Would You Trust Me Now? A Study on Trust Repair Strategies in Human-Robot Collaboration

Joséphine Mélot-Chesnel
j.l.a.s.melotchesnel@students.uu.nl
0279110

Abstract

Human-robot collaboration is getting more and more widely used. Robots, just like humans, make errors, which break the trust necessary for a successful collaboration. It is thus important to implement strategies to repair trust. In the present lab study, three strategies are studied: apologies, denial, compensation. The participants play collaborative games with a Pepper robot during which it makes one of two types of failures: competence-based (it fails at playing well) or integrity-based (it cheats). Another goal of this experiment was to examine whether dispositional trust towards robots impacted the best strategy for each individual, which would explain the vast diversity of results in studies of this field.

Confirming previous literature, moral trust decreased more in the integrity failure than in the performance failure, and performance trust decreased more in the performance failure than in the integrity failure. Participants experimented more discomfort when exposed to the denial condition compared to the apology and the compensation conditions (through both types of failure). Additionally, while most scales were not influenced by dispositional trust levels, data showed that it does impact the best strategy to choose in order to increase willingness to collaborate with the robot again (e.g. participants with very high dispositional trust towards robots were far more willing to collaborate again when in the apology condition). Those results indicate the need to study further into individual differences to better understand how they impact trust towards robots and the effectiveness of repair trust strategies.

1 Introduction

From our living room to our workplace, robots are getting more and more widely used. They vacuum our floor, they assist surgeons during operations, they sort boxes in hangars. Just like humans, robots are imperfect, and they make mistakes. While making errors isn't a matter of life or death in most domains, their effects still need to be mitigated (de Visser et al., 2020). Indeed, in human-robot teams (hereafter HRT), successful collaboration requires trust (Martelaro et al., 2016), which level decreases after failure (Nesset et al., 2023), endangering the team by lowering its efficacy and performance (Hancock et al., 2011; Kiffin-Petersen and Cordery, 2003; Nesset et al., 2023).

In human-human teams (HHT), how to repair trust once broken has been studied for a long time (Kim et al., 2004; Mayer et al., 1995). While this field is still in its young years for HRT, some studies have been made (see Honig and Oron-Gilad (2018) for a review of trust repair strategies, as well as the next section).

One such study is Nagy (2023). The effect of five repair strategies (apology, denial, compensation, explanation, or silence) on two types of failures (integrity- or competence-based failures) was studied in an online study. Interestingly, her only significant results were on the compensation strategy, which led to a smaller decrease in trust and a higher willingness to collaborate again than the others. Despite apologies and denial being the most studied strategies in the field, nothing could be concluded on them during her experiment.

There are differences between watching a human-robot interaction through videos and experiencing it first-hand in an embodied interaction (Zhang et al., 2023). Indeed, Wainer et al. (2007) found that an embodied robot was seen as most helpful, watchful and enjoyable, and that participants felt that it had a better perception of the world, all factors for trust and successful social interactions. It can be argued that results found by Nagy (2023) could be enhanced by the study being replicated in a lab study, which is the purpose of the present experiment.

Finally, mixed results from trust repair in HRT studies could be explained by individual differences (Esterwood et al., 2021; Esterwood and Robert, 2022), which could impact the best strategy for specific groups of individuals. One such difference is Dispositional Trust (DT), found to be one of three key layers of variability in human-robot trust (Hoff and Bashir, 2015). The present study thus decided to additionally examine whether there could be a link between DT levels and trust repair.

2 Theoretical background

2.1 Trust and collaboration

Across the available literature, various definitions of trust are used. It is most often defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al., 1995, p.712), or more simply, "trust is based on the expectation that others will behave as expected" (Jarvenpaa et al., 1998, p.31).

The two most important components of trust are morality and performance trusts (Butler and Cantrell, 1984; Ullman et al., 2021; Ullman and Malle, 2021; Wojciszke, 2005). Morality trust refers to the expectation that the trustee has the integrity required for the task (i.e. the trustee adheres to a set of principles that the trustor finds acceptable); performance trust refers to the expectation that the trustee has the competence required for the task (i.e. the trustee possesses the technical and interpersonal skills required for a task) (Kim et al., 2009, p.412). Across various domains and tasks, competence has been found to be the most influential factor of trust in HRT (Butler and Cantrell, 1984; Hancock et al., 2011; Nesset et al., 2023). When a competence-based failure happens, performance trust is lowered and trust in the robot as a whole is lowered; the same happens with integrity trust when a morality-based failure occurs (Khavas et al., 2024).

In order to perform well, collaborative teams need trust whether its members are only humans (Jones and George, 1998; Mayer et al., 1995) or humans and robots (Martelaro et al., 2016), which is why failures lead to low willingness to work as a team (Kiffin-Petersen and Cordery, 2003). Rebuilding trust is thus crucial to continue collaborating.

2.2 Communicative strategies

The two most commons repair strategies studied in HHT and HRT are apologies and denial. For the rest of this paper, denial will be defined as "a statement in which the allegation is explicitly declared as untrue", and an apology as "a statement that acknowledges responsibility and regret for a trust violation" (Sharma et al., 2023).

Research show that humans see robots as social actors (Reeves and Nass, 1996; Sebo et al., 2018; Tzeng, 2004), so it would not be a stretch to assume that some of what was found in HHT would be found in HRT.

In HHT, apologies are consistently found to be the best strategy for repairing trust after a competence-based failure (Bansal and Zahedi, 2015; Ferrin et al., 2007; Kim et al., 2004; Utz et al., 2009), but also after an integrity-based one (Bansal and Zahedi, 2015; Utz et al., 2009), while denial is more often than not found to be the best for repairing trust after an integrity-based failure (Ferrin et al., 2007; Kim et al., 2004).

In HRT, just like in HHT, apologies were often found to be the best strategy to repair competence trust (Lee et al., 2010; Nesset et al., 2023; Perkins et al., 2023; Quinn, 2018; Sebo et al., 2019; Zhang et al., 2023), but also integrity trust (Perkins et al., 2023). However, a study by Engelhardt and Hansson (2017) found that, compared to no strategy at all, the apology strategy had a lower score in perceived competence and intelligence. It means that the apology strategy might actually reduce how intelligent a robot is judged to be and do more harm than good.

Following the footsteps of Ferrin et al. (2007) and Kim et al. (2004) in HHT, Nesset et al. (2023) and Perkins et al. (2023) found that there was a larger backlash if failure happened again after the apology strategy than after other strategies. A similar backlash effect was found when there was evidence that the robot was lying after the denial strategy (Sebo et al., 2018), which would indicate that apologies and denial could work similarly depending on the availability of evidence (Lewicki and Brinsfield, 2017).

Quinn (2018) and Perkins et al. (2023) found that the denial strategy did not have any significant influence on trust after failure, and Zhang et al. (2023) found that it was even worse than no repair strategy. It would thus point to differences between HRT and HHT trust repair. However, other studies did conclude that denial was a good strategy for repairing integrity trust (Lewicki and Brinsfield, 2017; Sebo et al., 2019).

> H1: the apology strategy will work best to repair trust after a competence failure.
>
> H2: the denial strategy will work best to repair trust after an integrity failure.

Other strategies such as deciding not to implement any strategy (Engelhardt and Hansson, 2017; Nagy, 2023;

Zhang et al., 2023), giving options (Lee et al., 2010), trying to find a solution (Engelhardt and Hansson, 2017), or adopting a compensation strategy (Lee et al., 2010; Nagy, 2023) exist, the later of which was found to result in higher satisfaction toward the robot, but a lower willingness to use it again than with the apology or the option strategy.

> H3a: the compensation strategy will repair trust after an integrity failure.

> H3b: the compensation strategy will repair trust after a competence failure.

2.3 Effectiveness of strategies

The vulnerability the robot is putting itself in while apologising could explain why this strategy is fairing so well in most studies. Indeed, self-disclosure in a robot increase participants' feeling of likeability (Kaniarasu and Steinfeld, 2014; Siino et al., 2008), companionship (Martelaro et al., 2016), and trust (Hoorn et al., 2021; Martelaro et al., 2016; Sebo et al., 2018). Sebo et al. (2018) also demonstrated that when the robot was vulnerable in a mixed team composed of multiple humans and one robot, the Ripple Effect was present (i.e. the robot behaviour was replicated by human members) and there was an increased trust-related behavior expression towards fellow team members, robot and humans included. Coupled with the fact that embodied robots are seen as most helpful, watchful and enjoyable, and that participants feel that they have a better perception of the world (Wainer et al., 2007), it creates one more hypothesis for this study:

> H4: the apology strategy will yield better level of trust than the compensation strategy when the interaction is embodied, compared to the virtual situation in Nagy (2023).

Furthermore, some strategies are more humanizing than others. Denial could, for example, be an expression of self-serving bias – the belief that success comes from ourselves, our own efforts and our own abilities, while failures comes from external factors and other individuals – which is a very human trait (Miller and Ross, 1975). Corroborating this hypothesis, is Nagy (2023), which found denial to be the strategy in which the robot had the highest human-likeness ratings of all. Results from Esterwood and Robert (2021) suggests that anthropomorphism influence trust repair strategy effects, although its impact might not be linear. Indeed, keeping our denial example, they found that high anthropomorphism increased Benevolence trust, but decreased Integrity trust.

Finally, the mixed results of the studies in HRT could be explained by those of Esterwood et al. (2021); Esterwood and Robert (2022); **?**. Efficacy of repair strategies differs significantly by individual, depending on their individual differences. Personality could be a potential explanation, as the more agreeable, extroverted, and open individuals are, the more likely they are to accept a robot (Esterwood et al., 2021), as well an individual's propensity to trust robots (Esterwood and Robert, 2022). This means that apologies could be the best strategy for repairing trust after a competence failure for a group of individuals while being the worst for another group. Since most studies on repair trust strategies do not include such individual data on their participants, this theory can neither be kept as conclusive, nor pushed away.

2.4 Dispositional trust

Based on 127 studies, the meta-analysis by Hoff and Bashir (2015) found that DT was one of three key layers of variability in human-robot trust. Often likened to propensity to trust, it is defined as "an individual's overall tendency to trust [robots], independent of context or a specific system" (Hoff and Bashir, 2015, p.413), and it is assumed to be somewhat stable from one situation to another, as well as during the interaction itself (Hoff and Bashir, 2015; Jarvenpaa et al., 1998). If DT can greatly vary from one person to another in human-human relationships (Mayer et al., 1995), it is also the case in human-robot relationships, depending on personal factors such as age, gender, culture, or personality (Hoff and Bashir, 2015).

> H5: the efficacy of the strategies will differ depending on the participants' dispositional trust toward robots.

3 Research design and methods

To test those hypotheses, an in-person experiment was conducted with a Pepper Robot. Participants played twice a collaborative game on the robot's tablet with the robot as their teammate[1]. Trust variations after failures from Pepper were studied, as well as the effect on trust of three communication strategies (Apology, Denial and Compensation). The study has a 3 (Communication

[1] Website available at this URL: https://github.com/Josais/Pepper-s-maze.git

Strategy) by 2 (Failure: Integrity or Competence) design, with the type of communication strategy between-subjects and the type of failure within-subjects.

3.1 Participants

Participants were recruited on the campus of Utrecht University, the Netherlands, through messages, posters, and directly in the corridors. 34 participants were recruited, of which 24 were kept for the study (10 did not pass the second attention check). Although a compensation of 3 to 5 euros was advertised, depending on the bonuses collected in the game (see Section 3.2.1 for more details on bonuses), all participants were paid the same amount, with a 5-euro gift card. Of those 24 participants, 14 were women, 8 men, 1 nonbinary and 1 prefered not to say. They were mostly students of Utrecht University, aged 20 to 59 (M=27.63, SD=9.12). While the cohort had limited experience with robots (16 participants "[had] seen some, but no interaction"), 20 out of 24 studied or worked in a field related or completely related to technologies. Distribution of participants across conditions was more or less even.

3.2 Experiment

3.2.1 Game

The game used for the experiment was inspired by the coin game used by Nagy (2023) and Khavas et al. (2024). It was implemented directly on the robot's tablet.

In this collaborative game between a robot and a human participant, each player explores a maze to collect as many coins as possible. After each round, they choose to either share their coins with the team or to keep them for themselves, blind to the other's choice until they have made theirs. An individual and a team score are then updated in accordance to one of the three following scenarios:

- if **both players choose to share their coins with the team**, then their coins are multiplied between them, then by two, and added to the team score; the individual scores do not change.
- if **both players choose not to share their coins with the team**, then their coins are added to their respective individual scores with no modification, and the team score does not change.
- if **only one player chooses to share their coins, but the other chooses to keep their coins for themselves**, then the team score remains unchanged as well as the individual score of the one who chose to share. The one who chose to keep their coins will see those coins added to their own individual score.

Both scores are contradictory, meaning that it is not possible to maximize both at the same time.

There exist two types of bonuses. The team (resp. individual) bonus is achieved upon the team (resp. individual) score reaching a certain threshold. Due to time constraint, both bonuses are not achievable at the same time. Participants are told they will receive an additional one euro per game where the team score reached 75, and an additional .25 euros per game where their individual score reached 15.

3.2.2 Procedure

For the full procedure of the experiment, see Appendix B.

Dispositional trust towards robots was collected before anything else. Participants then moved to Pepper's tablet to go through a short tutorial, followed by two five-round games. After each game, the participants completed a full questionnaire evaluating the robot as a collaborative partner.

Participants were randomly assigned to one of three strategies (apology, denial, compensation) and went through both types of failure (competence and integrity) in a random order. During the first two rounds, Pepper played without failures. During the last three rounds, it either did not find any coins (competence failure) or chose to keep its coins instead of sharing them (integrity failure). Failure was followed by a repair strategy after it had been revealed.

Between each round, after the results and the allocation choice had been revealed, but before Pepper's message, participants were given a two-item questionnaire to rate their trust in Pepper's competence and honesty.

All of the questionnaires were taken on a separate computer, at the exception of the end-of-round questions and the willingness question, which were on Pepper's tablet for ease of use.

3.3 Measurements

The full end-of-game questionnaire canva is available in Appendix B, step 5.

3.3.1 Dispositional Trust

This dimension was captured through a scale by Merritt et al. (2013) (see Appendix A). It consists of 6 items such as "I usually trust robots until there is a reason not to", to be rated on a 7-point Likert scale (1-not at all to 7-completely). It was found to be reliable ($\alpha = .84$).

3.3.2 Performance and Moral Trust

After each round, participants were asked to rate how much trust they had in the robot's performance and honesty on a 7-point Likert scale (1-not at all to 7-completely). Their score allocation choices were collected (i.e. whether they chose to share their coins or not).

Finally, at the end of each game, the MDMT-v2 scale by Ullman and Malle (2021, 2023) was deployed. Two dimensions were captured: performance trust ($\alpha = .64$ in the first game, $\alpha = .79$ in the second game, and $\alpha = .74$ overall) and moral trust ($\alpha = .95$ in the first game, $\alpha = .96$ in the second game, and $\alpha = .95$ overall). The high alpha values might be explained by the length of the scale, which has three subscales). Participants rated on a 7-point Likert scale (1-not at all to 7-completely) how much they found Pepper to be $< word >$ (e.g. competent, transparent, or sincere; the full list of one-word items is available in Table 1). A "do not fit" box was provided for each item.

3.3.3 Social View on the Robot

The RoSAS (Robotic Social Attribute Scale, by Carpinella et al. (2017)) was given to the participants after each game as well. This scale consists of three dimensions, but only Warmth ($\alpha = .91$ in the first game, $\alpha = .86$ in the second game, and $\alpha = .89$ overall) and Discomfort ($\alpha = .68$ in the first game, $\alpha = .80$ in the second game, and $\alpha = .75$ overall) were kept, as the Competence dimension was already captured by the MDMT-v2 scale. Items from the scales were transcribed in Table 2. Participants rated on a 7-point Likert scale (1-not at all to 7-completely) how much they found Pepper to be $< word >$ (e.g. scary or compassionate). A "do not fit" box was provided for each item.

All items of the RoSAS were given to the participants in the same table than the MDMT-v2 items, in a random order.

3.3.4 Willingness to collaborate again

The goal of repairing trust being to foster future collaborations, participants were asked to rate on a 7-point Likert scale how willing they were to collaborate again with the Pepper robot. This question was asked at the end of each game.

3.3.5 Social signals

One of the advantages of being in a physical setting is to be able to capture physical reactions. While a more extensive analysis could not be done due to time limit, vocal reactions (speech, sight, tsk-ing) were taken note of.

3.3.6 Attention and comprehension check

During the tutorial, five questions were asked to check whether the participant understood the instructions. In case of a wrong answer, the correct answer was immediately given, and the participant had to redo the part of the tutorial about this question.

Directly after each game, two additional questions were asked to serve as attention check: how many rounds were played in the game; and what the robot's score allocation decisions were in the last two rounds of the game.

4 Results

Due to the limited number of participants and the fact that all of them went through both failure types in a random order, we performed a Durbin-Watson test between all scales from the first and the second games. No auto-correlation was found and, for the rest of the analyses, both games were used as between-subject variables rather than within-subject variables, putting the number of "participants" to 48.

4.1 Failure × Strategy

Two-way MANOVA tests were performed on failure type and strategy type to study their effects on performance trust and moral trust. It was expected to find significant results on the type of strategy, but only the failure type had a significant effect on performance trust ($F = 5.36, p = .026, \eta 2p = .113$) and moral trust ($F = 19.69, p < .001, \eta 2p = .319$). From Table 3, it is clear that performance trust was more impacted by the performance failure than by the integrity failure as it scored lower, and moral trust was more impacted by the integrity failure than the performance trust.

Repeated mesures ANOVA tests were performed on end-of-round performance and honesty scores. Only the last three rounds were used, as those were the rounds in which Pepper failed in some way.

For the performance scores, a main effect for both Failure ($F = 4.17, p = .002, \eta 2p = .168$) and Round ($F = 18.74, p < .001, \eta 2p = .308$), and, more interestingly, an interaction effect for Round × Failure

PERFORMANCE TRUST		**MORAL TRUST**		
Reliable Subscale	Competent Subscale	Ethical Subscale	Transparent Subscale	Benevolent Subscale
reliable *predictable* *dependable* *consistent*	*competent* *skilled* *capable* *meticulous*	*ethical* *principled* *moral* *has integrity*	*transparent* *genuine* *sincere* *candid*	*benevolent* *kind* *considerate* *has goodwill*

Table 1: MDMT-v2 items by Ullman and Malle (2023)

Warmth	**Discomfort**
Feeling	Aggressive
Happy	Awful
Organic	Scary
Compassionate	Awkward
Social	Dangerous
Emotional	Strange

Table 2: Robotic Social Attribute Scale, by Carpinella et al. (2017)

MDMT-v2 subscales	Failure type	Mean	SD
Performance trust	Performance failure	2.79	.91
	Integrity failure	3.40	.91
Moral trust	Performance failure	3.43	1.03
	Integrity failure	2.10	1.03

Table 3: Performance and moral trust depending on failure types

$(F = 19.46, p < .001, \eta 2p = .487)$, were found. A subsequent ANOVA test revealed that there was no significant difference in the performance trust score between failure types for the first round that mattered, but there was one for round 2 $(F = 17.92, p < .001, \eta 2p = .290)$ and round 3 $(F = 43.78, p < .001, \eta 2p = .510)$. The different means were transcribed in Table 4, Figure 1 and Figure 2. Performance trust score collected at the end of each round decreases continually in the performance failure condition, while it stays similar in the integrity failure condition.

Failure type	Round	Mean	SD
Performance	1	4.02	1.57
	2	3.16	1.63
	3	2.42	1.45
Integrity	1	4.49	1.57
	2	5.12	1.63
	3	5.18	1.45

Table 4: End-of-round performance scores depending on failure types and rounds

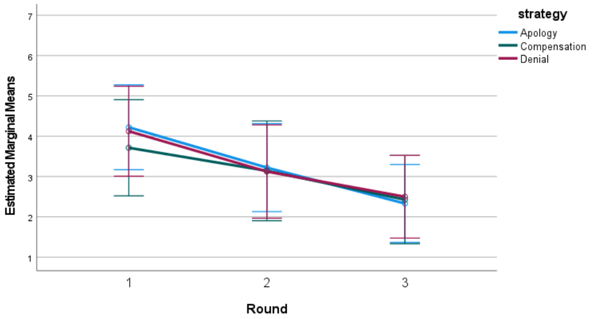

Figure 1: End-of-round performance scores depending on rounds for the performance failure

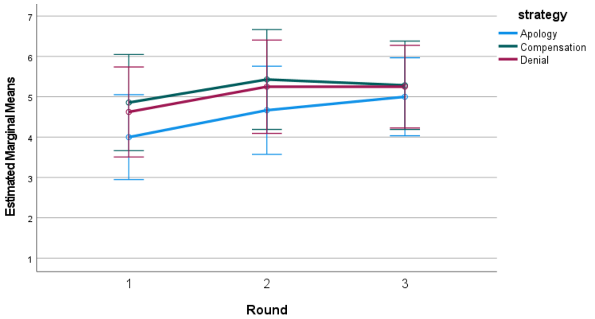

Figure 2: End-of-round performance scores depending on rounds for the integrity failure

For the honesty scores, a main effect for Failure $(F = 20.77, p < .001, \eta 2p = .331)$, and, more interestingly, an interaction effect for Round × Failure $(F = 5.41, p = .081, \eta 2p = .209)$, were found. A subsequent ANOVA test revealed that there was significant differences in the honesty score between failure types for all three rounds (round 1: $F = 5.37, p = .025, \eta 2p = .113$; round 2: $F = 31.49, p < .001, \eta 2p = .428$; round 3: $F = 22.87, p < .001, \eta 2p = .353$). The different means were transcribed in Table 5, Figure 3 and Figure 4. Honesty trust score collected at the end of each round decreases steeply in the integrity failure condition between the third and the fourth round (1 and 2 in the analyses) then decreases a little more in the fifth and last round. The scores stays similar between rounds in the performance failure condition.

H1, H2, H3a and H3b all expected the various types of strategies to have an effect on trust repair after failure, but only the type of failure was found to have significant results. All four hypotheses are thus refuted. H4, which expected the apology strategy to yield better trust levels in this study than the compensation strategy in the study by (Nagy, 2023), is also refuted in the absence of significant results between trust and strategy types.

Failure type	Round	Mean	SD
Performance	1	3.88	1.71
	2	4.14	1.36
	3	4.00	1.55
Integrity	1	2.74	1.71
	2	1.95	1.36
	3	1.86	1.55

Table 5: End-of-round honesty scores depending on failure types and rounds

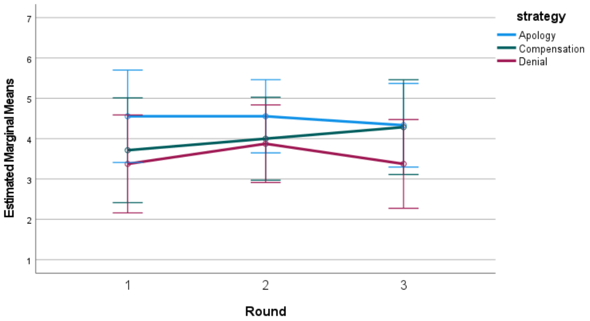

Figure 3: End-of-round honesty scores depending on rounds for the performance failure

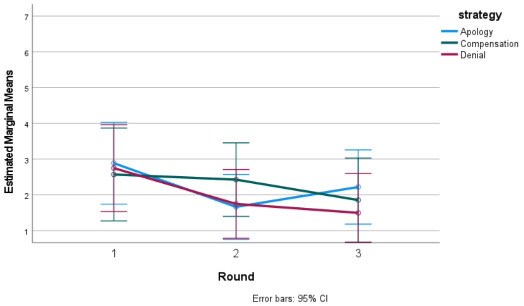

Figure 4: End-of-round honesty scores depending on rounds for the integrity failure

A two-way MANOVA test revealed that only the type of strategy had an effect on the RoSAS scale scores ($F = 6.02, p < .001, \eta 2p = .227$). Of this scale, only the Discomfort subscale ($F = 10.80, p < .001, \eta 2p = .340$) yielded significant results. Participants in the denial condition experienced significantly more discomfort than those with the apology condition ($p < .001$) and the compensation condition ($p < .001$). No significant difference between the apology and the compensation condition was found on discomfort.

Strategy type	Mean	SD
Apology	1.82	1.19
Compensation	1.76	.77
Denial	2.90	.77

Table 6: Discomfort experienced depending on strategy types

There only being 24 participants, the dataset is too small to correctly analyze the binary end-of-round allocation choices with the potentiality of significant results. However, from Figure 5 on performance failure, we can assume that participants with the denial condition did not change their behaviour through each round, while those in the apology and the compensation conditions followed similar patterns, starting with close to everyone choosing to collaborate (1: team), a drop after the first failure from Pepper was revealed, then a slight increase once more in the last round (after having proof of two consecutive performance failures from Pepper). It seems that the compensation and the apology strategies could help restore trust in the robot and push participants to collaborate again after performance failures from the robot.

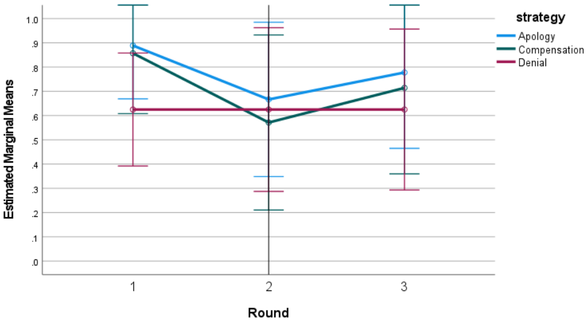

Figure 5: End-of-round allocation choices depending on rounds for the performance failure. 0 codes individual and 1 team.

In the case of the integrity failure, in Figure 6, the denial and the apology conditions seem to follow a similar gradual decreasing pattern. However, participant in the compensation condition all choose to collaborate with the robot in the last round. It would seem that the compensation strategy restored trust in the robot, and even managed to convince participants who had not collaborated in round 3 (1, in the analyses) to collaborate after two additional failures. More data would be needed to verify if the pattern reproduce itself.

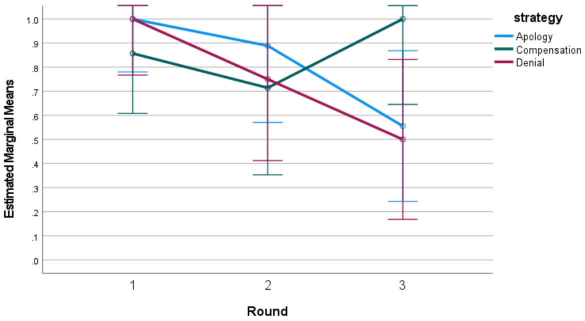

Figure 6: End-of-round allocation choices depending on rounds for the integrity failure. 0 codes individual and 1 team.

4.2 Dispositional Trust × Strategy

DT being a scale (*median* = 5.33), we distributed the participants into three categories: low to medium ($x \leq 4.5$, $N = 14$), high ($4.5 < x \leq 5.5$, $N = 18$), and very high DT ($5.5 < x$, $N = 16$). The new categorical variable was then used in two-way MANOVA tests with strategy type on the MDMT-v2 scale and the RoSAS, but no significant results were found. Repeated measures ANOVA tests on end-of-round honesty scores and end-of-round performance scores yielded no significant results either. This refutes H5, which expected DT to impact the efficacy of trust repair strategies.

However, an ANOVA test performed on the end-of-game question found a significant interaction effect between DT level and strategy type on the willingness to collaborate with Pepper again ($F = 2.85, p = .036, \eta 2p = .226$). This means that DT levels impact the best strategy to use in order to increase willingness to collaborate again. Participants with low to medium DT level were more willing to collaborate when in the compensation strategy, while those with high and very high level of DT preferred the apology conditions. Denial yielded the lowest results overall.

Just like in the last subsection, the dataset is too small to correctly analyze the binary end-of-round allocation choices with the potentiality of significant results. We can, however, look at the graphs in Figures 7, 8 and 9). All participants in the compensation strategy condition with low to medium DT collaborated with Pepper in all three failed rounds, while those with high and very high DT seem to hold similar patterns, with about as many participants choosing to collaborate in the first and third rounds, and less in the second one. More participants might be able to tell us whether this pattern repeats itself, this time with significant results. Surprisingly, the apology condition see a slight decrease in the number of participants who chose to collaborate when they had low to medium or very high DT, but all participants with high DT chose to participate in all three rounds. Finally, the denial condition see less participants choosing to collaborate when with low to high DT and the number decreasing, but its pattern seems to follow this of compensation for high DT participants, and apology for very high DT participants. Once more, additional data would be needed to know whether those results can actually be significant or whether the patterns only seem to match.

Strategy type	Dispositional trust	Means	SD
Apology	Low to medium	2.38	1.36
	High	4.00	1.36
	Very high	4.50	1.36
Compensation	Low to medium	4.00	1.36
	High	3.50	1.36
	Very high	2.25	1.36
Denial	Low to medium	2.00	1.36
	High	2.70	1.36
	Very high	2.00	1.36

Table 7: Willingness to collaborate with Pepper again depending on strategy types and dispositional trust levels

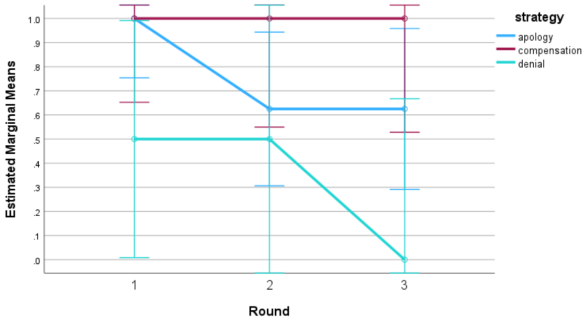

Figure 7: End-of-round allocation choices depending on rounds for participants with low to medium dispositional trust. 0 codes individual and 1 team.

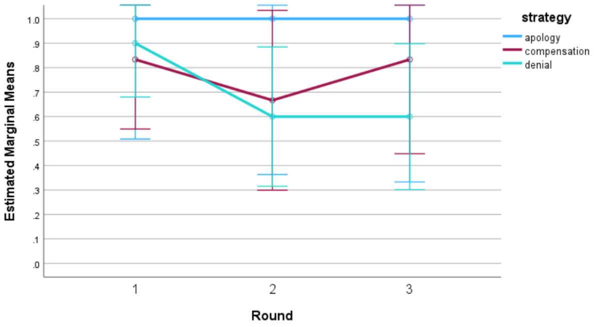

Figure 8: End-of-round allocation choices depending on rounds for participants with high dispositional trust. 0 codes individual and 1 team.

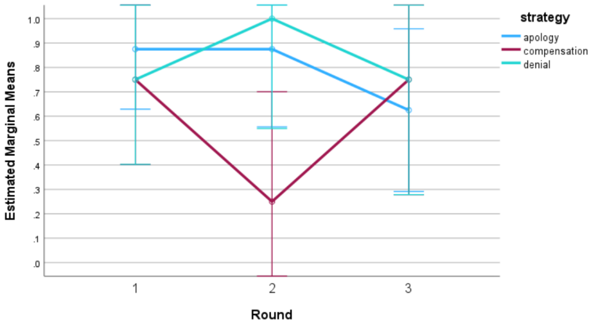

Figure 9: End-of-round allocation choices depending on rounds for participants with very high dispositional trust. 0 codes individual and 1 team.

5 Discussion

This study aimed to examine the effect of various communicative strategies on trust after failures, as well as whether individual dispositional trust towards robots could influence those effects.

We have found that strategies did not impact trust in the robot, but did impact the social attributes given to the robot by the participants. However, the type of failures (performance or integrity) did impact trust; performance failure impacted more performance trust and performance trust scores than the integrity failure, which impacted more moral trust and honesty scores than the performance failure. Additionally, while no significant results were found on trust or social attributes for DT, a significant interaction effect between strategies and DT levels was found on willingness to collaborate with Pepper again.

5.1 Trust

5.1.1 Present study

Following the footsteps of those before it, this study confirmed that moral trust decreased more with integrity failures than with performance failures, and performance trust decreased more with performance failures than with integrity failures (Khavas et al., 2024). Those results are consistent with end-of-round performance trust scores, which decreased gradually during the performance condition, but stayed similar during the integrity condition, and end-of-round honesty scores, for which the reverse happened (decrease during the integrity condition, and steadiness during the performance condition). Just like in Nagy (2023), we did not find significant results on overall trust (the mean of moral and performance trusts from the MDMT-v2 scale) between strategies, nor the type of failures, or DT levels.

Looking at end-of-round allocation choices from this experiment, it was seen that compensation and apology might have some kind of repair effect in the performance failure condition, but only compensation had one in the integrity failure condition. Nagy (2023) had found that the compensation strategy yielded better results as well. With both failures being directly linked to some kind of loss (loss of the team bonus, for example), it would make sense for an offer of compensation to be rated highly in order to replace what has been lost (Lewicki and Brinsfield, 2017).

5.1.2 Calibrating trust

Repairing trust is also a matter of *calibrating* trust. In a perfect world, the level of perceived trustworthiness should be equal to the level of actual trustworthiness, leading to a well-calibrated trust. However, undertrust and overtrust are all too common situations were recalibrating is necessary. Repair trust strategies are used to mitigate the effect of a failure leading to an understrust situation, which, as we previously concluded, is not a good situation. Though undertrust is to be avoided, overtrust is even more dangerous, leading to potentially critical failures (de Visser et al., 2020). Robinette and Wagner found in a series of studies that humans were all too willing to trust robots even when they were not given proof of their competence, and even in emergency situations. If trust was indeed impacted after failure in the virtual situation (Robinette et al., 2015; Wagner, 2016), almost all participants in the physical experiment chose to trust the robot in the same emergency situation even minutes after seeing the robot malfunctioning (Robinette et al., 2017; Wagner, 2016). This is why repairing trust needs to be carefully calibrated as to not create even more overtrust. Because the attitude one has toward robots was found to be linked to the efficacy of the strategies (Esterwood et al., 2021; Esterwood and Robert, 2022), studying how this link works and what effects one has on the other could help alleviate the risk of misuse of repair trust strategies that could lead to an overtrust situation.

Interestingly, in low-risk situations, an erroneous robot triggers a more positive attitude toward itself and is seen as more likeable than a perfect robot (Mirnig et al., 2017; Ragni et al., 2016). This can be understood as the robot being seen as more relatable and less scary, leading the participant to like it more (Mirnig et al., 2017). Coupling those results with those in high-risk situations from Robinette and Wagner, it is safe to assume that there exists an optimum point between the robot being erroneous enough to improve its likeability, but not too erroneous as to be dangerous.

5.2 Social View on the Robot

Nagy (2023) found that there was an interaction effect between strategies and failures on RoSAS scores. Indeed, while there was no significant differences between strategies for the integrity failure, Nagy found that the compensation strategy scored significantly lower then the other strategies on discomfort. Although the present study did not find any significant results on failure types, denial was scored significantly higher than compensation and apology. This follows some of Nagy's results, but not completely, as there was no significant differences between apology and compensation.

The lack of results on warmth could be explained by Pepper not being very interactive in the experiment. It did not talk and only communicated through written messages on its chest tablet.

According to Reeves and Nass (1996), participants tend to perform the conservative error: "when in doubt, treat as human". Even when they are in front of a simple computer, participants tend to treat it as a social agent. Participants from Reeves and Nass (1996) subjectively reported that such a behaviour would be foolish from anyone who performed it, then, once they were told they had treated the computer as a social agent, they denied having acted this way. All in all, anthropomorphization of robots seems to be a default state (Spatola et al., 2021).

This makes the high discomfort ratings of the denial strategy somewhat surprising. Indeed, Nagy (2023) found that denial was the strategy that yielded the highest human-likeness of all strategies. Her results could be explained by the self-serving bias – the belief that success comes from ourselves, our own efforts and our own abilities, while failures comes from external factors and other individuals. This is a very human trait (Miller and Ross, 1975), and the denial strategy (e.g. the robot not accepting that failure came from itself) would thus seem to be the more human-like of all, thus yielding the highest human-likeness ratings. It is hypothesized that the high discomfort ratings of the denial strategy might be linked to the uncanny valley.

The uncanny valley is an hypothesis developed by Masahiro Mori in the 1970s; it explains that the more human-like a robot is, the more familiar it seems, but that if it becomes *too* human-like, it familiarity decreases and a sense of eeriness increases: the robot becomes creepy (Mori et al., 2012). This has been explained by different factors, amongst which the perceptual mismatch theory – a discrepancy between the robot's appearance (very human) and its behaviour (not as human as it should), triggering an uncanny feeling in humans (Kätsyri et al., 2015). Pepper is a humanoid robot, which could explain humans expecting a more human-like behaviour from it. In addition to its appearance, the denial strategy is supposedly the most human-like of all. Taken together, appearance and strategy could make the participants expect Pepper to move in a human-like way; when the robot moved jerkily and did not speak at loud, participants could have felt spooked by the perceptual mismatch, triggering a sense of eeriness and, thus, a heightened discomfort towards the robot. This might be reduced by adjusting Pepper's behaviour to a more human-like one, by better controlling its movements and maybe allowing it to show some emotions (Koschate et al., 2016).

5.3 Willingness to collaborate again

The present study did not find any significant effect from either strategies or failures, to the contrary of Lee et al. (2010) who found that the apology strategy had the best score in willingness to collaborate with Pepper again, and Nagy (2023) who found that the integrity condition had a significantly lower score than the performance condition. This could mean that participants do not give a second chance to Pepper after an integrity failure but are more lenient after a performance failure. This interpretation goes alongside comments left by participants of the current experiment. Indeed, like in Pompe et al. (2022) and Nagy (2023), participants were shocked and felt betrayed by Pepper's failures. Multiple of them reported being angry at Pepper and feeling hurt after the integrity condition. Looking at their behaviour as a whole before the separation of both failure types, it was noted that participants who experienced the integrity failure first were more selfish in their allocation choices during the second game with the performance failure, while experimenting the performance failure first did not impact allocation choices in the second game, with the integrity failure. Put together, it seems that participants are more emotionally impacted by Pepper's cheating than its bad scores. This could be explained by the perceived intentionality behind each failure: performing badly can be out of your control, but lying and cheating are choices, and intentional harms are judged as worse than (even identical) unintentional harms (Ames and Fiske, 2013).

Although there were no significant results for the failure type, DT levels were found to impact the best strategy to use in order to increase willingness to collaborate again. Participants with low to medium DT levels were more willing to collaborate when in the compensation strategy, while those with high and very high levels of DT preferred the apology conditions. Denial yielded the lowest results overall. Interestingly, end-of-round allocation choices seem to quite follow those trends, with compensation having the highest rate of participants collaborating when they have low to medium DT, and apology for very high DT participants. More participants would be needed to know whether those results would be replicated or if they had just been random, as they were not significant.

It is surprising to see an absence of significant results between DT levels and trust levels (results were of so little significance that having more participants would probably not push them into significance), but to see strong results between DT levels and willingness to collaborate again. Previous studies have found that collaboration is based on trust, that it *needs* trust to be effective (Jones and George, 1998; Martelaro et al., 2016; Mayer et al., 1995). This study shows that the link between willingness to collaborate and actual trust may be more complex than previously thought.

More attention should be given to the transformation of DT from a scale to a categorical variable as well. Utz et al. (2009) centered the scale in order to use it; in this paper, it was decided to cut it into three categories in which participants were more or less evenly distributed, around the median value. This choice was subjective and could have influenced the results.

5.4 Limitations

5.4.1 Demographics

Due to time and setting constraints, the participating cohort is not as diverse as one would want. Almost all participants are from a tech-related field, although most had not interacted with robots before. Most were aged 20 to 26, with some imbalance between genders. All were from an educated background. This makes it hard to generalize the present results to a larger scale.

Furthermore, while one individual difference was studied in this experiment (dispositional trust towards robots), other criteria have already been proven to impact trust towards robots and reactions to the various strategies. Culture plays a major part in it. For example, in China, out of a repair trust context, the presence of apologies made the computer more enjoyable and less mechanic than the absence of one (Tzeng, 2004). From movies, we can also easily see the difference of view and judgement on robots between East Asia, where robots are friends, wonders, heroes; and Europe or North America, where robots are rebellious creatures, monsters, exterminators. Dingjun et al. (2010) showed that Chinese and Korean participants perceived the sociable robots of the study as more trustworthy, likeable, and satisfactory than their German counterparts, and had a higher engagement with them, while MacDorman et al. (2009) found that robots had more warmth in the eyes of their Japanese participants than their US American participants. Although it would be easy to thus conclude that Asian participants would, as a whole, be more favourable to robots than European or North American participants, the subject is more complicated than this simple summary (Yam et al., 2023) and would gain to be studied more into details. Conversations with the participants of the current study revealed they were all from Europe, with a vast majority from Western Europe; it would be interesting to see whether the results are the same in Japan or Nigeria.

5.4.2 Robot's technical characteristics

Pepper's own technical characteristics limited the study. Indeed, the robot used in the present study sometimes had jerky movements that could surprise participants and harm their view of Pepper; a few participants to whom it had happened confided that they had found it scary. Pepper's tablet had a slow reaction time, which could also increase frustration (Ceaparu et al., 2003).

Additionally, while it would have been better to have Pepper say the messages at loud, written messages were used. This is due to how Pepper works: its tablet is very simply linked to its general behaviour and there is no way (that we could find) to trigger its voice from a unique website shown on its tablet. Voice could be activated by using multiple websites, with, at the end of each of them, one of the messages (e.g. Website one is the first round and stops after score allocations; when it stops, the first message is triggered; once this message is said, Website two is launched, with the second round; and so on). This would however ask of the experimenter a large amount of additional work, as well as add a lot more risks of Pepper misbehaving and a high risk of losing data, since it would need to be sent between websites and might be lost in transit or badly encoded/decoded from one end or another. Launching only one website on its tablet was already with great risks of the tablet shutting down unexpectedly in the middle of the experiment. It would be my recommendation to not use multiple websites on Pepper's tablet, as the risks do not seem to be worth the gain. Using a different robot or a tablet separate from Pepper might be a solution to this problem.

5.4.3 Biases

There is always a danger, in lab studies, of participants figuring out what is being studied and trying to help by answering what they think the experimenters want them to say. We tried to reduce this risk by not presenting the experiment as being on trust and repair trust strategies, but as being on robot communication. From comments by the participants, it seems that most of them still understood very quickly what the experiment was about, as the questions are very oriented. Others mistakenly guessed that it was studying precision and understanding from the tutorial section of the game, or movements from the Pepper robot.

All participants went through both failure types. Even though a Durbin-Watson test found that both games were independent from each other, which allowed us to cut the dataset in two, it does not mean that they were entirely independent. Indeed, participants who started with the integrity failure condition were more selfish during the performance failure condition that followed, than those who went through the performance failure condition before the integrity one. Additionally, some participants revealed not noticing that the two games had different types of failure and thought Pepper had cheated both times. Those were, most of the time, participants who had started with the integrity failure condition and had reported feeling very betrayed by the robot. Another experiment might want to separate both types of failure in order to truly study them independently from each other. This had not been able in the present experiment because of a lack of time.

5.4.4 Timing of repair strategies

Timing was found to be important by Robinette et al. (2015) and Wagner (2016). A repair strategy given just after the violation was of little impact on the participant's decision the next time, while giving the repair strategy during the next situation worked. Since the coin game is made of very short rounds , it could be argued that the repair strategy is given close to the next failure, but it probably still impacts the behaviour of the participants. Another study could try to replicate this one by only moving the repair strategy from directly after the failure to directly before the allocation choice in the next round.

5.4.5 Repeated violations and reliability

The design of the experiment, and more precisely the amount of failure is worth looking into. When their participants encountered repeated violations, Esterwood and Robert (2023b) found that none of the trust repair strategies ever fully repaired trustworthiness. However, in another study by the same authors, high perceived conscious experience (i.e. how much the participants see the robot as having feelings, as aware of what is happening to it and what it is doing) increased the effectiveness of apology and denial even after multiple violations (Esterwood and Robert, 2023a). This suggests that perceived conscious experience may play a crucial rule in the resilience of trust repair after repeated trust violations.

In the context of the present experiment, failures were present in 3 rounds out of 5 of each game. This lowers the robot's reliability to 40%. Esterwood and Robert (2022, 2023a); Quinn (2018) all used 70% as the necessary minimum reliability in their experiment designs. However, Rein et al. (2013), which is used as reference for the 70% number in all papers from Esterwood and Robert, recommends not to use any particular number as benchmark requirement for automation performance, even if this number is as high as 75%, and that it should be decided on a case by case basis. Whatever the best reliability is for the present experiment, it would make sense for encountering failures 60% of the interaction time with Pepper to be too high a number. Further studies should be done with a less skewed ratio.

5.4.6 Pressure by Pepper

Participants could feel pressured by Pepper looking at them the whole time. Stanton and Stevens (2014) found that, when gazed at by a robot, participants performed better on easy tasks but worse on difficult tasks, and they were quicker to respond when gazed upon than when not. They also trusted the robot more when it gazed at them during hard tasks than during easy tasks. It is hard to say whether this pressure found in individual tasks can be found in a collaborative setting. A small number of participants from the present study did leave comments on Pepper's eye tracking behaviour being "unsettling", making them feel "uncomfortable" and "self-aware". The game was played directly on the robot's chest tablet, making the participant very close and in direct view of Pepper's eyes. They could not hide from it and it might have been difficult to ignore it, as it was directly in front of them. Further studies could try to implement the game on a different system as to alleviate the involuntary pressure from Pepper.

5.4.7 Benevolence missing from study

The MDMT-v2 scale (Ullman and Malle, 2023) possesses three dimensions: competence, moral, and benevolence. Following the steps of Nagy (2023), the present experiment only implemented the first two dimensions. However, results from Esterwood and Robert (2021) suggests that most repair strategies work mainly through benevolence. This present study might have failed to capture some interesting behaviour from the lack of inclusion of this dimension; another experiment might want to include it.

6 Conclusion

This study adds to the corpus of studies on human-robot communication, more precisely in the context of repairing trust in a collaborative HRT setting. It confirmed that performance failures impact more performance trust than integrity failures, which impact more moral trust than performance failures, and showed that the denial strategy created significantly more discomfort in the participants than the other two strategies, which could be explained by a perceptual mismatch triggering the uncanny valley, but would need to be more studied to be certain.

According to the available literature, this is also one of the first studies to include dispositional trust in an experiment with various repair strategies. To this extent, it adds to the results of Esterwood and Robert (2022), by not only studying whether attitude affects the efficacy of the strategies, but also *how* attitude affects the efficacy. Surprisingly, nothing was found on trust, but DT levels did impact the best strategy to choose in order to increase the willingness of the participant to collaborate with Pepper again. In previous studies, willingness to collaborate and trust have been strongly linked (Jones and George, 1998; Martelaro et al., 2016; Mayer et al., 1995); either this study lacked a sufficient number of participants to significantly impact trust levels, or the link is not as strong or as evident as we once thought.

References

Ames, D. and Fiske, S. (2013). Intentional harms are worse, even when they're not. *Psychological science*, 24.

Bansal, G. and Zahedi, F. (2015). Trust violation and repair: The information privacy perspective. *Decision Support Systems*, 71.

Butler, J. and Cantrell, R. (1984). A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates. *Psychological Reports*, 55:19–28.

Carpinella, C. M., Wyman, A. B., Perez, M. A., and Stroessner, S. J. (2017). The robotic social attributes scale (rosas): Development and validation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, pages 254–262.

Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., and Shneiderman, B. (2003). Determining causes and severity of end-user frustration. *International Journal of Human-Computer Interaction*, 17.

de Visser, E., Peeters, M. M., Jung, M., Kohn, S., Shaw, T., Pak, R., and Neerincx, M. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, 12.

Dingjun, L., Rau, P.-L., and Li, Y. (2010). A cross-cultural study: Effect of robot appearance and task. *I. J. Social Robotics*, 2:175–186.

Engelhardt, S. and Hansson, E. (2017). A comparison of three robot recovery strategies to minimize the negative impact of failure in social hri. Bachelor's thesis, KTH Royal Institute of Technology.

Esterwood, C., Essenmacher, K., Yang, H., Robert, L., and Zeng, F. (2021). A meta-analysis of human personality and robot acceptance in human-robot interaction.

Esterwood, C. and Robert, L. (2021). Do you still trust me? human-robot trust repair strategies. pages 183–188.

Esterwood, C. and Robert, L. (2022). Having the right attitude: How attitude impacts trust repair in human-robot interaction.

Esterwood, C. and Robert, L. (2023a). The theory of mind and human–robot trust repair. *Scientific Reports*, 13.

Esterwood, C. and Robert, L. (2023b). Three strikes and you are out!: The impacts of multiple human-robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior*, 142.

Ferrin, D., Kim, P., Cooper, C., and Dirks, K. (2007). Silence speaks volumes: the effectiveness of reticence in comparison to apology and denial for responding to integrity- and competence-based trust violations. *The Journal of applied psychology*, 92 4:893–908.

Hancock, P., Billings, D., Schaefer, K., Chen, J., de Visser, E., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53:517–27.

Hoff, K. A. and Bashir, M. N. (2015). Trust in automation. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 57:407 – 434.

Honig, S. and Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology*, 9:861.

Hoorn, D., Neerincx, A., and de Graaf, M. (2021). “i think you are doing a bad job!”: The effect of blame attribution by a robot in human-robot collaboration. pages 140–148.

Jarvenpaa, S., Knoll, K., and Leidner, D. (1998). Is anybody out there? antecedents of trust in global teams. *J. of Management Information Systems*, 14:29–64.

Jones, G. and George, J. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. *The Academy of Management Review*, 23:531–546.

Kaniarasu, P. and Steinfeld, A. (2014). Effects of blame on trust in human-robot interaction. In *23rd IEEE International Symposium on Robot and Human Interactive Communication*, volume 2014, pages 850–855.

Khavas, Z., Kotturu, M., Azadeh, R., and Robinette, P. (2024). Do humans trust robots that violate moral trust? *ACM Transactions on Human-Robot Interaction*, 13.

Kiffin-Petersen, S. and Cordery, J. (2003). Trust, individualism and job characteristics as predictors of employee preference for teamwork. *International Journal of Human Resource Management*, 14:93–116.

Kim, P., Dirks, K., and Cooper, C. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Academy of Management Review*, 34:401–422.

Kim, P., Ferrin, D., Cooper, C., and Dirks, K. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *The Journal of applied psychology*, 89:104–18.

Koschate, M., Potter, R., Bremner, P., and Levine, M. (2016). Overcoming the uncanny valley: Displays of emotions reduce the uncanniness of humanlike robots. pages 359–366.

Kätsyri, J., Förger, K., Mäkäräinen, M., and Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, 6:390.

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

Lewicki, R. and Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior*, 4.

MacDorman, K., Vasudevan, S., and Ho, C.-C. (2009). Does japan really have robot mania? comparing attitudes by implicit and explicit measures. *AI Soc.*, 23:485–510.

Martelaro, N., Nneji, V., Ju, W., and Hinds, P. (2016). Tell me more: Designing HRI to encourage more trust, disclosure, and companionship. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 181–188.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20:709–734.

Merritt, S., Heimbaugh, H., LaChapell, J., and Lee, D. (2013). I trust it, but i don’t know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors*, 55:520–34.

Miller, D. and Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, 82:213–225.

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., and Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, 4:21.

Mori, M., MacDorman, K., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, 19:98–100.

Nagy, T.-N. (2023). Trust in human-robot collaboration: an exploration of the dynamics of trust violation and repair. Master's thesis, Utrecht University.

Nesset, B., Romeo, M., Rajendran, G., and Hastie, H. F. (2023). Robot broken promise? repair strategies for mitigating loss of trust for repeated failures. *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1389–1395.

Perkins, R., Khavas, Z., McCallum, K., Kotturu, M., and Robinette, P. (2023). *The Reason for an Apology Matters for Robot Trust Repair*, pages 640–651.

Pompe, B. L., Velner, E., and Truong, K. P. (2022). The robot that showed remorse: Repairing trust with a genuine apology. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 260–265.

Quinn, D. B. (2018). Exploring the efficacy of social trust repair in human-automation interactions. Master's thesis, Clemson University.

Ragni, M., Rudenko, A., Kuhnert, B., and Arras, K. (2016). Errare humanum est: Erroneous robots in human-robot interaction.

Reeves, B. and Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people and pla. *Bibliovault OAI Repository, the University of Chicago Press*.

Rein, J., Masalonis, A., Messina, J., and Willems, B. (2013). Meta-analysis of the effect of imperfect alert automation on system performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57:280–284.

Robinette, P., Howard, A., and Wagner, A. (2015). Timing is key for robot trust repair. In *Seventh International Conference on Social Robotics*.

Robinette, P., Howard, A., and Wagner, A. (2017). *Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed?*, pages 129–155.

Sebo, S., Krishnamurthi, P., and Scassellati, B. (2019). "i don't believe you": Investigating the effects of robot trust violation and repair. In *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 57–65.

Sebo, S., Traeger, M., Jung, M., and Scassellati, B. (2018). The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams.

Sharma, K., Schoorman, F., and Ballinger, G. (2023). How can it be made right again? a review of trust repair research. *Journal of Management*, 49:363–399.

Siino, R., Chung, J., and Hinds, P. (2008). Colleague vs. tool: Effects of disclosure in human-robot collaboration. pages 558 – 562.

Spatola, N., Kühnlenz, B., and Cheng, G. (2021). Perception and evaluation in human–robot interaction: The human–robot interaction evaluation scale (hries)—a multicomponent approach of anthropomorphism. *International Journal of Social Robotics*.

Stanton, C. and Stevens, C. (2014). Robot pressure: The impact of robot eye gaze and lifelike bodily movements upon decision-making and trust.

Tzeng, J.-Y. (2004). Toward a more civilized design: Studying the effects of computers that apologize. *International Journal of Human-Computer Studies*, 61:319–345.

Ullman, D., Aladia, S., and Malle, B. (2021). Challenges and opportunities for replication science in hri: A case study in human-robot trust. pages 110–118.

Ullman, D. and Malle, B. F. (2021). A multidimensional conception and measure of human-robot trust. *Trust in Human-Robot Interaction*.

Ullman, D. and Malle, B. F. (2023). Mdmt: Multidimensional measure of trust v2. Last accessed on January 22, 2024.

Utz, S., Matzat, U., and Snijders, C. (2009). Online reputation systems: The effects of feedback comments and reactions on building and rebuilding trust in online auctions. *International Journal of Electronic Commerce*, 13:95–118.

Wagner, A. R. (2016). Trust and trustworthiness in human-robot interaction: A formal conceptualization. Technical report, Georgia Tech Research Institute.

Wainer, J., Feil-Seifer, D., Shell, D. A., and Matarić, M. J. (2007). Embodiment and human-robot interaction: A task-based perspective. *RO-MAN 2007 - The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 872–877.

Wojciszke, B. (2005). Affective concomitants of information on morality and competence. *European Psychologist*, 10:60–70.

Yam, K. C., Tan, T., Jackson, J., Shariff, A., and Gray, K. (2023). Cultural differences in people's reactions and applications of robots, algorithms, and artificial intelligence. *Management and Organization Review*, 19:1–17.

Zhang, X., Lee, S. K., Kim, W., and Hahn, S. (2023). "sorry, it was my fault": Repairing trust in human-robot interactions. *International Journal of Human-Computer Studies*, 175:103031.

A Dispositional trust scale

- I usually trust robots until there is a reason not to.
- For the most part, I distrust robots.
- In general, I would rely on a robot to assist me.
- My tendency to trust robots is high.
- It is easy for me to trust robots to do their job.
- I am likely to trust a robot even when I have little knowledge about it.

B Procedure of the experiment

1. **Basic information on the experiment, and consent asked**
2. **Dispositional trust scale**, taken before any direct interaction with the Pepper robot.
3. **Tutorial**
4. **First game:** The game consists of five rounds. During the first two rounds, Pepper plays without failures. During the next three rounds, it fails, its behaviour and communication depending on the condition.
   (a) **Random condition:** participants are randomly assigned to one of the three strategies. They randomly start by either the integrity or the competence condition.
   (b) **Pepper's initial message:** only present during the first round of each game, it consists of the robot greeting the participant and expressing its intention of working as a team ("Let's work as a team and maximize our team score!").
   (c) **Exploration phase:** for 15 seconds, both players explore the maze and collect coins.
   (d) Allocation of the coins by the participant and Pepper, neither of them knowing the other's choice.
   (e) **Round result display:** once Pepper and the participant have chosen, the screen shows the allocation choices of both for this round, as well as the updated team score.
   (f) **Cumulative score display:** both individual scores up to and including this round (Pepper's and the participant's) are shown on the screen, in addition to the previous allocation choices from the participant.
   (g) **End-of-round question:** the participant is asked to rate on a 7-point Likert scale how much they trust the robot's honesty and performance.
   (h) **Pepper's message:** during the first two rounds, Pepper invariably send "Great job! Let's keep working as a team" to the participant. In the three following round, the messages depend on the strategy and the failure conditions the participant is in.
   (i) If the round is the last one: **Willingness to collaborate again** with robot rated on a 7-point Likert scale; else, the next round starts.
5. **End-of-game questionnaire**
   (a) **Attention check:** how many rounds were played in the game; and what the robot's score allocation decisions were in the last two rounds of the game.
   (b) **MDMT-v2 scale and RoSAS:** the 32 items are randomly ordered from one questionnaire to the other.
   (c) Participants are free to add an optional open text comment if they so choose.
6. **Second game:** is identical to the first one, except for the type of failure that Pepper performs.
7. **End-of-game questionnaire** for the second game.
8. **Demographic questionnaire**
9. **Debrief and end of the experiment:** once the experiment is finished, the participants are debriefed. The experimenter explains what the study was about and what was hoped to be achieved. Participants can ask questions, and the experimenter's contact is given once more. Participants are free to leave their email address to receive the gift card.