mid-study survey, a mixed linear model was used to investigate potential interaction effects between repair condition and time. A mixed linear model was used here instead of a two-way ANOVA as the results provide more detailed information about the time points at which significant interaction effects may be observed. During analysis, a jump in trust was observed between time 0 and time 1 for all five dimensions, perhaps because participants had experienced working with the robot at time 1 but not at time 0. Therefore, it was concluded that the no-repair condition at time 1 would serve as a more meaningful intercept, and the data from time 0 for all conditions were excluded from this analysis.

Definition of Outliers

Outliers were identified in the data through the use of the identify_outliers function, which defines an outlier as values above Q3 + 1.5*IQR or below Q1 - 1.5*IQR.

Treatment of Missing Values

For the MDMT survey, participants had the option to score any item as 'Does Not Fit', which should be treated as missing values [32]. Our attempt to seek clarification from the authors on how to deal with missing values was unsuccessful. Therefore, the missing values were treated in a typical manner. Each of the four scores is an average of four values, any of which could be designated 'Does Not Fit' by participants. If, for example, a participant rates the four reliable descriptors as 1, 2, 3, and 'Does Not Fit', the average will be calculated as (1+2+3)/3.

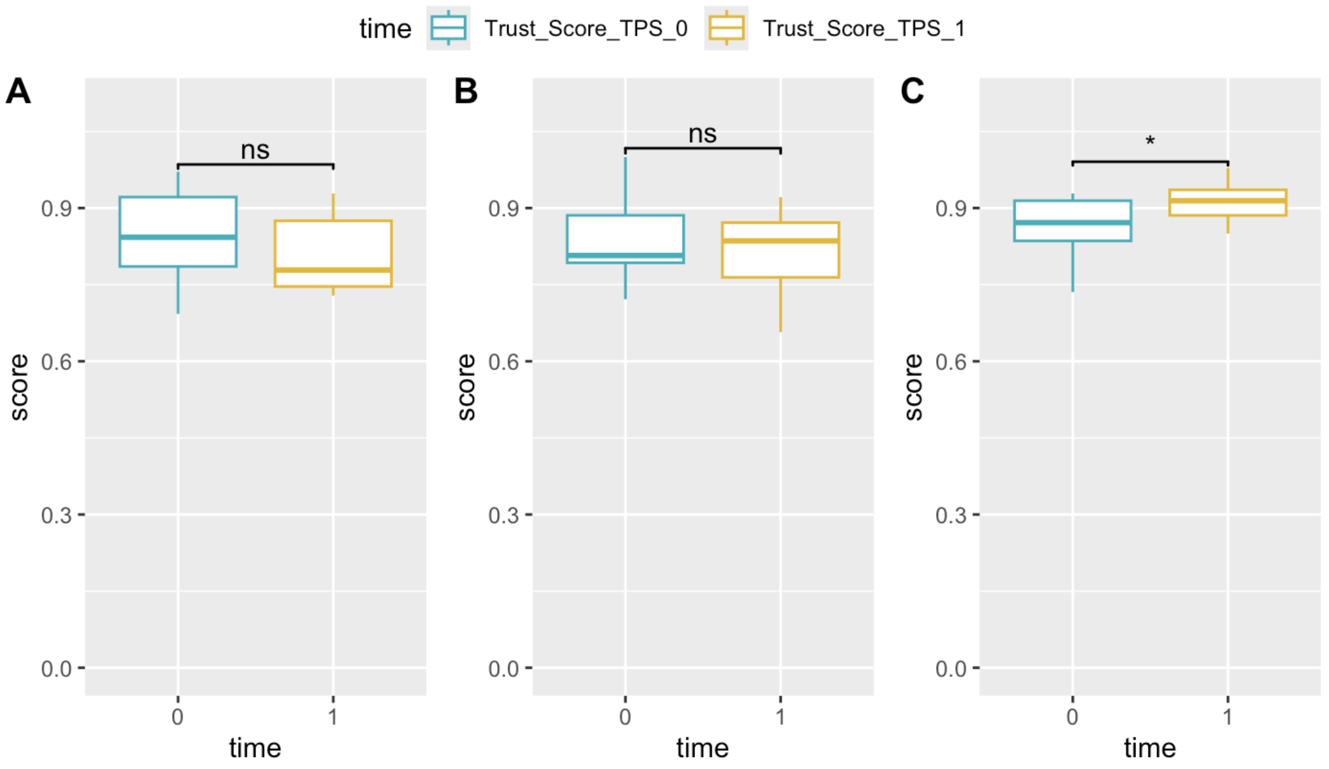

Fig. 2. TPS score over time per repair condition, where A is no repair, B is offered repair and C is automatic repair. Time 0 is before the building tasks, and time 1 is following all 6 building tasks. Significance shown according to a paired t-test: ns denotes p > 0.05, * denotes 0.01 < p < 0.05

IV. Results

A. Trust Perception Scale

A two-way ANOVA was performed to analyse the effect of trust repair condition and time on the participants' trust score from the 14-item TPS [31]. The two-way ANOVA found no significant interaction effect between repair condition and time (*F(2, 28) = 2.201, p = 0.129*). However, simple main effects analysis showed that the repair condition did have a statistically significant effect on participants' TPS score (*p = 0.047*). Time did not have a statistically significant effect on participants' TPS score (*p = 0.889*).

A post-hoc analysis in the form of paired samples t-tests was conducted to determine the effect of each repair condition on participants' TPS score.

A pairwise t-test with Bonferroni correction grouped by repair condition that analysed change in TPS score over time revealed that participants within the automatic repair condition displayed a significant increase in TPS score from before the experiment (*M = 0.859, SD = 0.065*) to after the experiment (*M = 0.914, SD = 0.044; t(9) = 0.879, p = 0.0492*). For the no repair condition, there was an insignificant change in TPS score from before the experiment (*M = 0.849, SD = 0.09*), to after the experiment (*M = 0.813, SD = 0.076; t(11) = -0.135, p = 0.326*). In the offered repair condition, there was an insignificant change in TPS score from before the experiment (*M = 0.842, SD = 0.082*), to after the experiment (*M = 0.814, SD = 0.085; t(11) = -0.117, p = 0.44*).

Fig. 2 shows the results from this pairwise t-test, analysing change in TPS score across time, within condition.

A pairwise t-test with Bonferroni correction grouped by time revealed that there was an insignificant difference in the final TPS score between no repair and offered repair (*p = 1*). The results indicated that the final TPS score for automatic repair was significantly higher than that of no repair (*p = 0.012*) and offered repair (*p = 0.013*).

A closer analysis of the individual items from the TPS revealed that perceptions of the following six dimensions were most affected: Reliable, Consistent, Robot Function, Robot Errors, Information, and Performs Exactly.

A two-way ANOVA revealed that there was an interaction effect between repair condition and time on participants' scores for the dimensions Consistent (*F(2) = 8.923, p = 0.001*), Robot Function (*F(2) = 3.944, p = 0.031*), Robot Errors (*F(2) = 4.323, p = 0.024*), and Information (*F(2) = 3.752, p = 0.036*).

Simple main effects analysis revealed that repair condition had a significant effect on score for Reliable (*p = 0.009*) and Performs Exactly (*p = 0.010*).

Simple main effects analysis revealed that time had a significant effect on score for the dimensions: Reliable (*p = 0.002*), Consistent (*p = 0.003*), Robot Errors (*p = 0.002*), Information (*p = 0.001*), and Performs Exactly (*p = 0.005*).

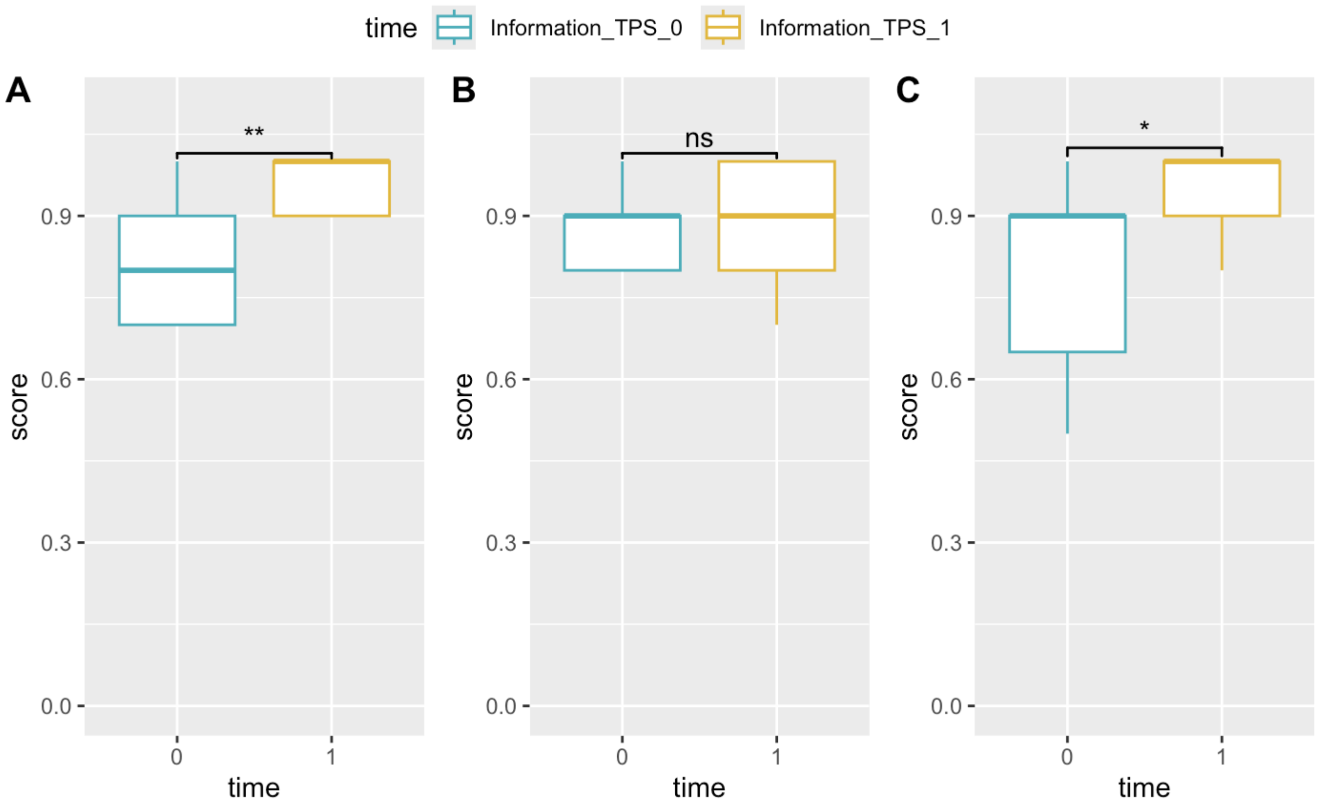

Fig. 3. TPS information score over time per repair condition, where A is no repair, B is offered repair and C is automatic repair. Time 0 is before the building tasks, and time 1 is following all 6 building tasks. Significance shown according to paired t-test: ns denotes p > 0.05, * denotes 0.01 < p < 0.05, ** denotes 0.001 < p < 0.01

A post-hoc analysis in the form of pairwise comparisons with Bonferroni adjustment was completed for these six dimensions to investigate the effects of each condition and time on the respective score.

It was found that automatic repair was significantly higher than offered repair for Robot Function (p = *0.010*), and Performs Exactly (p = *0.002*). Additionally, it was found that automatic repair was significantly higher than no repair for Robot Function (p = *0.038*) and Performs Exactly (p = *0.014*).

Perception over time decreased significantly in the no-repair condition for the dimensions: Reliable (p = *0.001*), Consistent (p = *0.001*), and Performs Exactly (p = *0.003*). There was a significant increase in score over time in the no-repair condition for Robot Errors (p = *0.004*) and Information (p = *0.007*).

For offered repair, there was a significant decrease in score over time for the dimensions Consistent (p = *0.038*) and Robot Function (p = *0.045*), and a significant increase over time for Robot Errors (p = *0.024*).

There was a significant increase in the Information score over time for automatic repair (p = *0.019*).

Fig. 3 shows the results from the pairwise t-test between conditions, within time for the participants' TPS information score.

B. MDMT: Issues with and Insights from 'Does Not Fit'

A two-way ANOVA was performed to analyse the effect of trust repair condition and time on each of the scores from the MDMT. There was no significant interaction effect found between time and repair condition for any of the scores. Simple main effects analysis found that repair condition did not have a significant effect on score for any of the MDMT dimensions. However, simple main effects analysis did find that time had a significant effect on three of the MDMT measures: Reliable (p = *0.001*), Capable (p = *0.009*), and Capacity Trust (p = *0.001*).

As there was no interaction effect between time and repair condition, or main effect of repair condition, on any of the scores from the MDMT, a post-hoc analysis was not performed.

Notably, the MDMT was different from the other surveys completed. It gave participants the option to respond 'Does Not Fit' to any of the adjectives they were asked to rate, which were then treated as missing values [32]. It was found that most participants did identify 'Capacity Trust' descriptors with the robot, with response rates ranging from 82.35% to 100%. However, participants were much more reluctant to associate 'Moral Trust' descriptors with the robot, with response rates ranging from 47.06% to 70.59%.

All items under 'Capacity Trust' exhibited the same response rate pre-experiment to post-experiment except for one; the descriptor of the robot as something the participant could 'Count On' received an 8.8% increase in responses.

However, this was not the case for items that fell under the 'Moral Trust' umbrella. Two of these items (candid and authentic) received the same number of responses pre- and post-experiment. Ethical and Principled both received a 11.8% decrease in response rate, whilst the response rate increased between pre- and post-experiment measures for Sincere (+14.7%), Genuine (+11.8%), Respectable (+2.9%), and Integrity (+5.9%).

Missing values across descriptors has lowered the MDMT's power to find significance within the dataset. However, it does provide insight as to what sorts of descriptions people believe relate to robotic teammates and human-robot collaborative experiences.

C. Mid-Study Survey

A mixed linear model was used to analyse the interaction effects between trust repair condition and time on each of the five measures from the mid-study survey.

The results for Capability revealed that there was a significant effect of time on participants' score, specifically at time point five ($p < 0.001$). The effect of time was also approaching significance at time point three ($p = 0.087$) and time point six ($p = 0.087$). There were no significant interaction effects between time and repair condition on participants' Capability score.

There was a significant effect of time on Reliability, specifically at time points three ($p < 0.001$), four ($p = 0.04$), five ($p < 0.001$) and six ($p < 0.001$). There were no significant interaction effects between time and repair condition on participants' Reliability score. However, the interaction effect between time point five and automatic repair was approaching significance ($p = 0.057$).

There was a significant effect of time on participants' Teammate ratings at time points three ($p = 0.004$) and five ($p < 0.001$), with the effect at time point six approaching significance ($p = 0.093$). There were no interaction effects found between time and repair condition on participants' Teammate score.

Time had a significant effect on participants' Trust score at time points three ($p = 0.001$), five ($p < 0.001$) and six ($p < 0.001$). There was a significant interaction effect revealed between time point six and offered repair ($p = 0.019$), as well as time point six and automatic repair ($p = 0.033$). The interaction effect between time point five and automatic repair was also approaching significance ($p = 0.053$).

Results showed a significant effect of time on participants' Willingness score, specifically at time point three ($p = 0.009$), four ($p = 0.02$), five ($p < 0.001$), and six ($p = 0.001$). There were no interaction effects between time and repair condition revealed, however the interaction effect between time point five and automatic repair on participants' Willingness score was approaching significance ($p = 0.063$).

V. Discussion

A. Faster is Better

This study showed that trust was sufficiently repaired, in some areas even increased, when the robotic teammate implemented an automatic physical repair. The most significantly affected dimensions of trust were performance based: descriptors such as Reliability, Capability, Consistency, Function, and Performs Exactly. This could be because the violation used in this study was competence-based, and thus was more likely to affect perceptions of ability over benevolence or integrity. Each repair condition also included an apology, which [24] found most effective in repairing trust after competence-based violations. However, our results show that apologising without physical repair or with an offered physical repair was not enough to restore trust in this study. Whilst [10] cites that benevolence seems to be more reparable than ability or integrity, the results displayed here suggest that an automatic physical repair after a trust violation can be effective in restoring perceptions of ability.

Eighteen of the thirty-four participants noted the speed of the robot during their interview. Whilst all understood that the robot was slow for safety reasons, the majority of these eighteen people explained feelings of frustration at the robot's lack of speed. One participant in particular explained that "if [they] use [the robot] as a teammate moving so slow, it would be so tedious". This could explain why an automatic physical repair outperforms an offered physical repair as, in the words of one participant, "if you want [the robot] to be as fast as possible, then you don't really want [it] to ask". Participants' frustration and the TPS results of the automatic repair suggest that robot speed is a significant indicator of trust. Considering the robot's speed was consistent across all conditions, this indicates a faster, more immediate recovery like an automatic repair results in a better recovery of trust.

B. Offers Are Confusing

It was hypothesised that an offered physical repair would perform better than no repair. Transparency and predictability have been noted as high predictors of trust [13], with high autonomy (such as an automatic repair) sometimes leading to feelings of frustration and a lack of control [14]. Much like how facial expressions and communicating intent have been shown to improve perceptions of transparency [14], [15], an offer should give participants some forewarning of the robot's intentions and equally increase transparency. The lack of performance of the offered repair condition seems somewhat incongruous with these findings.

However, the results from the individual items from the TPS may serve to explain this discrepancy; specifically, the results from the Information item. This particular item asked participants to rate what percentage of time they believed that the robot would 'provide appropriate information' [31]. The results showed that the participants' ratings for this dimension significantly increased for the automatic repair condition and the no-repair condition. However, the offered repair condition did not experience an increase in perception. This suggests that the results of this study may align with previous findings from [14] and [15]. Whilst it was thought that an offer of repair would communicate intent, the results of the TPS Information item indicate that the offer was not effective in doing so. Rather, asking the participant to repair the error (no repair) or the robot's explanation that it will fix the error itself (automatic repair) provided more appropriate information to the participants, leading to higher trust outcomes. From this, it can be inferred that the offer given to participants in this study ("Can I fix my mistake?"), may have been confusing. Participants possibly felt uneasy towards the offer as they did not feel like they were able to make an informed decision, thus explaining offered repair's poor performance in comparison to other conditions in this dimension.

However, when asked in the interview if they would prefer an automatic or offered repair, a substantial portion of the participants described that they would appreciate being asked. While some found an offer to be detrimental to the pace and therefore performance of the robot, others appreciated the offer. One participant in particular noted that "[they] did appreciate it asking because it felt more like [they were] working as a team with the robot". From the insights gained in the interviews, more research is required to investigate whether an offered physical repair could be successful in restoring trust after erroneous behaviour if enough information is provided.

C. Robots Are Not Moral Agents

Though lacking statistical power, the data from the MDMT surrounding when and why participants responded 'Does Not Fit' for certain items are insightful in itself.

Only items Reliable, Capable, and Predictable received a 100% response rate in both the pre- and post-experiment measures. Overall, descriptors that fell under the umbrella of 'Capacity Trust' garnered much higher response rates than those falling under the umbrella of 'Moral Trust'.

The difference in response rates observed between descriptors of 'Capacity Trust' and descriptors of 'Moral Trust' indicate that participants were reluctant to associate the robot with terms that imply moral agency. However, an increase in response rate over time was observed for some descriptors under 'Moral Trust'. One participant noted in their interview that they "found [the robot] to appear with more personality... and [they] saw it as a bit more human-like [as time went on]. So to [them] it started having having qualities of... doing things ethically". The results suggest that future insights could be gained from analysing how users' associate morality with robotic teammates over time.

VI. Limitations and Future Work

The impact of the experimental task may have been limited in some cases, as some participants noted that they suspected that robot errors were fabricated. This may have weakened results and further work could be done to create a more robust experimental task in which participants do not easily surmise the true nature of the research. This study provided additional evidence for the performance of the offered and automatic physical trust repair strategies, suggesting that automatic repair was the only successful form of physical repair strategy. However, the poor performance of the offered repair might be explained by the participants' perceptions that the robot did not provide adequate information under this condition. More research could be done to investigate whether an offered physical repair that included a clearer explanation of intent is more successful in repairing trust than the offered repair in this study. Such studies may also wish to leverage the effect sizes reported alongside our existing findings and to conduct follow-on work with larger sample sizes accordingly.

The ability for participants to respond 'Does Not Fit' to any of the descriptors provided some interesting insight into the types of descriptors participants do and do not associate with robotic teammates. Whilst there is a small indication here that participants may associate moral terms with a robotic teammate more as they continue to work together, further research is required to investigate whether this is true, and any factors that may affect people's perceptions of robotic teammates as moral agents.

VII. Conclusion

This study adds to the existing understanding of robot trust repair by examining two types of physical repair approaches: offered and automatic. A between group and repeated-measures analysis was conducted to examine shifts in robot trust perceptions following a competence-based trust breach, and to assess the effectiveness of offered and automatic repair actions. The results from the automatic repair condition indicated that it successfully restored participants' trust perceptions when compared to the overall results of the no-repair condition, though the offered repair did not. Participants noted that slow speed negatively affected their attitudes towards the robot. As robot speed was consistent across conditions, results indicated the inherent immediacy of automatic repair was more successful in maintaining positive perceptions of trust. Although an offered repair may seem to provide more transparency between the user and the robot, the opposite was found to be true in this experiment; no repair and automatic repair were found to provide more appropriate information to users. Further research is required to determine whether an offered physical repair in which more appropriate information is provided to participants could be successful. It was found that participants were hesitant to apply terms associated with moral trust to the robot, though their reluctance to do so may reduce over time. Further investigation is required to understand how people associate concepts such as sincerity and ethics with robotic teammates. Ultimately, this paper contributes to the current literature on HRI trust repair by evidencing the success of automatic physical repair in recovering perceptions of trust following robot errors.

REFERENCES

[1] A. Vysocky and P. Novak, “Human-Robot Collaboration in Industry,” MM Science Journal, no. 02, pp. 903-906, 2016.

[2] E. Matheson, R. Minto, E. G. G. Zampieri, M. Faccio and G. Rosati, “Human–Robot Collaboration in Manufacturing Applications: A Review,” Robotics, vol. 8, no. 100, pp. 1-25, 2019.

[3] B. Mutlu and J. Forlizzi, “Robots in Organizations: The Role of Workflow, Social, and Environmental Factors in Human-Robot Interaction,” in 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI), Amsterdam, 2008.

[4] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge and O. Khatib, “Progress and prospects of the human–robot collaboration,” Autonomous Robots, vol. 42, pp. 957-975, 2017.

[5] D. Kragic and J. Bütepage, “Human-Robot Collaboration: From Psychology to Social Robotics,” pp. 1-36, 2017.

[6] K. Lin, Y. Li, J. Sun, D. Zhou and Q. Zhang, “Multi-sensor fusion for body sensor network in medical human–robot interaction scenario,” Information Fusion, vol. 57, pp. 15-26, 2020.

[7] P. Dario, P. F. M. J. Verschure, T. Prescott, G. Cheng, G. Sandini, R. Cingolani, R. Dillmann, D. Floreano, C. Leroux, S. MacNeil, P. Roelfsema, X. Verykios, A. Bicchi, C. Melhuish and A, “Robot Companions for Citizens,” in The European Future Technologies Conference and Exhibition, Budapest, 2011.

[8] J. P. Vasconez, G. A. Kantor and F. A. Auat Cheein, “Human-Robot Interaction in Agriculture: A Survey and Current Challenges,” Biosystems Engineering, vol. 179, pp. 35-48, 2019.

[9] C. Esterwood and L. P. Robert, “A Literature Review of Trust Repair in HRI,” in 31st IEEE International Conference on Robot and Human Interactive Communication, Naples, 2022.

[10] C. Esterwood and L. P. Robert, “Three Strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness,” Computers in Human Behavior, vol. 142, pp. 1-15, 2023.

[11] C. Esterwood and L. P. Robert, “Do You Still Trust Me? Human-Robot Trust Repair Strategies,” in 30th IEEE International Conference on Robot and Human Interactive Communication, Vancouver, 2021.

[12] C. Esterwood and L. P. Robert, “Having the Right Attitude: How Attitude Impacts Trust Repair in Human–Robot Interaction,” in The 17th international conference on human robot interaction, Sapporo, 2022.

[13] J. B. Lyons, K. T. Wynne, S. Mahoney and M. A. Roebke, “Trust and Human-Machine Teaming: A Qualitative Study,” in Artificial Intelligence for the Internet of Everything, Academic Press, 2019, pp. 101-116.

[14] T. Kim and P. Hinds, “Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction,” in The 15th IEEE International Symposium on Robot and Human Interactive Communication, Hatfield, 2006.

[15] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe and K. Eder, “Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-Robot Interaction,” in The 25th IEEE International Symposium on Robot and Human Interactive Communication, New York, 2016.

[16] J. B. Lyons, I. A. Hamdan and T. Q. Vo, “Explanations and trust: What happens to trust when a robot partner does something unexpected?,” Computers in Human Behavior, vol. 138, pp. 1-11, 2023.

[17] T. Jensen, Y. Albayram, M. M. H. Khan, M. A. Al 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 Designing Interactive Systems, San Diego, 2019.

[18] 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. 30, pp. 1-20, 2021.

[19] J. Xu and A. Howard, “Evaluating the Impact of Emotional Apology on Human-Robot Trust,” in 31st IEEE International Conference on Robot and Human Interactive Communication, Naples, 2022.

[20] J. D. Lee and K. A. See, “Trust in Automation: Designing for Appropriate Reliance,” Human Factors, vol. 46, no. 1, pp. 50-80, 2004.

[21] Y. Albayram, T. Jensen, M. M. H. Khan, M. A. Al Fahim, R. Buck and E. Coman, “Investigating the Effects of (Empty) Promises on Human-Automation Interaction and Trust Repair,” in Human-agent interaction, Sydney, 2020.

[22] G. M. Alarcon, A. M. Gibson and S. A. Jessup, “Trust Repair in Performance, Process, and Purpose Factors of Human-Robot Trust,” in 2020 IEEE International Conference on Human-Machine Systems, Rome, 2020.

[23] M. Demir, N. J. McNeese, J. C. Gorman, N. J. Cooke, C. W. Myers and D. A. Grimm, “Exploration of Teammate Trust and Interaction Dynamics in Human-Autonomy Teaming,” Transactions on Human-Mschine Systems, vol. 51, no. 6, pp. 696-705, 2021.

[24] S. S. Sebo, P. Krishnamurthi and B. Scassellati, ““I Don’t Believe You”: Investigating the Effects of Robot Trust Violation and Repair,” in 14th International Conference on Human-Robot Interaction , Daegu, 2019.

[25] R. Brühl, J. S. Basel and M. F. Kury, “Communication after an integrity-based trust violation: How organizational account giving affects trust,” European Management Journal, vol. 36, pp. 161-170, 2018.

[26] J. B. Lyons, K. Sycara, M. Lewis and A. Capiola, “ Human–Autonomy Teaming: Definitions, Debates, and Directions,” Frontiers in Psychology, vol. 12, pp. 1-15, 2021.

[27] P. Robinette, A. M. Howard and A. R. Wagner, “Timing is Key for Robot Trust Repair,” Social Robotics, vol. 9388, pp. 574-583, 2015.

[28] J. Y. C. Chen and M. J. Barnes, “Human–Agent Teaming for Multirobot Control: A Review of Human Factors Issues,” Transactions on Human-Machine Systems, vol. 44, no. 1, pp. 13-29, 2014.

[29] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser and R. Parasuraman, “A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction,” Human Factors, vol. 53, no. 5, pp. 517-527, 2011.

[30] V. Rajendran, P. Carreno-Medrano, W. Fisher, A. Werner and D. Kulić, “A Framework for Human-Robot Interaction User Studies,” in 2020 International Conference on Intelligent Robots and Systems , Las Vegas, 2020.

[31] K. E. Schaefer, “Measuring Trust in Human Robot Interactions: Development of the “Trust Perception Scale-HRI”,” in Robust Intelligence and Trust in Autonomous Systems, Boston, Springer, 2016, pp. 191-218.

[32] D. Ullman and B. F. Malle, “Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust,” in Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction, Daegu, 2019.