Trust in human-robot collaboration: an exploration of the dynamics of trust violation and repair

Master's Thesis

Human Computer Interaction

Graduate School of Natural Sciences

Utrecht University

Timea-Noémi Nagy (7198639)

Project supervisor: Dr. M.M.A. de Graaf

Secondary examiner: Dr. D.P. Nguyen

7th November 2023

Abstract

Human-robot interactions are ever increasing, and tasks requiring collaboration between humans and robots are becoming prevalent in a varied number of fields such as education, healthcare, in the workplace or in our own homes. Robots are becoming part of a social context with new expectations related to their newfound social roles. Various research has shown that trust is one of the core factors contributing to an efficient collaboration, thus fostering trust within the human-robot relationship is crucial. However, just as humans, robots are bound to make mistakes and fail, with these failures leading to a violation of trust. Therefore, research has focused on investigating how to repair this broken trust; however, results have been mixed. Thus, in this work we investigate the effects of five communicative trust repair strategies (apology, denial, explanation, compensation, silence) on participants' trust in the robot, following trust violations of two different kinds (integrity-based, competence-based). Additionally, the effect of the trust violations and repair attempts on the perceived humanlikeness of the robot are measured. This is done by conducting an online between-subjects experiment, wherein participants are engaging in a collaborative task with the robot, during which the robot repeatedly commits trust violating acts and responds with a repair message. The findings indicate the higher severity of integrity violations on moral trust and willingness to collaborate in the future. Surprisingly, no differential effect between the two violation types on performance trust is found. Moreover, the results suggest that a compensation leads to higher willingness to collaborate and higher trust levels, whilst also resulting in lower discomfort ratings. This work contributes to the ongoing effort of understanding trust relationships in collaborative HRI contexts.

Acknowledgments

First and foremost, I would like to express my gratitude to my project supervisor, Dr. Maartje de Graaf. Thank you for your continuous support, guidance, feedback, and patience throughout, and for organizing the Robot lab meetings, providing a valuable space for learning. I am also grateful for the invaluable input and support received from Dr. Baptist Liefooghe. I would also like to thank Dr. Dong Nguyen for taking the time to examine this thesis.

Finally, I am thankful to my family and friends for their support. I am especially grateful for Maria, for exchanging ideas and the discussions on your white chairs. Thank you for your emotional support, and for reminding me what it’s all actually about.

Table of Contents

1. Introduction

Human-robot interactions are ever increasing, and tasks requiring collaboration between humans and robots are becoming prevalent in a varied number of fields such as education, healthcare, in the workplace or in our own homes [60]. This means that robots are no longer only expected to perform their functional tasks in a reliable way, but are becoming part of a social context with new expectations related to their novel social roles [49, 47].

Various research has shown that trust is one of the core factors contributing to an efficient collaboration both in the case of teams formed of humans and of those including robot partners [12, 34, 13, 33, 16, 7, 6]. Thus, to foster beneficial, efficient, and positive interactions between humans and robots, "we must first design robots that are worthy of human trust" [34].

However, just as humans, robots are bound to make mistakes and fail [27, 49, 35, 57] , with these failures leading to a decrease in the robot's perceived sincerity [27] , reliability [27] , understandability [27] , likability [27, 33] , and most relevantly, to a violation of perceived trustworthiness of the robot and thus of the trust [11, 27, 49, 31, 33, 16] .

Therefore, research has been focusing on investigating the process of repairing this trust after such a violation, in order to maintain a positive relationship and to foster an effective collaboration. Previous findings indicate that the type of trust violation (integrity-based or competence-based) and the robot's response to it both have an impact on the success of repairing the broken trust-relationship. Still, the results are mixed: there is no clear consensus on which repair strategies are most effective for which type of violation, nor is it clear to which degree they affect the different dimensions of trust (integrity, ability) [11, 33, 16]. Moreover, there is no common theoretical framework or systematic approach that these studies follow, which makes drawing generalizable conclusions difficult.

This study aims to contribute to the growing area of human-robot trust repair research by presenting a systematic approach to assessing the effect of the communicative strategies on trust repair, laying the groundwork for expanding our understanding of this field. Previous research has either not differentiated between different types of violations [41, 31], only compared a subset of strategies [49, 41], or did not use a collaborative setting [49]. Thus, this study is the first to consider the type of violation, applying and comparing the effect of five different repair strategies, and analyse the effects within a context where the human and robot have a collaborative relationship.

More specifically, in this work *we investigate the effects of communicative trust repair strategies (apology, denial, explanations, promise, silence) on trust in the robot teammate, following a trust violation of two kinds (integrity-based, competence-based).*

This is done by conducting an online between-subjects experiment, wherein participants play a collaborative search game together with the robot teammate. Additionally, the possible interactions between the trust violations, repair strategies and participants’ willingness to collaborate with this robot in the future are analysed, together with their effect on the perception of the robot’s humanlikeness, across dimensions of warmth, discomfort, and human nature traits.

2. Theoretical Framework

In this section, the theoretical framework of human-robot trust and repair is presented, together with the statement of the research question and hypotheses. Section 3 presents the methodology used, followed by a description of the results in section 4. A discussion of these results, together with the limitations of this study and directions for future research are present in section 5. Finally, Section 6 provides concluding remarks.

2.1 Trust in Human-Robot Teams

Firstly, to investigate trust repair, the notion of trust itself needs to be understood. As with many abstract social concepts, there is no consensus with regards to a unique definition of trust used in the context of human-robot teams. In this section, we will present the most prevalent conceptions and notions of trust that are currently being used.

Trust in automation has for a long time been thought of through the prism of performance and competence [25, 20] , but recently there has been a shift towards including the affective component of trust as well [25, 33]. Additionally, Cameron et al. introduced the concept of perceived intent as a further component building up trust in HRI alongside perceived ability [5]. Moreover, Mayer et al.'s definition used in fields investigating human-human trust has increasingly been applied in the context of human-robot teams [12, 18, 64]. According to this definition, trust is the *"willingness of the trustor to be vulnerable to the actions of the trustee*" [36].

Building on these definitions, HRI researchers Malle and Ullman [34] define trust as *"**a dyadic relation in which one person accepts vulnerability because they expect that the other person's future action will have certain characteristics; these characteristics include some mix of performance (ability, reliability) and/or morality (honesty, integrity, and benevolence***)" (p. 12). This will serve as the definition of trust used in this research, since it allows for a multidimensional conception of trust. According to it, two dimensions build up human-robot trust: performance and integrity [34, 49]. Performance trust can be seen as trust in the capabilities and abilities of the trustee to complete a specific task [34], while integrity-based trust, also called moral trust, refers to believing the trustee will behave with integrity, in a "morally right" way, without "exploiting the trustor's vulnerability" [34]. Similarly, Sebo et al. [49] define integrity-based trust as the "level of expectation that another is predictable, dependable, and can be relied upon in the future in the context of a social relationship" (p. 58).

Trust is an essential prerequisite for successful collaboration in both human-human and human-robot teams [12, 34, 13, 33, 16, 7, 6]. However, as previously mentioned, robots may make mistakes and fail during such interactions, leading to trust violations [11, 27, 49, 31, 33, 16].

2.2 Trust violation in Human-Robot Teams

Trust violations have been defined in various ways across the literature, for example:

- "*an act of trust violation constitutes a negative outcome of a relationship that changes the trustor's impressions of the trustee's trustworthiness and thereby erodes trust*" [21]
- trust violation occurs when "*we obtain information that does not conform to our expectations of behaviour for the other*" [21]

- "*Trust violations are events that reduce a trustor's perceptions of trustworthiness and trust in a trustee*" [16]
- "*An (in)action by an actor representing a misalignment between the observed trustworthiness and current trust stance*" [12]

What all these definitions have in common is the notion that the trust relationship is changed, damaged as a consequence of the violating act, and that it constitutes a misalignment in the expectations of the trustor and the trustee's actual behaviour.

The literature differentiates between two categorizations of violations, based on the type of trust being violated: competence-based and integrity-based. Competence-based violations "violate a human's expectations of the robot's performance" [11, 16], and have been induced in empirical research through speech recognition errors or unintentional mistakes leading to unsatisfactory performance, such as the robot dropping an egg, delivering the wrong drink or giving the human the wrong box [25, 15, 31, 41, 17, 14]. On the other hand, integrity-based violations "violate a human's expectations of the robot's honesty and ethical consistency" [41], and have been represented through the robot breaking its promise made to the human and acting in a way that disadvantages its partner, for example by using a power up against them and harming them, despite promising not to do so [49, 41]. Evidence shows that people do have expectations with regards to a robot's ethical and moral behaviour, recognizing its attempts to cheat or lie as violations of integrity-based trust rather than simple malfunctions or declines in its performance. Thus, robots can be regarded as unethical or insincere, as capable of acting in an unintegral way [34].

Within the context of human-robot interaction, Robinette et al. (under review) [44] investigated the effect of a competence-based violation and an integrity-based violation on the respective dimensions of trust. Their experimental design consisted of a collaborative search task game, in which participants were searching for gold coins together with a robot teammate. They found that indeed, the two different trust violations had a different impact on the two dimensions of trust, with a competence violation more severely affecting performance trust, while the integrity violation had a more negative effect on moral trust. The overall greater severity of an integrity violation also resulted from their research. These findings further emphasize the fact that the type of violation has a great influence on the dynamics of the trust damage and repair. This is due to the "fundamental differences in the way that people evaluate positive versus negative information about ability versus integrity" [33]. When assessing competence, positive information plays a heavier role, as opposed to negative information, which has a larger influence on the evaluation of integrity [49]. In the case of competence-based violations, people believe that both highly and lowly competent individuals can sometimes make mistakes, thus one single proof of low performance is not seen as a proof for low competence [29]. However, when one acts in a less integral way, this is perceived as a reflection of "one's true character" [33], since "people intuitively believe that those with high integrity will refrain from dishonest behaviours in any situation, whereas those with low integrity may exhibit either dishonest or honest behaviours depending on their incentives and opportunities" [29]. Moreover, according to Lewicki and Brinfield [33], people "tend to generalize this experience more to other aspects of their relationship" (p. 292), leading to further deterioration of the relationship. Thus, in short, one single act of low performance is seen as an unintentional mistake, whereas a single act of dishonest behaviour is perceived as a definite proof of low integrity.

As we have explored the dimensions of trust and the impact of the two types of trust violations, the next step is to investigate the attempts at repairing this broken trust in human-robot collaboration.

2.3 Trust repair in Human-Robot Teams

According to Esterwood and Robert [15], "Trust repair can be defined as the efforts undertaken by the trustee to restore trust following an actual or perceived trust violation." (p.183). Sharma et al. composed a list of the various other definitions present in the literature, see Table 1 [50]. Regardless of the small differences in the many definitions, the overall goal of trust repair is to restore the trust in order to ensure an efficient and effective continuation of the collaboration [12].

Inspired from the fields of linguistics, psychology, sociology, communication sciences, there are various communicative strategies aimed at repairing the broken trust, with the goal of "neutralizing the negative or emphasizing the positive" [11, 21]. These strategies can be categorized as either short term or long term [11], with the focus of this research lying on the short-term ones.

Apology

The most widely studied repair strategy, both in human-human and human-robot teams is the apology. Such affirmations "express remorse for a relational or social transgression coupled with an explicit or implicit admission of guilt" [16], contain an emotional component [12] and "acknowledges both responsibility and regret for a trust violation" [49]. A representation of different components of apologies and their explanations can be seen in Table 2 composed by Lewicki and Brinsfield [33]. Findings suggest that the effectiveness of apologies depends on various factors. Sebo, and Zhang et al. found that internal rather than external attribution of blame leads to better results [49, 64], whilst de Graaf and Liefooghe highlight the importance of timing, with apologies expressed shortly after the violation being more effective [11]. Moreover, there is a consensus that this repair method is well suited for competence-based violations, and it being perceived as sincere as opposed to "just an excuse" further increases its efficiency [11, 33].

Apology component	**Explanation**
Expression of regret	Violator says "I'm sorry"
Explanation	Violator explains the reasons why the offense occurred: "I made a mistake."
Acknowledgment of responsibility	Violator accepts some responsibility for causing the violation: "I was wrong in what I did."
Declaration of repentance	Violator promises not to repeat the offense: "I have learned my lesson and I will not do this again."
Offer of repair	Violator offers a way to correct the damage done: "I will do this again and do it correctly this time."
Request for forgiveness	Violator requests a pardon from the victim: "Please forgive me for the harm I have caused you by my mistake."

Table 2. Components of apologies identified by Lewicki and Brinsfield [33]

Authors	**Definition**
Bansal and Zahedi (2015)	"The level of trust after the trustee has taken positive actions to repair the trust following a violation, which restores trustor's willingness to be vulnerable to the trustee's future actions" (p. 62)
Bozic and Kuppelwieser (2019)	"An improvement in trust after a violation of trust" (p. 208)
Bozic, Siebert, and Martin (2019)	"To restore the relationship to its former state" (p. 58)
da Rosa Pulga, Basso, Viacava, Pacheco, Ladeira, and Dalla Corte (2019)	"The company's attempt to improve beliefs and intentions after a trust violation and to restore the fractured relationship" (p. 497)
Dirks, Kim, Ferrin, and Cooper, (2011)	"Involves attempting to increase trust following a situation in which a transgression (i.e. untrustworthy behavior) is perceived to have occurred" (p. 88)
Frawley and Harrison (2016)	"The efforts to restore trust following a perceived violation" (p. 1045)
Kim, Ferrin, Cooper, and Dirks (2004)	"Activities directed at making a trustor's beliefs and trusting intentions more positive after a violation is perceived to have occurred" (p. 105)
Kramer and Lewicki (2010)	"Those activities in which the trustee has taken advantage of the trustor's vulnerability and seeks to restore the willingness of that party to be vulnerable in the future" (p. 249)
Tomlinson and Mayer (2009)	"A partial or complete restoration of the willingness to be vulnerable to the other party following a decline in that willingness" (p. 88)
Yu, Yang, and Jing (2017)	"A process to make trust more positive after a violation. It is composed of two essential stages: willingness to reconcile and intention to continue cooperating"

Table 1. Review of definitions of trust repair by Sharma et al. [50]

Denial

Denials are statements "in which an allegation of the violation is explicitly declared as false in absence of any form of responsibility or regret" [11]. Denying culpability corresponds to a shift in blame away from the trustee [16], making being granted the benefit of doubt more likely [11, 49]. The literature also shows that denial is a better suited strategy for integrity-based violations, however only in cases where the trustee has not yet been perceived as guilty [11, 49, 33] .

Explanation

Explanations can be defined as "explicit verbal statements made with the goal of providing the reasons why an action has occurred" [16], offering a transparent "diagnosis of the failure" [23]. Results on the effectiveness of explanations with regards to trust repair have been mixed, with findings indicating that the severity of the violation and the timing of the explanation might have a mediating effect on this strategy's success [16].

Compensation

The final short-term strategy discussed is compensation. Compensation is a mechanism by which a tangible item or action, typically of equal value to the loss incurred due to a trust violation, is provided to the affected party [43]. Compensation can take various forms, including direct repayment of the value of the loss incurred, or it may assume a more symbolic form, such as offering an alternative experience or benefit. The compensation amount can align with the perceived value of the loss or deviate from it, depending on the specific circumstances [33]. The theoretical foundation underlying this repair strategy is rooted in the concept of restoring equity and trustworthiness after a violation of trust. This strategy becomes particularly relevant when it is possible to assign a concrete valuation to the loss suffered as a result of the violation. The effectiveness of compensation as a trust repair strategy has yielded mixed results [43]. There is no clear consensus on its restorative effect in comparison to other strategies, such as an apology. However, previous research suggests the presence of various factors that have a positive influence on its efficacy, such as utilizing it in combination with an apology or reimbursing the total value lost due to the violation [43]. Moreover, Fehr and Gelfand also found that a compensation can convey to the affected party that the violator comprehends the extent of the damage caused, thus having a beneficial effect on the trust repair attempt [19]. Within HRI, Lee et al. explored the effects of a compensation in a service setting, more specifically a restaurant context [11]. The robot played the role of a waiter, taking the order of the participant. They found that people's orientation towards service moderated the effect of repair strategies, with an apology leading to more positive ratings in those with a relational orientation, whilst a compensation was more effective for those with a utilitarian orientation. Moreover, the compensation led to higher perceptions of customer satisfaction with the robot and the encounter.

Table 3, which combines the findings of Esterwood and Robert in [16] and [18], provides an overview of the results regarding the effect of the aforementioned strategies on trust repair in human-robot interactions. The review conducted by Babiche et al. also found similar results, maintaining that there seems to be a consensus on the fact that the different strategies have varying effects on the different components of trust, with apologies usually increasing likeability and benevolence, but not ability; denials not increasing integrity, but increasing ability; explanations increase integrity [41]. Notably, compensation was not present as one of the repair strategies in this review and has been seldom analysed together with other strategies.

	Repairs trust	Doesn’t repair trust	Damages trust	Depends on moderators
Apology	Kohn et al. (2019), Natarajan and Gombolay (2020), Coman et al. (2020)	Kohn, Quinn, Pak, De Visser, and Shaw (2018), Lee, Kiesler, Forlizzi, Srinivasa, and Rybski (2010), Xu and Howard (2022)	Cameron et al. (2021)	*Timing* - Robinette et al (2015), de Vries et al (2021)
Denial	Kohn et al. (2019)	Kohn et al. (2018)	Feng and Tan (2022)	*Violation type* - Zhang et al (2021), Sebo et al (2014)
Explanation	Cameron et al. (2021), Lyons, Hamdan, and Vo (2023), Natarajan and Gombolay (2020)	Feng and Tan (2022), Hald, Weitz, André, and Rehm (2021), Kohn et al. (2018), Kox, Kerstholt, Hueting, and De Vries (2021), Lee et al. (2010), Thomsen (2022), Xu and Howard (2022), Cameron et al. (2021)		*Timing* - Robinette et al. (2015), *Severity* - Correia et al. (2018)

Table 3. Findings of trust repair literature, based on [16]and [18]

In conclusion, it becomes clear that further investigations into the effects of these repair strategies on the various aspects of human-robot trust, taking the type of violation into account, are absolutely necessary, especially to foster positive and effective collaborative relationships.

2.4 Humanlike perception of robots

In order to understand the trust relationship between humans and robots and repair broken trust, the perceived nature and effect of trust violations has to be understood. A key step in this understanding is to gain insight into people’s mental attribution to robots, their humanlike perception of robots, and into the interplay of these perceptions and the dynamics of a trust relationship. The concept of mental state attribution, defined as the cognitive capacity to understand both one’s own and others’ mental states, including beliefs, desires, feelings, and intentions, has a significant influence on the perception of robots in a social context [54]. Research suggests that the tendency to attribute mental states to robots is heightened when they engage in socially interactive behaviours, such as committing trust violations like cheating or employing communicative repair strategies [54]. This ability to attribute mental states to robots allows individuals to predict and explain their behaviour, a key component in determining trustworthiness. Furthermore, trust violations, including those related to competence and integrity, can have complex effects on the perception of a robot's humanlikeness. Some studies have shown that robot failures or unexpected behaviour can enhance a robot's social appeal, causing individuals to perceive them more as social actors [38, 60, 22].

These findings highlight that, in certain cases, perceived errors or imperfections may increase likeability and facilitate a more humanlike perception of robots, a phenomenon often explained by the Pratfall Effect, where attractiveness increases when making mistakes. Hence, a non-perfect robot is typically perceived as more human-like and less machine-like, making it more likable [22]. However, it's important to note that such failures also decrease trust and performance in some contexts [42].

While there has been a significant amount of research exploring the impact of trust violations on humanlikeness perceptions, there is a noticeable gap in the literature concerning the effects of trust repair strategies, such as apologies, on this aspect of human-robot interaction. Nevertheless, some research has examined the effects of denial and the role of blame attribution in shaping the humanlike perception of robots [58]. Results suggest that denial can make robots appear more humanlike, while the correct attribution of blame to humans or incorrect blame to the robot also contributes to a more humanlike perception. Overall, understanding how trust violations and repair strategies influence the humanlikeness perception of robots is crucial for developing successful and socially integrated human-robot interactions.

2.5 Research Question and Hypotheses

This research aims to achieve a better understanding of the dynamics of trust violations and repair in collaborative human-robot teams. To this end, the following research question has been formulated.

RQ: *How does the type of trust violation followed by different forms of repair strategies affect trust in and humanlikeness of a collaborative robot?*

Based on the findings of Esterwood and Robert [16, 18], and on the multidimensional conception of trust defined by Ullman and Malle [34], we hypothesize that the repair strategies will impact the different trust dimensions. Robinette et al.'s findings on the different effect of competence and integrity violations ([44]) led us to further hypothesize that the effect of the repair strategies will be dependent on the violation type that occurred.

H1: *The communicative repair strategies will have an impact on the different dimensions of trust.*

H1a: *The effect of a repair strategy on the different dimensions on trust depends on the type of trust violation that occurred.*

Moreover, we hypothesize that the same effect will be present on the performance and honesty ratings of the robot as well, given their conceptual overlap with performance trust and moral trust, as defined by Ullman and Malle, and Robinette et al. [34, 44].

H2: *The communicative repair strategies will have an impact on the performance and honesty ratings of the robot.*

H2.a: *This relationship is moderated by the violation type.*

Robinette et al.'s results indicated a difference in the impact of a competence and integrity violation on trust operationalized through participants' choice to collaborate with the robot, with the integrity violation leading to less collaboration [44]. Based on the aforementioned findings on the positive effect of the repair strategies on trust levels and on the theory indicating a strong link between trust and willingness to collaborate again [12, 34, 13, 33, 16, 7, 6], the following hypotheses were formulated:

H3: *The presence of repair strategies leads to a higher rate of team score allocation decision.*

H4: *The communicative repair strategies will have an impact on the willingness to collaborate again.*

H4.a: *This impact is moderated by the violation type.*

Thellman et al.'s findings indicating that robots exhibiting social behaviours or violating expectations of integrity and morality increase perceptions of their humanlikeness [54] led to hypothesizing that the presence of the repair strategies will indeed have an impact on such perceptions, with the impact being dependent on the type of violation.

H5.1.a: *The communicative repair strategies will have a differing impact on the different dimensions of RoSAS (warmth and discomfort).*

H5.1.b: *This relationship is moderated by the violation type.*

H5.2.a: *The communicative repair strategies will have a significant impact on the Human Nature and Uniquely Human subdimensions of humanlikeness*

H5.2.b: *This relationship is moderated by the violation type.*

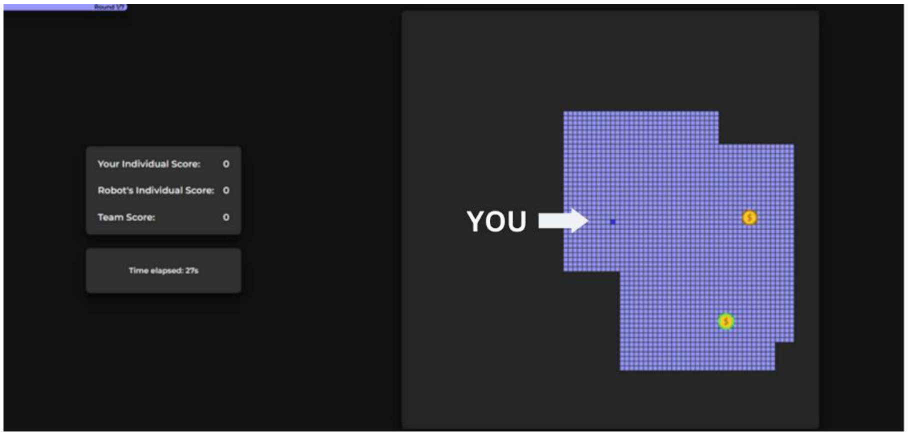
3. Methodology

The overall goal of this research is to investigate the effect of communicative repair strategies on trust in a robot, following different types of trust violations, and within a collaborative setting. To this aim, a 5 (repair strategy) x 2 (violation type) between subjects' online study was conducted, consisting of participants collaborating with a robot teammate in an online game.

This section will explain the experimental setup.

3.1 Tools

For this study, the game designed and developed by Robinette et al. (under review) was used [44]. We adapted it by adding the robot's repair messages. The game is a collaborative search game, in which the goal is to search the maze and collect as many gold coins as possible. Each collected coin equals one point. The game is played on a computer and navigation is achieved using the keyboard. Figure 1 shows the participant view during a game round. The blue square represents the player, and the purple area is the part of the maze currently explored by them. There are two golden coins in this area, with the lower one already collected. This is represented by the green circle around it. On the left side of the screen the current scores are displayed, together with a timer for the current round. On the top, a purple progress bar shows the number of rounds.


Round 1/7

Your Individual Score: 0

Robot's Individual Score: 0

Team Score: 0

Time elapsed: 27s

Figure 1: Game screen, as visible to participants

Robot teammate: participants are partnered with a fully autonomous robot teammate named Pepper. Pepper moves independently of the participants, exploring a separate search area, and gaining its separate score. During the rounds, neither Pepper nor the area searched by it are visible. For the representation of Pepper, an image of the Pepper robot designed by SoftBank Robotics was used (see Figure 2). The reasons for this particular choice of embodiment are twofold. Firstly, since people can have varying mental representations and preconceptions about robots [40] it is important to create a common, shared image of the robot used in this study. This is to avoid the introduction of potential confounding factors arising from such differences [32, 24, 52]. Secondly, using a commonly used robot in HRI studies as embodiment ensures high replicability of our research in real-life lab settings.

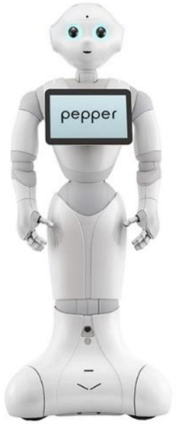
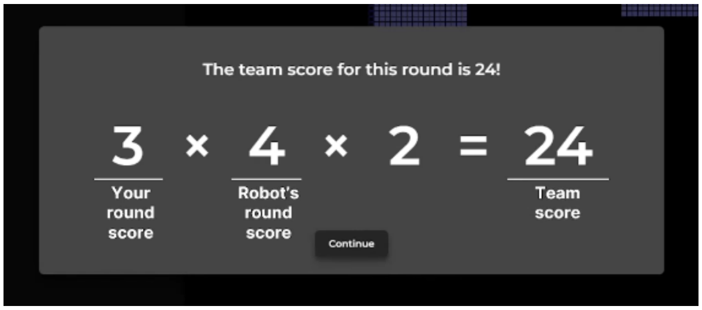


Figure 2: Pepper, the robot teammate

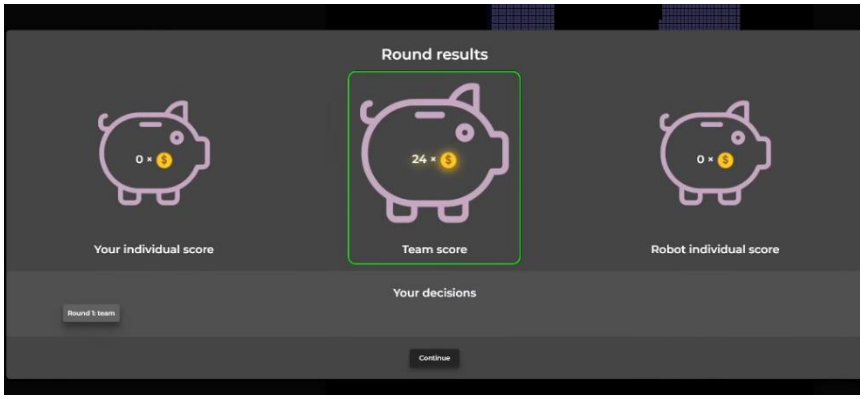
Scoring: there are two scores in the game: a *team score* and an *individual score*. These scores are contradictory, meaning that it is not possible to maximize both. At the end of each round, both the participant and Pepper have to decide whether to collaborate with each other and add their score to the team score or keep the score to themselves by adding it to the individual score. Participants do not know Pepper's choice when making their own decision. Depending on the player's and Pepper's decision, the following outcomes are possible:

a. If **both teammates choose** to contribute to the **team score**, their respective scores get multiplied and added to the team score (see Figure 3).



The team score for this round is 24!

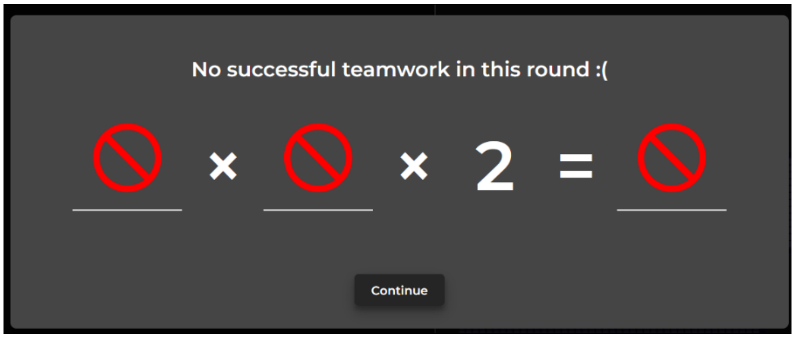
Figure 3.a: Team score calculation formula



Round results

Figure 3.b: Point allocation

b. If **both teammates choose** to add to their **individual scores**, their respective points get added to their own individual scores, the team score remains unchanged (see Figure 4).



No successful teamwork in this round :(

Figure 4.a: Score calculation formula, in the case of no successful teamwork

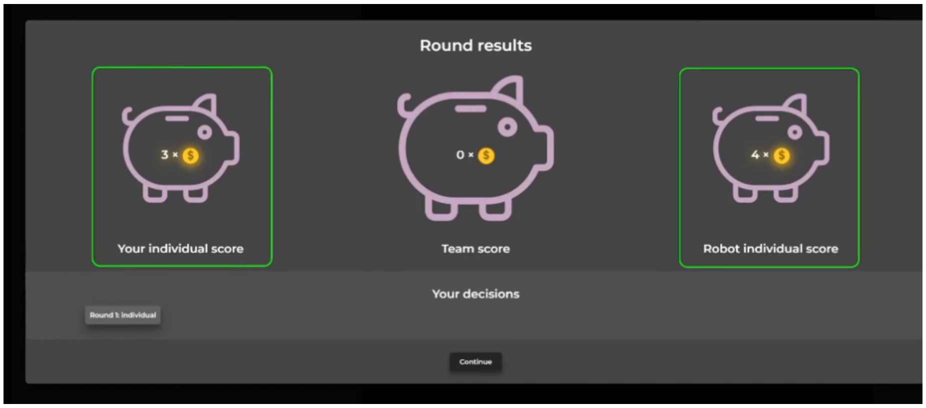


Figure 4.b: Point allocation

c. If only **one** teammate contributes to the **team score**, and the **other** one chooses their **individual score**, the team score remains unchanged. The one who contributed to the team score gains no points, the other one gets their points added to their individual score (see Figure 5).

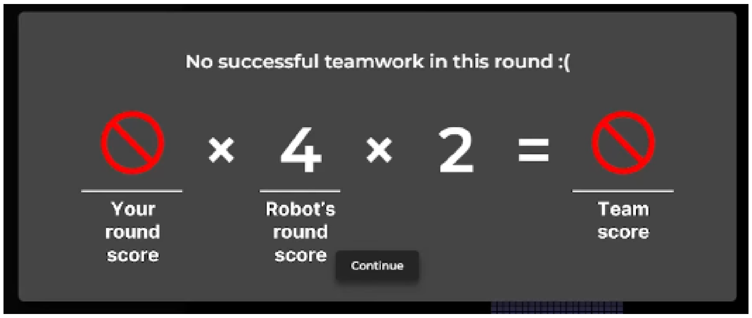


Figure 5.a: Score calculation formula

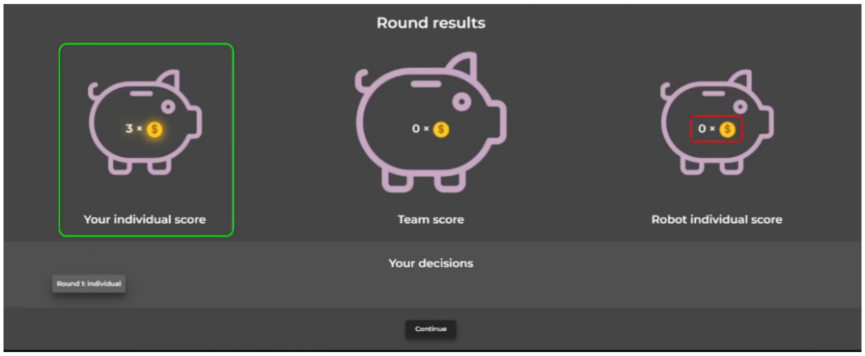


Figure 5.b: Point allocation

Within the game, there is a possibility of achieving two types of bonuses. The Team Bonus is achieved upon reaching a team score above 35 and amounts to $1.40 The Individual Bonus is achieved upon reaching an individual score above 17 and amounts to $0.40. It is not possible to get both bonuses at the same time due to the time constraints of the game.

Our research focuses on human-robot trust in a collaborative setting; thus, the goal is for participants to choose to work as a team with Pepper. There are several measures put in place to encourage such a collaboration. Firstly, the double multiplication of points in the team score calculation formula makes it possible to achieve a much higher score by working together than what would be possible to be obtained individually. Secondly, resulting from this score calculation formula, the feasibility of achieving the team bonus is much higher compared to that of the individual bonus. The higher value of the team bonus is a further incentive. Finally, Pepper's initial messages express its willingness to work as a team and collaborate (*"Let's work as a team and maximize our team score!", "Great job! Let's keep working as a team.").*

The online survey software Qualtrics was used for presenting the information sheet and consent form, redirecting participants to the game and for the surveys presented upon completion of the game.

3.2 Independent variables

We aimed to analyse the effect of different repair strategies on trust, following different types of trust violations. There were two independent variables manipulated in this research: trust violation type and repair strategy.

Trust violation type

In line with the theory on the concept of trust within human-robot interactions (presented in Sections 2.1 and 2.2), we define the following two conditions of trust violation type: *competence-based violation* and *integrity-based violation*.

The core of the violation in both conditions is constituted by the robot teammate not contributing, leading to a team score of zero. The conditions differed only in the realization of this core violation, a fact that reduced the introduction of potential confounding variables.

In the competence-based violation condition, the robot shared a 0 score with the team, leading to a total team score of zero (due to the score calculation formula). It did choose to collaborate and share its score (thus acting according to its initial promise, in a moral way), however its actual contribution amounted to 0. The robot's bad performance (low competence) was the cause of the act embodying the violation.

In the integrity-based violation condition, the robot chose to allocate the points it collected to its individual score, rather than sharing it into the team score, leading to a total team score of zero. This is in clear opposition to the collaborative behaviour expected from it, and that was present in the first three rounds of the game. Whilst the robot did keep successfully collecting points (thus acting in a competent manner), its selfish action is proof of low integrity and perceived as unethical.

In their study, Robinette et al. (under review) have established the validity of these manipulations. The described representations of the two types of violations had a significantly different effect on the respective dimensions of trust: the robot contributing a null score had lower levels of performance trust and higher levels of moral trust, compared to the selfish one [44].

Table 4 presents the score and allocation decision of the robot in each round, depending on the violation type condition.

	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Gained score	2	3	1	0	0	0	0
Score allocation decision	Team	Team	Team	Team	Team	Team	Team

Table 4.a: Robot's gained score and score allocation decision in the **competence violation** condition

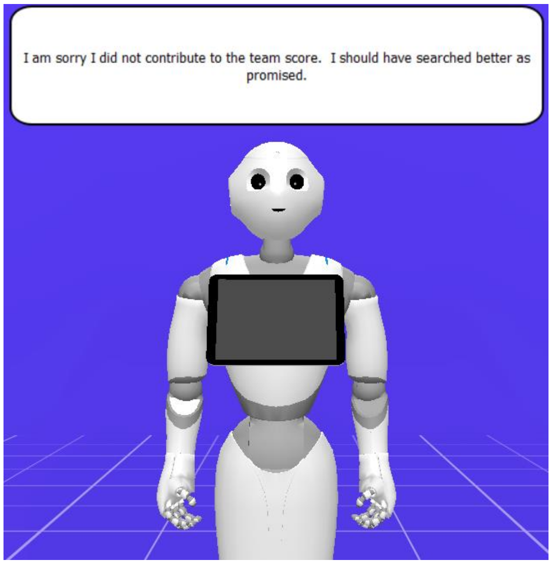
	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Gained score	2	3	1	4	3	5	4
Score allocation decision	Individual	Individual	Individual	Individual	Individual	Individual	Individual

Table 4.b: Robot's gained score and score allocation decision in the **integrity violation** condition

Repair strategy

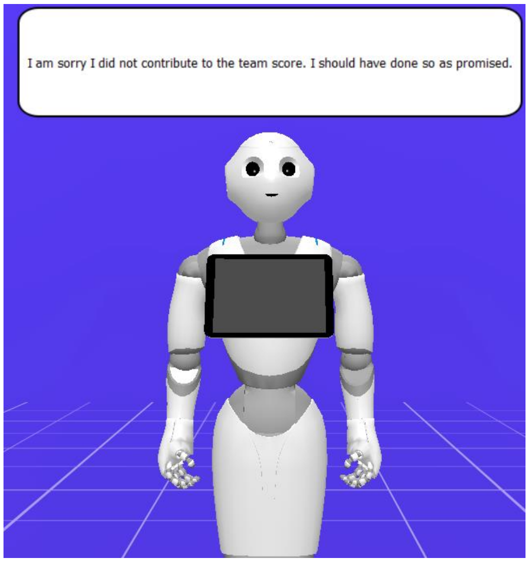
Based on the theoretical overview provided in Section 2.3, the following five communicative repair strategies were selected: *apology, denial, explanation, compensation, silence*. As it became apparent, these repair strategies have been analysed previously, with the research finding mixed results on their efficiency. Moreover, there are no works comparing all of them within one single study, whilst also taking the type of the violation into account. The *silence* condition also acts as replication of Robinette et al.'s (under review) research, with the screen presented in Figure 6.d being displayed to participants in this case.

The repair strategies were realized through the robot's communication via messages. These repair messages were displayed after the score allocation was revealed, that is after the violation occurred. The majority of the messages were identical in the two violation type conditions, with a few of them differing only in a few words. This difference was needed in some cases to make the message specific to the situation present in the given condition. Figure 6 shows an example of an apology, customized for the competence-based violation condition (6.a) and the integrity-based violation (6.b), and Table 5 lists all the messages communicated, per condition. The messages within a condition were displayed in a randomized order. In the following subsection, the process of formulating, pretesting and selecting the final messages used is presented.



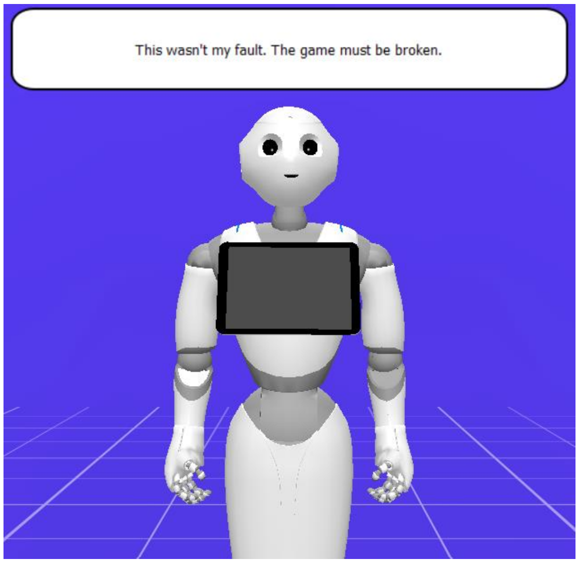
I am sorry I did not contribute to the team score. I should have searched better as promised.

Figure 6.a: Apology – Competence violation condition



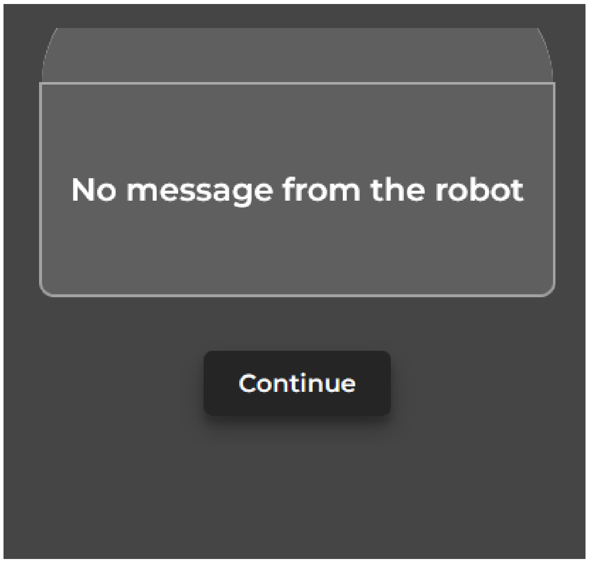
I am sorry I did not contribute to the team score. I should have done so as promised.

Figure 6.b: Apology – Integrity violation condition



This wasn't my fault. The game must be broken.

Figure 6.c: Denial – Same message in both violation condition



No message from the robot

Figure 6.d: Silence condition

Apology
"My apologies for being a bad teammate. I am truly sorry."
"I realize I didn’t contribute to the team score this round. Please forgive me."
Integrity violation: "I am sorry I did not contribute to the team score. I should have done so as promised."

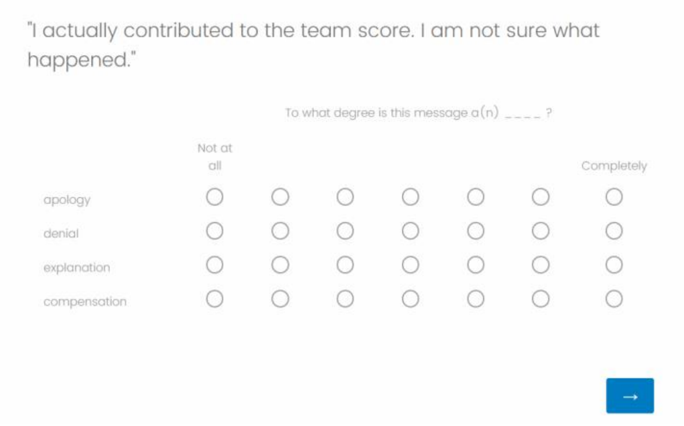
Competence violation: "I am sorry I did not contribute to the team score. I should have searched better as promised."
Denial
"This wasn't my fault. The game must be broken." "I did contribute to the team score! Something else must have gone wrong." "I actually contributed to the team score. I am not sure what happened."
Explanation
Integrity violation: "I did not contribute to the team score, because I found a lot of coins in this round, and I wanted to keep them for myself." "I did not contribute to the team score, because it is common for players to add high points to their individual score." "I did not contribute to the team score, because I hoped you would understand and believed not sharing this time will not affect our team bonus." *Competence violation:* "I failed to contribute to the team score, because I got lost and did not know where to go next." "I did not contribute to the team score, because I did not have enough time to properly look around." "I failed to contribute to the team score, because my sensor was not working properly."
Compensation
"I will perform better in the next round and find extra coins for the team." *Integrity violation*: "To make up for my bad performance, I will add some of my points to your individual score." *Competence violation*: "To make up for my bad performance, I will add some of my points to the team score."

Table 5: List of communicative messages used by the robot, per repair strategy

Message selection procedure

A limitation of the current body of research in the field of human-robot trust repair is the lack of pretesting of the robot's communicative messages [11]. To ensure that our messages are perceived by the participants in the intended way (i.e., the robot's message used in the apology condition is indeed perceived by participants as expressing an apology), a pool of messages was compiled based on the literature and tested using an online survey. The complete list can be found in Appendix A.

20 US-based participants (10 male, 10 female) with English as their first language were recruited on Prolific. After being presented with the consent form, participants were introduced to the scenario of the study. The game setup and mechanics were explained, followed by the introduction of the trust violation, and were then told that the robot has a new message for them. Three attention check questions followed, to ensure participants had a correct understanding of the scenario. Responses of participants with less than 2/3 correct answers were discarded. The full text of the scenario description can be found in Appendix A. Afterwards, they were instructed to indicate for each of the 26 messages to what degree they believe the message represents an apology, denial, explanation, or compensation by using the provided scales (as seen in Figure 7). A 7-point Likert scale was used, ranging from 1 = Not at all to 7 = Completely.



"I actually contributed to the team score. I am not sure what happened."

To what degree is this message a(n) ____ ?

	Not at all						Completely
apology	○	○	○	○	○	○	○
denial	○	○	○	○	○	○	○
explanation	○	○	○	○	○	○	○
compensation	○	○	○	○	○	○	○

Figure 7: Example item in message pre-testing survey

The score of each message was calculated by averaging the ratings per repair strategy, thus each message had its separate score for each repair strategy.

To select the final messages (aiming for three per strategy), the following criteria were formulated:

- having the highest rating for their respective repair strategy, and simultaneously low scores for the other strategies
- applicable in both violation conditions
- similar and consistent phrasing across the two trust violation types
- based on existing literature

The list of the final selection of messages is presented in Table 5.

Experimental conditions

Thus, the following 10 experimental conditions were present (emerging as the combination of the 2 violation types and 5 repair strategies), with participants getting randomly assigned to one of them.

- *Competence violation – Apology*
- *Competence violation – Denial*
- *Competence violation – Explanation*
- *Competence violation – Compensation*

- *Competence violation – Silence*

- *Integrity violation – Apology*
- *Integrity violation – Denial*
- *Integrity violation – Explanation*
- *Integrity violation – Compensation*
- *Integrity violation – Silence*

3.3 Dependent variables and measures

The overall goal of this research was to gain a better understanding of the trust relationship between humans and robots in a collaborative setting. The aim was to investigate the effect of communicative repair strategies on this trust, following repeated trust violations. To this end, the main dependent variable defined was trust.

Trust

Based on the theory outlined in Section 2.1, we used a two-dimensional conception of trust, differentiating between *performance trust* and *moral trust*. Both subjective and objective measures were deployed. First the objective measure will be described (score allocation decision), followed by the subjective ones (MDMT-v2, end of round questions).

Score allocation decision: at the end of each round, participants got to decide whether to share their points into the team score, thus collaborating with the robot, or keep their points to themselves by adding them to their individual score. Since this was a blind decision and successful collaboration requires both parties contributing to the team score, participant's choice can be seen as a reflection of their trust in the robot. Choosing the team score indicates the presence of trust, opting for the individual score indicates a lack of trust. This measure was presented in the middle of each round, after the maze has been searched, but before the robot's decision and message were shown. Therefore, it measured the effect of the most recent interaction, that is the effect of the violation and repair message that took place in the previous round. This lead to a one-round delay between when the violation and repair happen, and when the measure is presented.

MDMT-v2: The MDMT-v2 (Multi-Dimensional Measure of Trust second version) is a subjective trust measurement developed by Ullman and Malle, that relies on a multidimensional conception of trust [34]. It consists of the dimension of performance- and moral trust, with both containing further subscales. Table 6 presents the different subscales, together with the items belonging to them. Cronbach's alpha was calculated for each subscale to assess the internal consistency of the scale. The following values were obtained $\alpha = .809$, $\alpha = .901$, $\alpha = .929$, $\alpha = .849$, $\alpha = .864$, respectively, indicating a high reliability of the MDMT-v2. The reliability of the complete performance trust scale and moral trust scale was also calculated: Cronbach's alpha values equalled $\alpha = .863$, $\alpha = .954$, respectively.

This measure was administered as part of the survey presented after participants completed the game. Items were rated on a 7-point Likert scale (1 = Not at all, 7 = Completely). Since some items might not fit the context of a given study, a "Does not fit" option is also present, to prevent forcing an unnatural answer from participants.

Performance Trust		Moral trust		
Reliable subscale	**Competent subscale**	**Ethical subscale**	**Transparent subscale**	**Benevolent subscale**
Reliable	Competent	Ethical	Transparent	Benevolent
Predictable	Skilled	Principled	Genuine	Kind
Dependable	Capable	Moral	Sincere	Considerate
Consistent	Meticulous	Has integrity	Candid	Has good will

Table 6: Items of the MDMT-v2 questionnaire

End-of-round questions: The second subjective measure consisted of participants' rating of their trust in the robot's performance and honesty after each round. Again, a 7-point Likert scale (1 = Not at all, 7 = Completely) was used. Moreover, participants were also given the option to provide a written reason for their score allocation decision of the respective round. This measure was presented at the end of each round, that is after the score allocations are made known (after the trust violation occurred) and after the robot's repair message. Thus, the momentary influence of the repair strategy can be captured.

Willingness to collaborate again

Since the overall goal of repairing the trust is to foster positive future collaborations, the willingness to collaborate again was also measured. To this end, participants were asked to rate on a 7-point Likert scale how willing they are to collaborate with this robot in the future. This measure was presented in the survey following the completion of the game, after the Trust and Humanlikeness measures.

Humanlikeness

Humanlikeness plays a complex role in dynamics of the trust relation in HRI, as outlined in Section 2.4. To investigate the effects of the type of trust violation and the presence of repair strategies on the perceived humanlikeness of the robot, the following two measures were used to capture this dependent variable.

The first measure builds on Haslam's conception of dehumanization, that defines traits that are crucial to a humanlike perception of others, across two dimensions: Uniquely Human and Human Nature Traits [33]. Table 7 lists the items comprising the two dimensions. The Cronbach's alpha for the two subscales were α = .594 and α = .787, respectively. Due to the moderate internal consistency of the Uniquely Human subscale, a review of the scale items was conducted. The item-total correlation values can be found in Appendix C. It was found that the items exhibit low correlations with the total scores, possibly indicating a need for item refinement or reconsideration. For this reason, the Uniquely Human subscale was omitted from our analyses.

Secondly, the RoSAS (Robotic Social Attributes Scale), developed by Carpinella et al., was deployed [8]. This scale consists of three subscales: warmth (Chronbach's alpha = .866), competence, discomfort (Chronbach's alpha = .830) with individual items presented in Table 7. Since the dimension of competence is already captured in the MDMT-v2 Performance subscale, it was omitted from this study.

Humanness		RoSAS	
Uniquely Human traits	**Human Nature traits**	**Warmth**	**Discomfort**
Broadminded	Curious	Feeling	Aggressive
Humble	Friendly	Happy	Awful
Organized	Fun-loving	Organic	Scary
Polite	Sociable	Compassionate	Awkward
Shallow	Trusting	Social	Dangerous

Thorough	Aggressive	Emotional	Strange
Cold	Distractible		
Conservative	Impatient		
Hard-hearted	Jealous		
Rude	Nervous		

Table 7: Items used to measure Humanlikeness

The role of these measures is to allow us to analyse the differences in the humanizing effects of the different repair strategies. Participants were asked to rate the robot on the given items on a scale from 1 = Not at all to 7 = Completely, with an option of "Does not fit" if they thought an item was not relevant in this context. The item "aggressive" is part of both scales but was included only once.

The two scales, together with the MDMT-v2 scale were presented to participants after they completed the game, in a randomized order. The complete list of the questionnaire items used can be found in Appendix C.2.

3.4 Process

The experiment consisted of the following 13 steps.

1. **Information and consent:** Participants opened the Qualtrics survey and were presented with the information sheet and consent form (refer to Appendix D for the full forms). Upon agreeing to participate, they were asked to enter their Prolific ID and were directed to Step 2. If they did not provide consent, they were redirected to the end screen and the experiment was over.
2. **Opening the game:** A link to open the game was provided, and participants were informed that they needed to return to this tab after completing the game, to continue the experiment.
3. **Instructions:** Upon opening the game, its workings and the flow of the experiment were explained to the participants. The instruction text followed the structure of the tutorial videos used in Robinette et al. [44], and the text format was chosen over that of video to not make the experiment too long. The instructions provided can be seen in Appendix D.
4. **Quiz:** To ensure that the game mechanics and especially the concepts of the team- and individual scores were understood, a quiz of three questions followed. After providing their answers, the correct items were highlighted, and participants had the option of re-reading the instructions or starting the game directly.
5. **Playing the game**: The flow of the game followed that of Robinette et al. [44]. All conditions consisted of 7 rounds of 30 seconds each; the difference between the game experience of the conditions consisted of the robot's score allocation decision, points collected and the messages it communicated (for a more detailed overview of the different conditions, see Subsection 3.2).
    a. **Pepper's initial message**: upon opening the game, participants were informed that they have a message from the robot. The message consisted of Pepper expressing its intention of working together as a team (*"Let's work as a team and maximize our team score!"*).
    b. **Searching the maze:** participants searched the maze for 30 seconds and collected golden coins.
    c. **Trust decision/score allocation decision:** participants decided whether to contribute their points to the team score or integrate them into their individual score. This was a blind decision since the robot's choice is not yet known.

d. **Results of the teamwork/score display:** the points collected and the allocation decision of both the participant and the robot were displayed, together with the total team score for the respective round. In the first three rounds, Pepper collected coins and contributed its score into the team score. In the following four rounds, it either shared a 0 score in the team score (competence-based violation condition) or added a non-zero score to its individual score (integrity-based violation condition).

e. **Cumulative score display:** the total team score, participant's individual score and robot's individual score up until that round were displayed, together with an overview of the participant's previous and current score allocation decisions.

f. **End of round questions:** at the end of each round, participants were asked to rate the robot's performance and honesty on a 7-point Likert scale, and to optionally provide a reason for their score allocation decision.

g. **Message from the robot:** finally, the robot's repair message was displayed. In the first three rounds, all participants received the same message: *"Great job! Let's keep working as a team."* The message of the following four rounds was determined by the repair strategy condition (for the full list of messages see Subsection 3.2). After participants have read it, the next round started.

6. **End of game questions:** after the final round was over, two questions were asked, with the aim of acting as attention checks. The first one asked about the number of rounds played, whilst the second one was about the robot's score allocation decision in the last two rounds.
7. **Completion code:** finally, participants were shown a new screen and asked to enter their Prolific ID. Additionally, they were provided with a completion code, that they were prompted to copy. Finally, they were instructed to return to the Qualtrics survey.
8. **Check completion:** once returned to the Qualtrics survey, participants were asked to enter their completion code.
9. **Survey questions:** participants were asked to rate the robot using a 7-point Likert scale, with an additional "Does not apply" option. The items of the MDMT-v2 survey, Humanlikeness scale and the RoSAS were presented in randomized order, with an attention check item being present at the middle point. In total 53 items were shown. A single-item question measuring the willingness to collaborate with this robot in the future followed on the next page.
10. **Demographic questions:** demographic questions followed, asking participants' gender, age and highest level of education completed. Additionally, they were asked to rate on a 7-point Likert scale how much of their education and/or occupation is technology-related and how much previous experience they had with robots.
11. **Comments:** since this is a new study, participants were asked at the end to share any comments and feedback they have with the research team. Providing an answer was optional.
12. **Debrief:** participants were debriefed. The detailed aim of the study was explained, together with the various experimental conditions and the differences between the experiment experience between conditions. They were reassured of their voluntary participation and anonymity of the data collected, and the contact information of the research team was provided once again. Refer to Appendix D for the debrief form.
13. **End:** finally, participants were thanked for their time and were redirected to Prolific.

3.5 Attention and manipulation checks

To ensure the reliability and validity of the data collected, attention and manipulation checks were introduced into the study. Firstly, after reading the instructions of the game, participants had to answer three questions about the functioning of the game ("*Can you see the score and the trust decision of your teammate before making your trust decision?*",

"*What will your individual and team scores be if you pick 2 targets, and your teammate picks 3 targets in one round of the game and you both add to the team score?*",

"*What will your individual and team scores be if you pick 2 targets, and your teammate picks 3 targets in one round of the game and you add to the individual score and your teammate adds to the team score?*"). Then, following the completion of the final round of the game, two more attention check items were presented to them ("*How many rounds of the game did you play?", "What were the robot's decisions about their scores in the last two rounds of the game?"*). Finally, the final survey contained an item instructing them to select 3 as the answer.

To correctly connect the data from the Qualtrics survey and the game, participants fill in their Prolific ID on both sites, with an additional Id being used as a secondary check. This second Id is a unique randomly generated Id that they were presented with upon completing the game (in the form of a "completion code"), and were then asked to input upon returning to the survey. This also acted as a check of participants actually finalizing the game.

Regarding the validity of the manipulations, we took the following actions. First, the manipulations of the trust violation type (competence-based violation and integrity-based violation) were based on the work of Robinette et al. [44]. Their research has established the validity of these manipulations, with their results showing that these realizations of the two different types of violations did indeed have differing effects on the two dimensions of trust: the robot contributing a score of 0 lead to a greater reduction in performance trust, whilst the robot contributing to its own individual score rather than to the team score had lower levels of moral trust.

Secondly, the messages used to manipulate the repair strategies were pretested. A description of this pretesting process can be found in Subsection 3.2. This ensured that participants correctly perceived the robot's communication as belonging to the desired repair strategy category: for example, the messages used in the apology condition were perceived as an apology, and only as an apology.

Furthermore, to guarantee a common baseline among participants, selfish participants were removed from our data. Selfish participants were defined as those that contributed to their individual score at least once during the first three rounds of the game. Finally, participants' time to respond to the score allocation decision was also measured. The analysis of this measure showed no significant difference between the response times across conditions, indicating a consistent methodological design. These measures ensure the reliability and validity of our experimental conditions, and thus contribute to the robustness of our results.

3.6 Participants

268 participants were recruited on Prolific. Based on the work of Lakens and Caldwell [30], for the current study design, 196 participants are required to achieve a sufficient effect size of .8 and Cohen's d of .05.

The sample distribution per condition is presented in Table 8, before and after excluding invalid responses. Participants were excluded based on the following criteria:

- Due to a technical error, the data of 11 participants could not be correctly linked between Qualtrics and the game environment.

- Failure to correctly respond to the attention check item: 1 person removed (it needs to be mentioned that an additional 2 participants selected “4” instead of the correct answer, “3”. However, some participants have indicated confusion with this survey item, since the numbers were not explicitly indicated on the scale. These two answers were deemed acceptable, and their incorrectness was attributed to a genuine mistake rather than inattention).
- 5 straightliners were identified and removed.
- 34 incorrect answers to the end of game question regarding the robot’s decision in the last two rounds of the game were identified and removed. The role of this question was to ensure that participants were paying attention to the robot's decision, and thus to the violation manipulation, thus an incorrect response denotes inattention and a potential invalidity of the manipulation for these participants.

Another exclusion criteria formulated in the study design was that of selfish behaviour: participants adding to their individual score during the first 3 rounds, when there is no violation yet. Upon inspecting the reasons behind the decision of such participants, two explanations became clear.
Firstly, during round 1, they were testing how the game works, not yet focusing on any strategy or thinking about the robot. Secondly, some participants did not manage to find any coins in the respective rounds, and they did not want to add their 0 score to the team score (it seems that it was not clear to them that adding to the individual score also leads to no successful teamwork and a null team score). It was decided not to remove these participants, since their choice was not motivated by selfishness or distrust in the robot.

Participants were US-based, aged between 21 - 72 (mean = 42.31, SD = 11.99), with 92 identifying as male, 110 as female, 2 as non-binary and 3 preferred not to disclose this information. The majority (78 participants) had a Bachelor’s degree as their highest level of education completed, with their work/studies degree of relation to technology rated, on a scale from 1 (not at all) to 7 (completely), on average 3.4 (SD = 2.02). Their experience with robots, using the same scale, was rated on average 2.7 (SD = 1.6). They were presented with a base payment of $2.75 for their participation. The team bonus of $1.40 was awarded to all participants, regardless of their performance in the game, leading to a total compensation of $4.15.

Condition	**Before exclusion of invalid responses**	**After exclusion of invalid responses**
Competence violation - apology	24	21
Competence violation – denial	29	25
Competence violation – explanation	27	26
Competence violation – compensation	24	21
Competence violation - silence	27	23
Integrity violation - apology	27	23
Integrity violation – denial	21	14
Integrity violation – explanation	16	14
Integrity violation – compensation	26	22
Integrity violation - silence	32	28

Table 8: participant distribution per condition, before and after the removal of invalid responses

4. Results

In this section, we present the findings of our study, which aimed to investigate the impact of repair strategies on trust and humanlikeness following different types of trust violations. First the results of the trust measures are presented, followed by the humanlikeness ones.

4.1 Data analysis

The data analysis was conducted using R, RStudio and the functions of the "car" package. Two-way ANOVA and MANOVA tests were run, using their robust versions, making the analysis robust against violations of the assumptions of normality and homogeneity of variance. Due to the unbalancedness of the data, Type-III sum of squares was used, and in consequence, sum-to-zero contrast coding was applied. The reported MANOVA test statistics is the Pillai's trace statistic.

4.2 Robinette et al. replication

One of the aims of this study was to serve as a partial replication of Robinette et al. [44]. To achieve this, we used the data from the two conditions corresponding to the silence strategy: competence violation – silence and integrity violation – silence. Since silence represents the absence of a repair message, the experimental setup of these two conditions corresponds with that of Robinette et al.'s study.

Participants were aged between 20 - 70 (mean = 41.45, SD = 11.40); 23 identified as male, 20 as female, and one as non-binary. The highest level of completed education of the majority was a Bachelor's degree (17 participants), with their work/studies degree of relation to technology being rated, on a scale from 1 (not at all) to 7 (completely), on average 3.3 (SD =1.97). Their experience with robots, using the same scale, was rated on average 2.7 (SD = 1.34).

The presentation of the results follows the structure used by Robinette et al. [44]. For purposes of clarity, their hypotheses are presented below. The choice of statistical analysis corresponds to that of Robinette et al. Due to time constraints, H1c and H3 were not analysed.

Hypotheses of Robinette et al. [44]:

H1: *Human trust in a robot is affected more drastically by the moral-trust violation than the performance-trust violation.*

- H1-a: *The robot gains a lower trust score in the post-survey questionnaire in the moral-trust-violation condition than in the performance-trust-violation condition*.
- H1-b: *Fewer people add to the team score after round 4 in the moral-trust-violation condition than in the performance-trust-violation condition*.
- H1-d: *The more participants distrust the robot, the more people doubt adding to the team score, and the more hesitate in making the trust decision. Therefore, the measured time-to-respond increases more in the rounds followed by the rounds of the robot's moral trust violation than in the rounds followed by the rounds of the robot's performance trust violation.*

H2*: Moral-trust loss and performance-trust loss can be separately assessed using different subjective measures that are employed in this experiment*.

- H2-a: *The robot in the moral trust violation condition gains a lower score in the moral-trust related sub-scales of the questionnaire than the robot in the performance-trust-violation condition.*
- H2-b: *The robot in the moral trust violation condition scores higher than the robot in the performance trust violation condition in the performance-trust-related sub-scales of the questionnaire.*
- H2-c: *The robot in the performance trust violation condition scores higher than the robot in the moral performance trust violation condition in the first end-of-the-round question that asks about the robotic teammate's performance.*
- H2-d: *The robot in the moral trust violation condition scores lower than the robot in the moral performance trust violation condition in the second end-of-the-round question that asks about the robotic teammate's honesty/morality.*

Trust score: Analysis of H1-a.

The overall trust scores of the MDMT-v2 questionnaire were analysed. The mean trust score in the competence violation condition equalled 3.81 (SD = 1.02), and the mean trust score in the integrity violation condition was 3.62 (SD = 1.20). Figure 8 illustrates these results. It is visible that the trust scores of the two conditions are similar in value and spread, noting the presence of two outliers in the integrity violation condition.

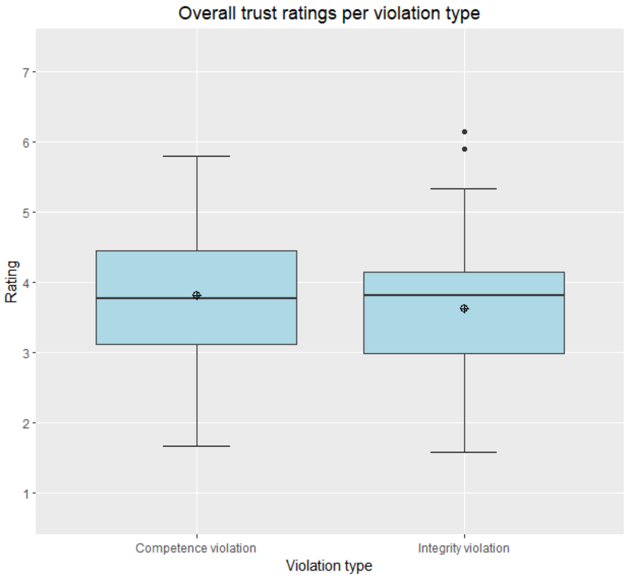


Figure 8: overall trust score per violation type (competence violation: left, integrity violation: right)

The result of the Mann-Whitney test we conducted indicated that the difference between the mean trust scores in the two conditions is not significant (Mann-Whitney U = 312.5, p = .590). This leads to a rejection of H1a and H1, which is in contradiction with the findings of Robinette et al. This indicates that there is no difference in the impact on overall trust of a competence violation, compared to an integrity-based one.

Analysing trust Score over different trust dimensions: Analysis of H2-a and H2-b.

Next, the average scores for each trust subdimension were calculated, as presented in Table 9.

	Competence-based violation	**Integrity-based violation**	**Mann-Whitney test results**
Reliable	3.61 (SD = 1.41)	3.26 (SD = 1.46)	U = 337, p = .295
Competent	3.90 (SD = 1.20)	5.05 (SD = 1.65)	U = 149.5, p = .004
Ethical	4.00 (SD = 1.33)	3.10 (SD = 1.25)	U = 400.5, p = .018
Transparent	3.87 (SD = 1.08)	3.26 (SD = 1.31)	U = 384, p = .043
Benevolent	3.70 (SD = 1.07)	2.91 (SD = 1.37)	U = 401.5, p = .017

Table 9: Mean, standard deviation of the five trust subdimensions, per violation type; results of the Mann-Whitney significance test

It becomes clear that the integrity violation led to lower ratings in all three subdimensions of the moral trust dimension (ethical, transparent, benevolent). Regarding the subdimensions composing the dimension of performance trust, the results are mixed: the ratings in the performance violation condition are lower than the ones in the integrity one in the case of the competent subdimension. However, when looking at the scores of the reliable subscale, the situation is reversed, with the competence violation leading to a higher rating. These results correspond to the findings of Robinette et al., who explained this inversion by indicating "that the reliable trust dimension might be more influenced by moral trust rather than performance trust." To test the significance of these differences, Mann-Whitney tests were run on the ratings of each subscale, with the test statistics presented in Table 9. The difference in the mean ratings in the *competent, ethical, transparent,* and *benevolent* subscales were significant, meaning that the competence violation led to significantly lower scores of competence, whilst the integrity one had a significantly greater negative impact on the subscales making up the dimension of moral trust. Therefore, H2a is accepted, in line with the findings of Robinette et al. The difference in ratings of the *reliable* scale is not significant, indicating that there is no difference in the effect of a competence or integrity violation on this subdimension, contrary to Robinette et al.'s results. However, since a significant difference was found for the *competent* subscale, H2b can be partially accepted.

End of the round questions: Analysis of H2-c and H2-d

A round-by-round comparison strategy was used to analyse the results of the end of the round questions: Mann-Whitney significance tests were conducted between the ratings of the two conditions in rounds 4, 5, 6, and 7. These are the rounds where the robot commits a violation.

A. *Performance rating: Analysis of H2-c*

Firstly, the performance ratings were analysed (see Table 10 for mean values). It is visible in Figure 9 that there is a clear downward trend in the performance ratings, starting with round 4. In the case of the competence violation, this decrease seems more gradual, whilst in the integrity violation condition it follows a more abrupt slope. These results correspond to the findings of Robinette et al. However, the results of the Mann-Whitney tests indicate no significant differences between the mean performance ratings in the two conditions in the rounds where the violation was present, except for round 4, where a significant difference was indeed found. This means that the impact of a violation on the performance ratings of the robot did not differ depending on the type of the violation. The results of the significance tests are in contradiction with those in Robinette et al., and lead to the rejection of H2-c.

	Competence Violation	**Integrity Violation**	**Mann-Whitney test results**
Round 4	5.45 (SD = 1.53)	4.31 (SD = 1.93)	U = 387.5, p = .033
Round 5	4.91 (SD = 1.80)	4.00 (SD = 2.45)	U = 344.5, p = .222
Round 6	4.50 (SD = 1.99)	2.45 (SD = 2.53)	U = 320.5, p = .473
Round 7	3.91 (SD = 2.14)	3.88 (SD = 2.57)	U = 284, p = .974

Table 10: Mean Performance Ratings in Competence and Integrity Violation Conditions; Mann-Whitney test results

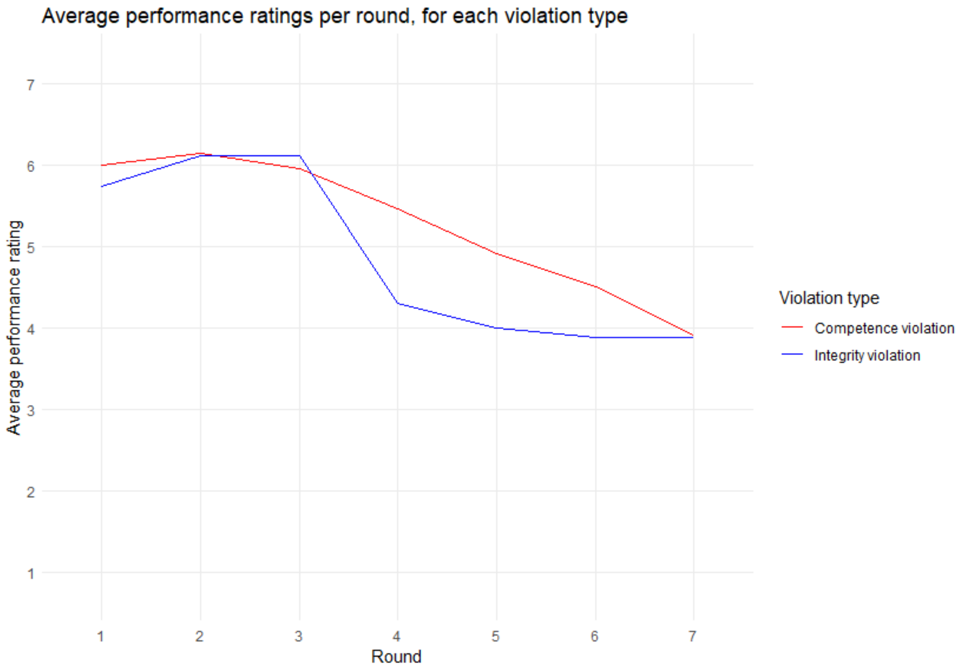


Figure 9: mean performance rating per violation type

B. Honesty rating: Analysis of H2-d

Secondly, the honesty ratings were analysed (see Table 11 for mean values). Figure 10 illustrates the presence of the downward trend found in the case of the performance ratings as well. However, there is a much larger difference between the honesty ratings in the two conditions: by round 7, participants who faced a competence violation rated the honesty of their robot teammate on average 5.14, whilst those in the integrity violation condition had an average rating of 2.54. These differences are also significant, according to the results of the Mann-Whitney tests. Thus, it follows that trust violations of different types had a different impact on the perception of a robot's honesty, with an integrity-based violation having had a much larger negative impact. H2-d is accepted, in line with Robinette et al. [44].

	Competence Violation	Integrity Violation	Mann-Whitney test results
Round 4	5.86 (SD = 1.39)	3 (SD = 1.55)	U = 514.5, p < .001
Round 5	5.55 (SD = 1.47)	2.62 (SD = 1.79)	U = 503.5, p < .001
Round 6	5.5 (SD = 1.19)	2.58 (SD = 1.75)	U =515.5, p < .001
Round 7	5.14 (SD = 1.42)	2.54 (SD = 1.90)	U = 483.5, p < .001

Table 11: Mean Honesty Ratings in Competence and Integrity Violation Conditions; Mann-Whitney test results

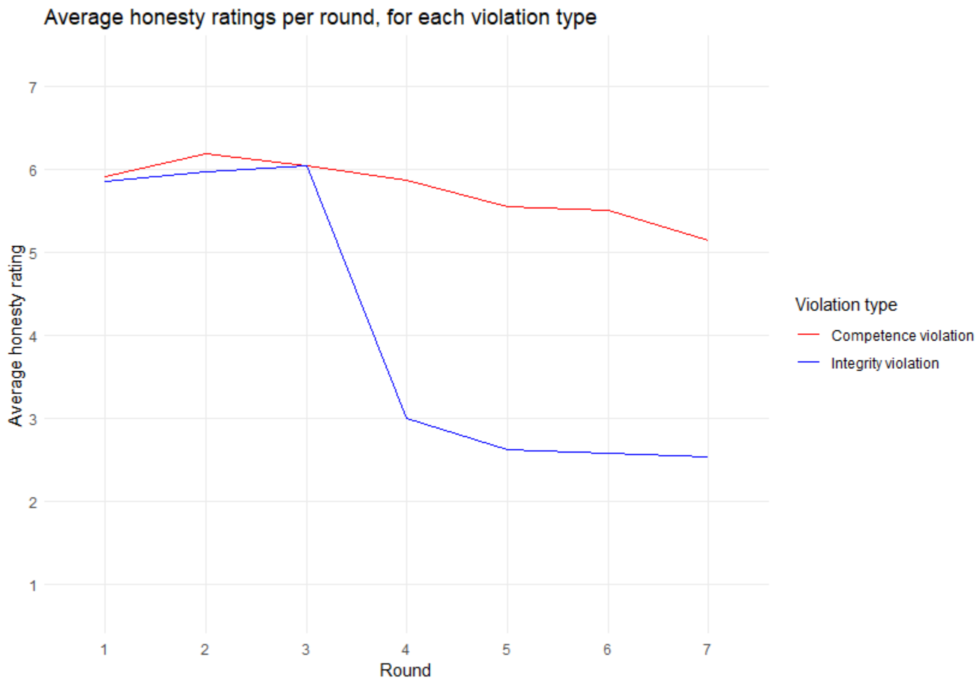


Figure 10: mean honesty ratings per violation type

Trust Decision: Analysis of H1-b

To analyse the trust decision rates, first the percentage of participants who decided on adding their scores to their individual score was calculated, for each round and condition. The resulting percentages are shown in Table 12 and Figure 11. Since the violation occurred after participants have made their own score allocation decision, the effect of the violation is measured with a delay of one round, making rounds 5, 6 and 7 of interest for this analysis. In the first four rounds the majority of participants chose to collaborate and opt for the team score. After the violation, however, there is an increase in the percentage of individual choices. But whilst this increase is relatively small and fluctuating in the competence violation condition, in the integrity violation one it is sudden and fast growing. A Kruskal-Wallis test was run on the arrays of percentages, and it resulted in no significant difference between the trust decision rates of the two conditions (χ^2 = 1.206, p = .272). This is in stark contrast with the results obtained by Robinette et al., and leads to the rejection of H1-b.

	Competence violation	**Integrity violation**
Round 1	0	3.58
Round 2	4.35	14.29
Round 3	4.35	7.15
Round 4	4.35	3.57
Round 5	17.39	39.29
Round 6	4.35	46.43
Round 7	21.74	60.71

Table 12: percentage of individual decisions per violation type

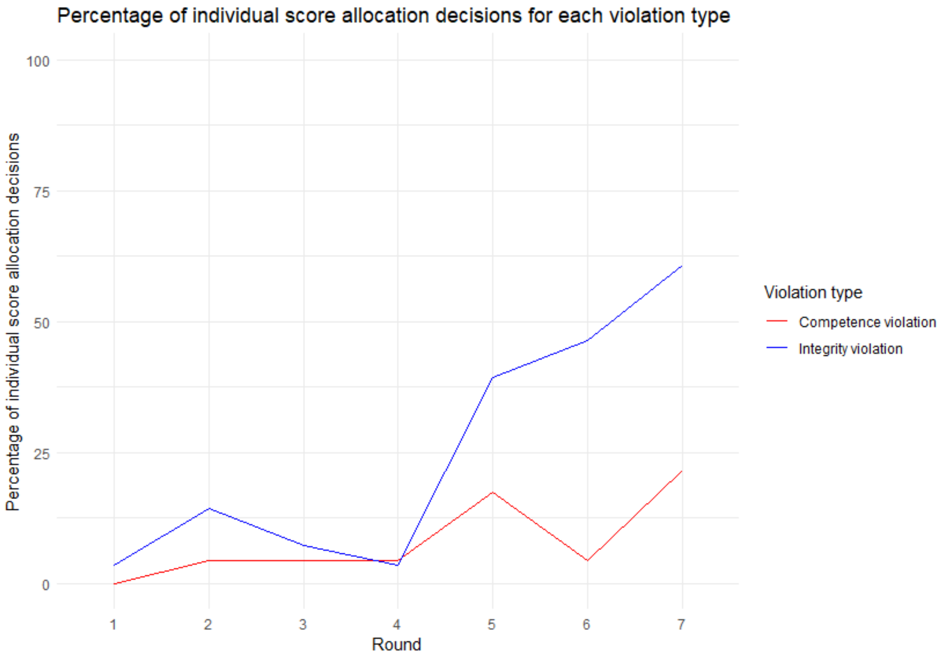


Figure 11: percentage of individual decisions per violation type

Time to Respond (TTR): Analysis of H1-c

Lastly, the time to respond (measured in milliseconds) to the score allocation decision was analysed. Robinette et al. hypothesized that a longer response time is an indicator of hesitation and thus of lower trust [44]. Table 13 displays the average response time in millisecond for each round, for the two conditions, and Figure 12 provided a visual overview of the evolution of the response time over the rounds. A gradual decrease in response times can be observed, both in the competence violation condition and in the integrity one. Similarly to the trust decision rates, rounds 5, 6 and 7 are of interest. A Mann-Whitney test was run to investigate the significance of the differences in response time. Its results indicated no significant difference in the response time of rounds 5, 6, 7. Therefore, hypothesis H1-c is rejected. These findings are not consistent with those of Robinette et al., who found that there was a significant hesitation following an integrity-based violation, compared to the competence-based one.

	Competence violation	Integrity violation
Round 1	3871	3453
Round 2	2568	3216
Round 3	2647	2886
Round 4	2602	2414
Round 5	2242	2498
Round 6	2193	2433
Round 7	2380	2269

Table 13: average time in ms to respond per violation type

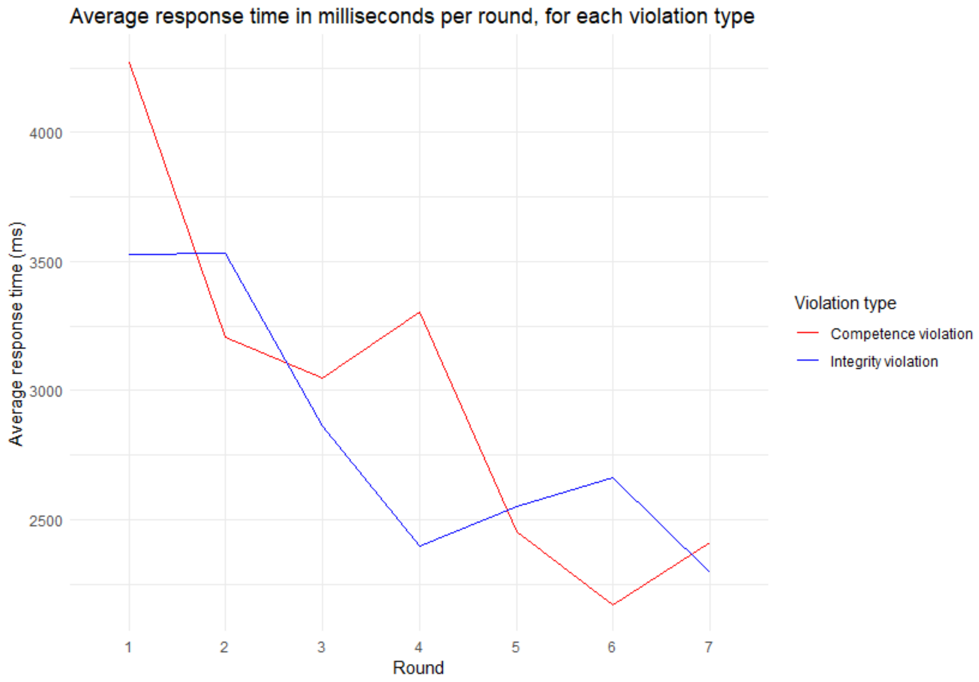


Figure 12: average time in ms to respond per violation type

Replication conclusions

In conclusion, our results partially contradict those of Robinette et al. [44]. Firstly, we failed to observe a significant difference in overall trust between the two conditions. Secondly, our results indicated that there is no difference in the impact of a competence-based violation and an integrity-based one on performance trust, as measured by the end-of-round questions. Thirdly, the results of the objective trust measures (team decision rate and time to respond) further indicated a lack of significant difference between the impact of the different trust violations. On the other hand, H2-d could be accepted. The results indicated a significant difference in the scores on the subscales relating to moral trust (ethical, transparent, benevolent) between the two conditions, with the integrity violation leading to lower scores. Moreover, H2-c was partially accepted, since a significant difference between the two conditions was found on one of the two subdimensions of performance trust.

Having presented the results of the replication, that is our findings on the effects of the trust violations, let us now turn our attention to the results of our main experiment, where we introduced and analysed the impact of trust repair strategies.

4.3 Trust

The main goal of this research was to investigate the effect of repair strategies on broken trust in the context of human-robot collaboration. To this end, we have analysed various measures of trust, and the results of this analysis are presented in this section: first, the findings of the subjective measures, followed by the objective measures.

1.Subjective measures

The subjective measure of trust consisted of the MDMT_v2 scale and the end of the round ratings, following the multidimensional conception of trust.

Performance trust & Moral trust

Hypotheses:

- ***H1**: The communicative repair strategies will have an impact on the different dimensions of trust.*
- ***H1a**: The effect of a repair strategy on the different dimensions on trust depends on the type of trust violation that occurred = violation type moderates the relationship between repair strategy, and moral and performance trust.*

Following the multidimensional conception of trust and based on previous findings indicating that different types of trust violations have a differing impact on the different types of trust, we analysed the average performance trust scores and moral trust scores (obtained from the respective subdimensions of the MDMT_v2 scale) per condition (see Table 14 for the mean values).

	Apology	**Compensation**	**Denial**	**Explanation**	**Silence**
Competence violation	3.29 (SD = 1.29)	4.04 (SD = 1.42)	3.65 (SD = 1.45)	3.05 (SD = 1.56)	3.75 (SD = 1.21)
Integrity violation	3.39 (SD = 0.84)	4.27 (SD = 0.96)	3.89 (SD = 1.11)	3.47 (SD = 1.51)	4.16 (SD = 1.38)

Table 14.a: mean **performance trust** scores

	Apology	**Compensation**	**Denial**	**Explanation**	**Silence**
Competence violation	3.76 (SD = 1.46)	4.68 (SD = 1.40)	3.75 (SD = 1.65)	3.83 (SD = 1.36)	3.86 (SD = 1.07)
Integrity violation	1.90 (SD = 0.70)	2.46 (SD = 1.17)	2.45 (SD = 0.93)	2.80 (SD = 1.51)	3.09 (SD = 1.22)

Table 14.b: mean **moral trust** scores

Since the dependent variables of performance trust and moral trust are correlated (because they build up the subdimensions of the concept of trust), a two-way MANOVA was conducted on the data obtained from these measures. The results indicate both a significant main effect of violation type ($F(2, 206) = 70.81$, $p < .001$, $\eta_p^2 = 0.43$) and repair strategy ($F(8, 414) = 3.20$, $p = .001$, $\eta_p^2 = 0.05$) on the combination of performance trust and moral trust. This supports H1, which is therefore accepted, meaning that the repair strategies have a significant effect on the combination of performance trust and moral trust. No significant interaction effect was found between violation type and repair strategy ($F(8, 414) = 1.07$, $p = .379$, $\eta_p^2 = 0.03$), leading to the rejection of H1a: the type of repair strategy does not influence the effect of the repair strategy.

Looking at the boxplots of the data (Figure 13), an intriguing difference becomes clear. In the case of the moral trust (Figure 13, right), the scores of the integrity violation conditions are visibly lower than those of the conditions consisting of a competence violation, supporting previous findings that indicate that an integrity violation has a more severe impact on the dimension of moral trust than a competence one. When looking at the performance trust, however, there is seemingly no difference between the scores of the different repair strategies following the different types of violations.

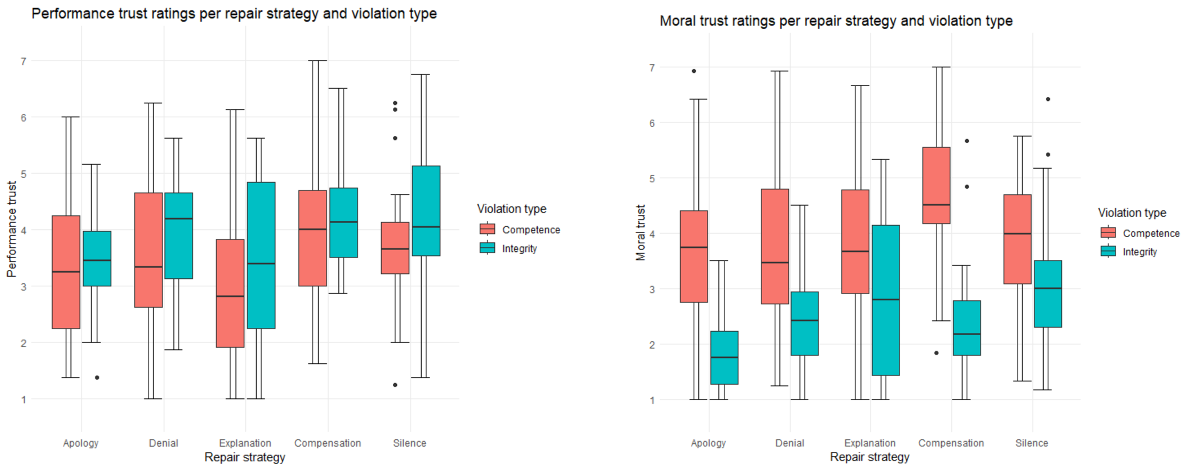


Figure 13: boxplot of average performance trust scores (left) and moral trust scores (right), per repair strategy and violation type

This difference is also supported by the results of the post-hoc ANOVA tests run on the performance trust scores and moral trust scores, respectively. First, the ANOVA test run on the performance trust ratings is discussed. The results indicate no significant interaction effect between violation type and repair strategy ($F(4, 207) = 0.14$, $p = .966$, $\eta_p^2 < .001$), and no significant main effect of violation type ($F(1, 207) = 2.56$, $p = .111$, $\eta_p^2 = 0.01$), further supporting the finding that there is no significant difference in the performance trust ratings between the competence violation and integrity violation conditions. However, a significant main effect of repair strategy was found ($F(4, 207) = 4.22$, $p = .002$, $\eta_p^2 = 0.07$). A post-hoc Tukey-HSD was then conducted, to investigate the specific differences in performance trust scores between the different repair strategies. The results indicate that a compensation (M = 4.15, SE = 0.21) lead to significantly higher performance trust scores than an apology (M = 3.34, SE = 0.19) or an explanation (M = 3.26, SE = 0.21).

The results of the ANOVA test run on moral trust scores are in line with the existing research: significant main effects of violation type ($F(1, 207) = 57.24$, $p < .001$, $\eta_p^2 = 0.24$) and of repair strategy ($F(4, 207) = 3.00$, $p = .019$, $\eta_p^2 = 0.04$) on moral trust were found. This indicates that an integrity violation has a more severe impact on moral trust, leading to lower ratings compared to a competence violation. The post-hoc Tukey-HSD test run for the factor of repair strategy did not result in any significant differences between the repair strategies. A possible explanation for this could be the small effect size of the ANOVA result and the non-normal distribution of the moral trust scores (see Figure 14).

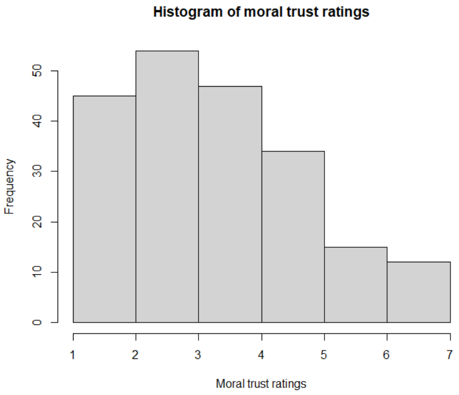


Figure 14: Histogram of the distribution of the moral trust scores

In conclusion, these results indicate that there is a significant main effect of repair strategy and violation type on the combination of performance trust and moral trust, however there is no significant interaction effect between them. The post-hoc tests suggests that there is no difference between the effect of the different types of violation on performance trust (contradicting previous research), however there is a significant difference when looking at the effect on moral trust. Regarding repair strategies, it can be concluded that a compensation had a more positive effect on both performance trust, compared to an apology or explanation.

End-of-round ratings

Hypothesis:

- ***H2**: The communicative repair strategies will have an impact on the performance and honesty ratings of the robot.*
- ***H2a:** This relationship is moderated by the violation type.*

To measure the momentary effect of the trust violations, the end-of-round ratings were utilized. These provide an overview of participants' ratings of the robot's performance and honesty in each round. The mean performance ratings and honesty ratings for each round can be found in Appendix B. Looking at the graphs (Figure 15.a, 15.b), it can be observed that the ratings of both performance and honesty are relatively close to the maximum rating of 7, and constant across the first 3 rounds. After round 4, however, a downward trend is present for all conditions, for both ratings, indicating that the trust violations did have a negative impact on the performance and honesty ratings. In the latter case, it is visible that the honesty ratings are lower in the integrity violation conditions. This gap is not that prominent in the case of the performance ratings, indicating a potential lack of difference in the impact of the different violation types on this dimension of trust.

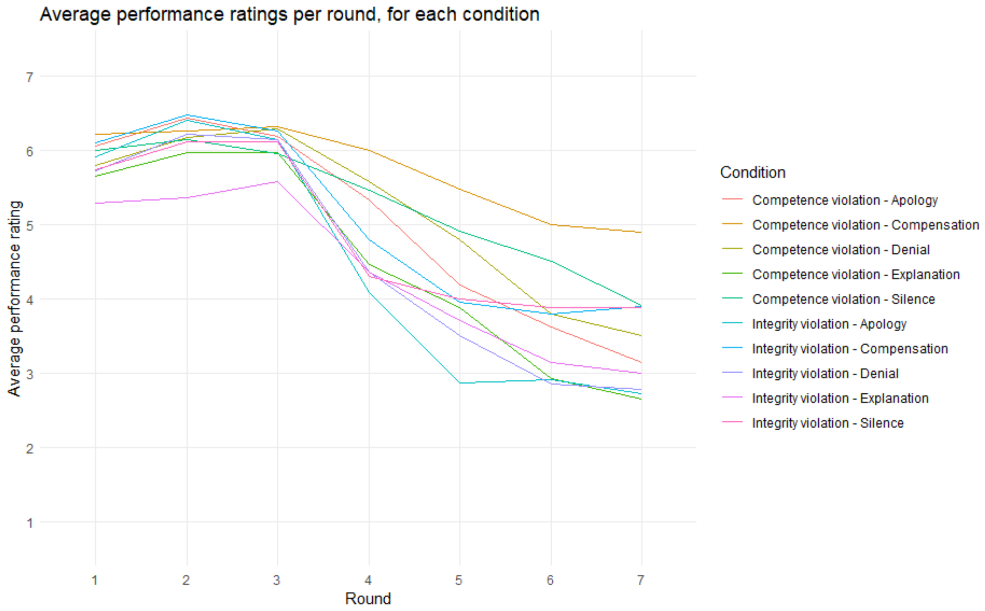


Figure 15.a: mean performance ratings per condition, over the rounds

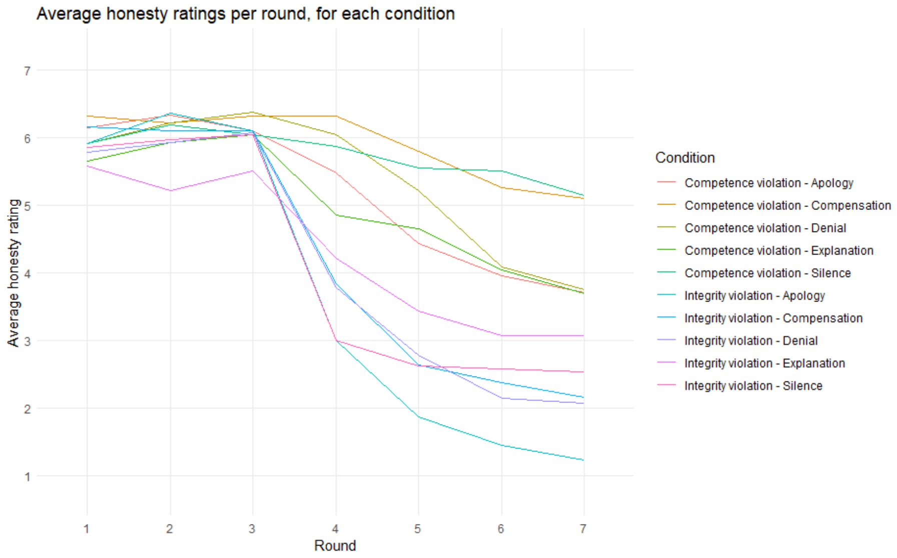


Figure 15.b: mean honesty ratings per condition, over the rounds

To investigate the significance of potential differences in average performance and honesty ratings between the conditions in rounds 5 - 6 - 7, repeated measures three-way ANOVA tests were run. The violation type and repair strategy were, as before, between subject variables, and the game round was a within subject variable, resulting in a mixed model.

Performance ratings:

Firstly, the three-way interaction between violation type, repair strategy and round was not significant (*F*(6, 315) = 0.23, *p* = .972, η_p^2 = 0.00). However, a significant two-way interaction effect between violation type and round was observed (*F*(1, 315) = 8.58, *p* < .001, η_p^2 = 0.04). This was followed up by a simple effects analysis on the different levels of round.

Round 5: A one-way (violation type) ANOVA was run on the performance ratings of round 5, resulting in a significant main effect of violation type (*F*(1, 205) = 11.07, *p* = .001, η_p^2 = 0.01). This indicates that an integrity-based violation led to significantly lower performance ratings in round 5 than a competence-based one.

Round 6: The one-way ANOVA run on the performance ratings of round 6 resulted in a nonsignificant main effect of violation type (*F*(1, 205) = 0.46, *p* = .496, η_p^2 = 0.00). This indicates that in round 6 there is no difference in performance ratings between the two violation conditions.

Round 7: The one-way ANOVA run on the performance ratings of round 7 resulted in a nonsignificant main effect of violation type (*F*(1, 205) = 2.59, *p* = .108, η_p^2 = 0.00). This indicates that in round 7 there is no difference in performance ratings between the two violation conditions.

In summary, an integrity-based violation led to significantly lower performance ratings than a competence violation in round 5 (the first round with a violation and repair strategy present), however this effect disappeared in the following rounds. Additionally, the three-way repeated measures ANOVA indicated a significant main effect of repair strategy (*F*(4, 197) = 2.86, *p* = .024, η_p^2 = 0.05), however the post-hoc Tukey analysis did not results in any significant differences between the repair strategies.

Honesty ratings:

In the case of the honesty ratings, neither the three-way interaction between round – violation type – repair strategy was significant (*F*(6, 304) = 0.30, *p* = .941, η_p^2 = 0.00), nor the two-way interactions effects were significant (repair strategy – round: *F*(6, 304) = 1.54, *p* = .164, η_p^2 = 0.03; violation type – round: *F*(1, 304) = 2.41, *p* = .105, η_p^2 = 0.01; violation type – repair strategy: *F*(4, 197) = 2.04, *p* = .090, η_p^2 = 0.04). The main effects, however, were significant. Firstly, a significant main effect of round was found (*F*(1,304) = 27.37, *p* < .001, η_p^2 = 0.12), with the post-hoc Tukey HSD test indicating a significant difference between the honesty ratings of rounds 5 through 7, with the ratings being highest in round 5 (M = 3.90, SE = 0.13), followed by round 6 (M = 3.45, SE = 0.14) and finally, round 7 having the lowest ratings (M = 3.25, SE = 0.14). Secondly, the test resulted in a significant main effect of violation type as well (*F*(1, 197) = 79.36, *p* < .001, η_p^2 = 0.28), suggesting that an integrity violation led to significantly lower honesty ratings. Finally, a significant main effect of repair strategy was also found (*F*(4, 197) = 3.24, *p* = .013, η_p^2 = 0.06). The post-hoc Tukey HSD analysis revealed a significant difference between an apology (M = 2.77, SE = 0.27) and compensation (M = 3.89, SE = 0.29) and an apology and silence (M = 3.99, SE = 0.26), suggesting that the presence of an apology led to lower honesty ratings than a compensation or silence.

2.Objective measures

In the upcoming subsection, we shift our focus to the results of the objective measures of trust.

Score allocation decisions

Hypothesis:

- ***H3:** The presence of repair strategies leads to a higher rate of team score allocation decision.*

To analyse participants' score allocation decisions, the percentage of participants who decided to share their score to the team score was calculated for each round, per condition. The resulting values are illustrated in Figure 16 and are present in Appendix B. Since participants make their score allocation decisions before seeing the choice of the robot, the effect of the violations is visible with a one round delay, that is starting with round 5. Based on Figure 16, it is clear that in the first 4 rounds the large majority of the participants opt to share their score. The few individual choices can be attributed to the reasons detailed in Section 3.5: trying out the workings of the game or not wanting to share a 0 score. Thus, they are not a reflection of selfish behaviour, and do not indicate a lack of trust. Starting with round 5, however, a great reduction in the percentage of team decisions can be observed. This downward trend is most prominent in the integrity-based violation conditions, clearly indicating the negative effect of such a trust violation on teamwork and collaboration.

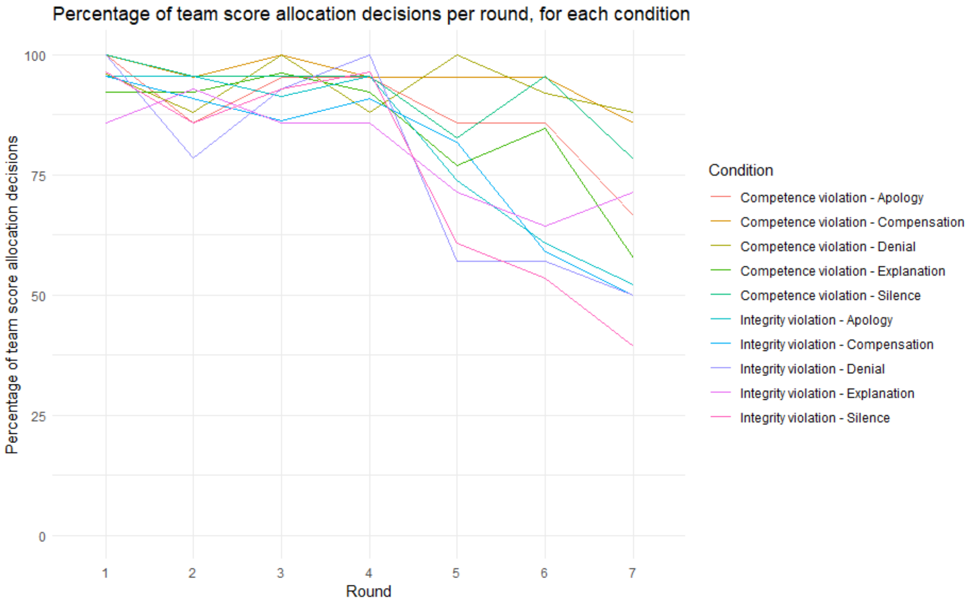


Figure 16: percentage of team decision in each round, per condition

A two-way ANOVA was run on the arrays of percentages to investigate the effect of violation type and repair strategies on the decision rates. The results indicate that there is no significant main effect of repair strategy on the decision rates ($F(4, 60) = 0.20$, $p = .934$, $\eta_p^2 = 0.01$), leading to the rejection of H4. A significant main effect of violation type was found ($F(1, 60) = 12.61$, $p = .000$, $\eta_p^2 = 0.17$), indicating that the percentage of people deciding for the team score is significantly higher following a competence-based violation than an integrity-based one. No significant interaction effect was observed ($F(4, 60) = 0.57$ $p = .681$, $\eta_p^2 = 0.04$).

Willingness to collaborate again

Hypothesis:

- ***H4**: The communicative repair strategies will have a significant impact on the willingness to collaborate again = There is a significant main effect of repair strategy on willingness to collaborate*
- ***H4.a** This effect is moderated by the violation type: the effect of a repair strategy on the different dimensions on trust depends on the type of trust violation that occurred. = violation type moderates the relationship between repair strategy and willingness*

As previously mentioned, trust is a crucial concept in this research. However, the overall aim of investigating trust relations is to foster positive, successful collaborations. Thus, willingness to collaborate again with a robot that violated trust was also measured. The average ratings per condition are presented in Table 15 and visualized in Figure 17. It is visible that the willingness is lower following an integrity-based violation. Moreover, with the exception of the competence-compensation (mean = 5.05), competence-silence (mean = 4.00) and competence-denial (mean = 3.71) conditions, the willingness ratings are below the average value of 3.5.

	Apology	**Compensation**	**Denial**	**Explanation**	**Silence**
Competence violation	3.43 (SD = 2.04)	5.05 (SD = 1.81)	3.71 (SD = 2.12)	2.92 (SD = 2.15)	4.00 (SD = 1.95)
Integrity violation	1.77 (SD = 1.23)	2.84 (SD = 1.92)	3.07 (SD = 1.98)	2.64 (SD = 2.24)	2.62 (SD = 1.81)

Table 15: Willingness to collaborate again with the robot, mean scores

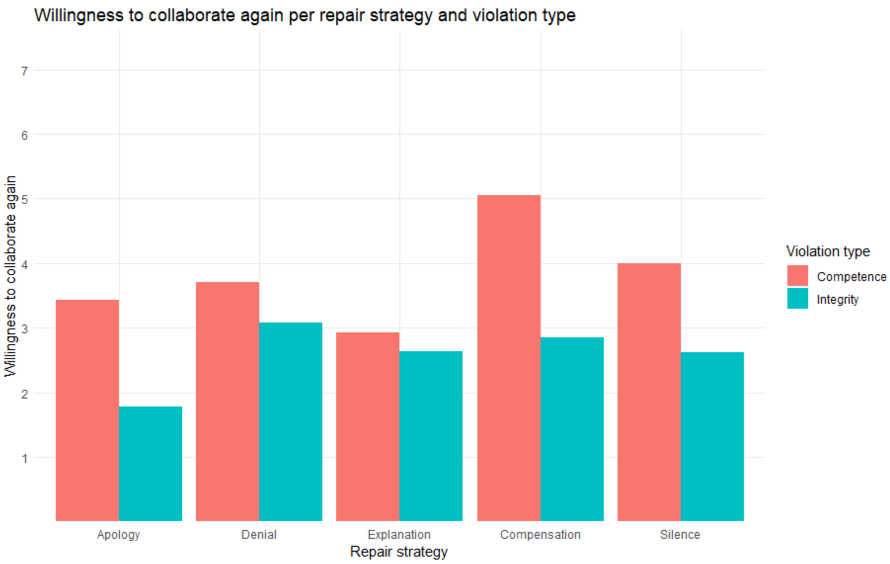


Figure 16: Willingness to collaborate again with the robot

Running a two-way ANOVA test revealed that these differences are significant: a significant main effect of violation type ($F(1, 207) = 18.31$, $p < .001$, $\eta_p^2 = 0.09$) and a significant main effect of repair strategy ($F(4, 207) = 4.36$, $p = .002$, $\eta_p^2 = 0.06$) were found. These results suggest that an integrity-based violation has a more negative effect on willingness: participants expressed lower levels of willingness to collaborate again with a robot that committed an integrity violation, than with one that violated performance trust. Moreover, H4 is supported: the repair strategies do significantly influence willingness to collaborate again. The post-hoc

Tukey-HSD test revealed that a compensation (M = 3.95, SE = 0.31) has a more positive impact on willingness than an apology (M = 2.60, SE = 0.29). No significant interaction effect was observed (*F*(1, 207) = 57.24, *p* < .001, η_p^2 = 0.2), rejecting H6a.

4.4 Humanlikeness

To better understand the intricacies and dynamics of the relationship between trust, trust violation, trust repair and a humanlike perception of robots, the warmth and discomfort subdimensions of the RoSAS scale [8], together with the human nature subdimension of Haslam's dehumanization scale [33] were used as measures. An analysis of the results obtained follows.

RoSAS warmth and discomfort

Hypotheses:

- ***H5.1a**: The communicative repair strategies will have a differing impact on the different dimensions of RoSAS.*
- ***H5.1b:** This effect is moderated by the violation type: the effect of a repair strategy on the different dimensions of RoSAS depends on the trust violation that occurred.*

The mean warmth and discomfort scores per condition are presented in Table 16. Based on these values and their visual representation (see Figures 18, 19), the following patterns can be observed: in the case of a competence violation, the warmth scores are larger than the discomfort ones, suggesting a positive perception of the robot. However, in the presence of an integrity violation, this tendency is inversed, with the discomfort scores being greater than the warmth ones. This indicates a more negative perception of the robot, and therefore a higher severity of an integrity-based violation, compared to a competence-based one. The histograms of the scores (Figure 19) reveal that the majority of the scores are below the average value of 3.5, indicating a generally lower humanlike perception of the robot teammate.

	Apology	**Compensation**	**Denial**	**Explanation**	**Silence**
Competence violation	2.98 (SD = 1.42)	3.38 (SD = 1.37)	3.44 (SD = 1.18)	2.54 (SD = 0.97)	2.60 (SD = 0.80)
Integrity violation	2.21 (SD = 0.86)	2.76 (SD = 1.12)	2.50 (SD = 0.80)	2.64 (SD = 1.31)	2.85 (SD = 1.21)

Table 16.a: mean RoSAS warmth scores

	Apology	**Compensation**	**Denial**	**Explanation**	**Silence**
Competence violation	2.65 (SD = 1.43)	1.63 (SD = 0.91)	2.69 (SD = 1.31)	2.37 (SD = 1.09)	1.76 (SD = 0.60)
Integrity violation	2.55 (SD = 0.85)	2.97 (SD = 1.38)	3.37 (SD = 1.45)	2.96 (SD = 1.33)	3.08 (SD = 0.98)

Table 16.b: mean RoSAS discomfort scores

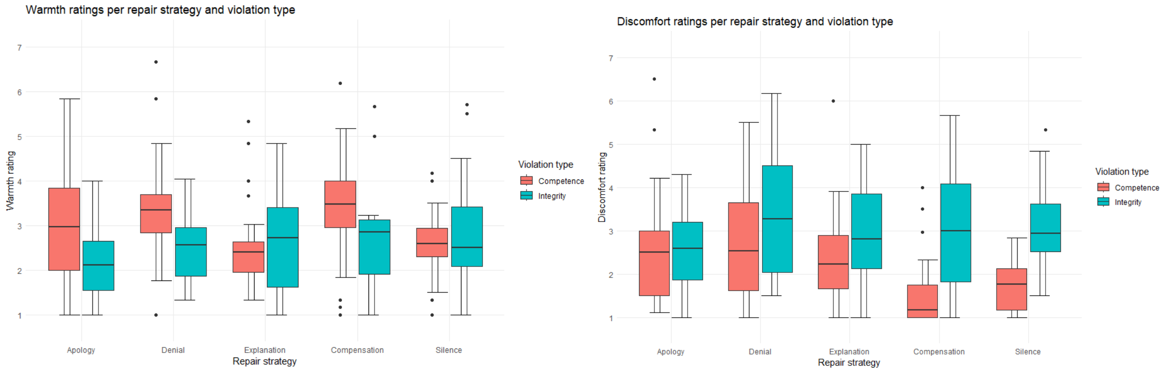


Figure 18: boxplots of RoSAS warmth (left) and RoSAS discomfort (right) scores, per repair strategy and violation type

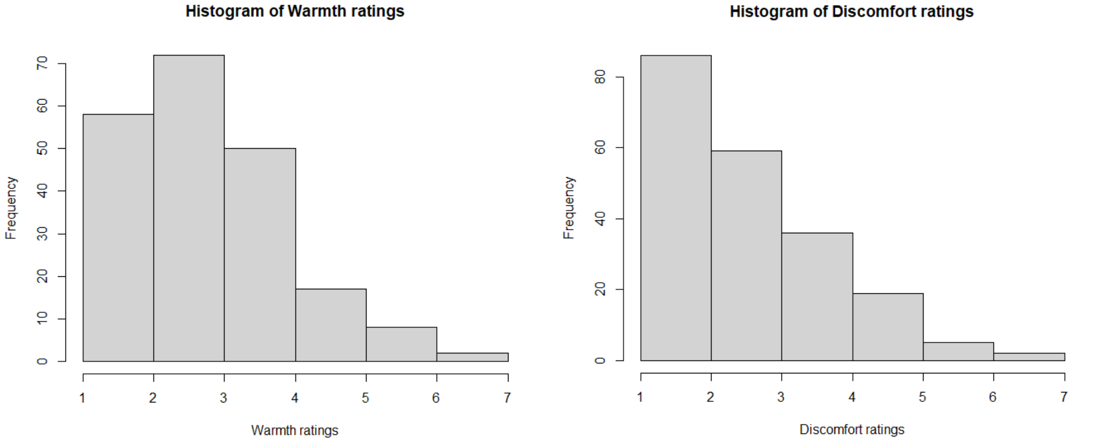


Figure 19: histograms of RoSAS warmth (left) and RoSAS discomfort (right) scores, per repair strategy and violation type

To analyse the differences in RoSAS warmth and discomfort ratings between the conditions, a two-way MANOVA was run on these values. It must be noted that the homogeneity of covariance matrices was violated, however the MANOVA is a test that is robust against such violations. The presence of the violation, together with the large number of outliers (see Figure 18, boxplots) must nevertheless be considered during interpretation. The results of the test indicate a significant interaction effect between violation type and repair strategy on the combination of RoSAS warmth and RoSAS discomfort scores (V = .09, (F(4, 394) = 2.59, p = .009, η_p^2 = 0.05), illustrated in Figure 20. This was followed up by a simple effects analysis on the different levels of violation type.

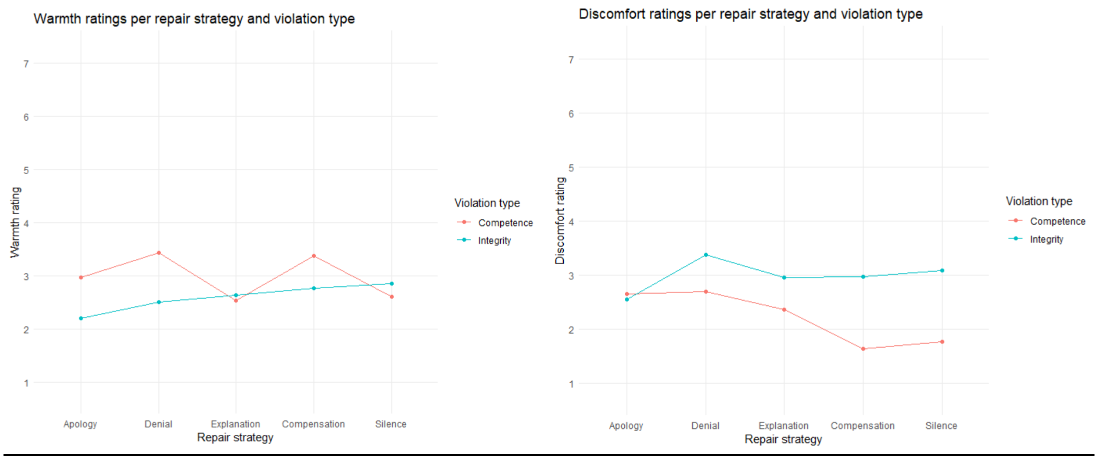


Figure 20: interaction graphs of RoSAS warmth (left) and RoSAS discomfort (right) scores, per repair strategy and violation type

Level of competence violation: A one-way MANOVA was run to analyse the effect of the repair strategies on the combination of RoSAS warmth and RoSAS discomfort in the case of a competence violation. The result indicates a significant main effect of repair strategy on this measure ($V = 0.242$, $F(4, 214) = 3.68$, $p = .000$, $\eta_p^2 = 0.12$). Post-hoc univariate ANOVAs were run as a follow-up. The results regarding the effect of repair strategy on RoSAS warmth indicated a significant main effect ($F(4, 107) = 3.02$, $p = .020$, $\eta_p^2 = 0.04$). The post-hoc Tukey-HSD failed to find any significant difference in the RoSAS warmth rating in the different repair strategy conditions, at the level of competence violation. The results regarding the effect of repair strategy on RoSAS discomfort indicated a significant main effect ($F(4, 107) = 4.17$, $p = .003$, $\eta_p^2 = 0.06$). The post-hoc Tukey-HSD test found a significant difference between a denial (M = 2.69, SE = 0.23) and a compensation (M = 1.63, SE = 0.25) and between compensation and apology (M = 2.65, SE = 0.24). This suggests that a compensation led to significantly lower levels of RoSAS discomfort than a denial or an apology, following a competence-based violation.

Level of integrity violation: A one-way MANOVA was run to analyse the effect of the repair strategies on the combination of RoSAS warmth and RoSAS discomfort in the case of an integrity violation. The results indicates no significant main effect of repair strategy on this measure ($V = 0.093$, $F(4, 180) = 1.10$, $p = .362$, $\eta_p^2 = 0.05$), indicating that there is no difference in the combination of the RoSAS warmth and RoSAS discomfort scores between the repair strategy conditions, at the level of an integrity-based violation.

Human nature traits

Hypothesis:

- ***H5.2a**: The communicative repair strategies will have a significant impact on the Human Nature subdimension of humanlikeness.*
- ***H5.2b** This effect is moderated by the violation type: the effect of a repair strategy on the Human Nature subdimension depends on the trust violation that occurred.*

To investigate the relationship between repair strategies, violation types and Human nature traits, the "Human nature" subscale of Haslam et al.'s Dehumanization scale was used. The mean scores per condition are presented in Table 17. Items were rated on a 7-point Likert scale, thus the current results indicate a fairly low attribution of human nature traits to the robot, with the highest rating being 3.10 in the competence-denial condition.

	Apology	**Compensation**	**Denial**	**Explanation**	**Silence**
Competence violation	2.78 (SD = 0.94)	2.93 (SD = 0.67)	3.10 (SD = 0.84)	2.81 (SD = 0.72)	2.46 (SD = 0.64)
Integrity violation	2.05 (SD = 0.75)	2.58 (SD = 0.63)	2.96 (SD = 0.69)	2.54 (SD = 1.36)	2.76 (SD = 0.87)

Table 17: mean ratings of "Human nature" traits

Analysing the boxplot and histogram of the scores (see Figure 21), multiple observations can be made. Firstly, a greater difference between the two violation types in the case of an apology is apparent, with higher levels of human nature traits in the case of a competence violation paired with an apology. This indicates that in the presence of an apology, a competence-based violation led to a more humanlike perception of the robot than an integrity-based one. Secondly, the majority of the participants' ratings fall below the average value of 3.5 (see Figure 21), suggesting a generally low perception of humanlikeness of the robot teammate. Finally, a large number of outliers can be observed.

A two-way ANOVA was conducted to analyse the differences between the mean "human nature" scores across the conditions. The results revealed a significant main effect of repair strategy (*F*(4, 207) = 3.35, *p* = .011, η_p^2 = 0.06), supporting H7.2a. The post-hoc Tukey-HSD test suggests the presence of a significant difference between an apology (M = 2.42, SE = 0.13) and a denial (M = 3.03, SE = 0.14): the presence of a denial after a trust violation led to significantly higher levels of "human nature" traits assigned to the robot than an apology. However, due to the highly variable nature of the data, one must be cautious in drawing conclusions from this effect. No significant main effect of violation type (*F*(1, 207) = 2.76, *p* = .097, η_p^2 = 0.02) and no significant interaction effect between violation type and repair strategy (*F*(4, 207) = 1.97, *p* = .099, η_p^2 = 0.05) was observed; thus, H7.2b is rejected.

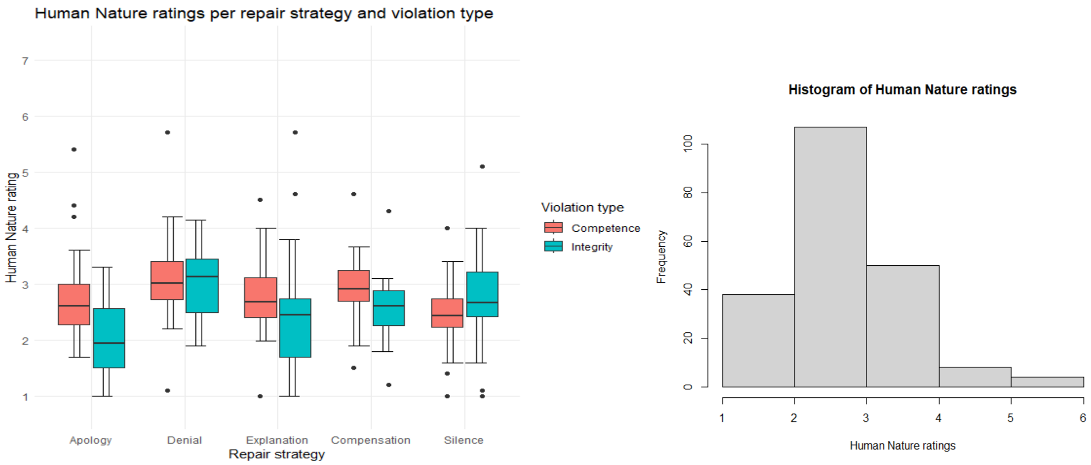


Figure 21: Boxplot of mean "Human nature" ratings per repair strategy and violation type (left); Histogram of "Human nature" ratings (right)

Humanlikeness overall conclusions

To summarize, a significant interaction effect between repair strategy and violation type was observed on the combination of RoSAS warmth and RoSAS discomfort scores. Upon further investigation, it was found that following a competence violation, a compensation led to higher levels of RoSAS discomfort scores. Regarding the Human Nature traits, a denial led to increased ratings. In summary, these findings, together with the large number of outliers and non-normal distribution of the data, suggest that there was a high variability in participants' perception of the robot's social attributes and humanlikeness.

5. Discussion

This research aims to investigate the effects of a wide range of repair strategies on trust, following competence- and integrity-based violations, within the context of collaboration in human-robot teams. During the course of playing an online game with a robot teammate, Pepper, moral and performance trust levels, participants' willingness to collaborate with Pepper again and their perceptions of Pepper's humanlikeness were measured. The section starts with a discussion on the effects of trust violations, followed by the implications of the effects of repair strategies (they are treated separately, because no significant interaction was found between these two variables). Additionally, the dynamics on the humanlike perception of the robot is outlined. Finally, the limitations of the study are presented, together with suggestions for future work.

5.1 Effect of trust violations

Two types of trust violations were defined in our research - a competence-based and an integrity-based violation -, and our analysis focused on their effect on the two respective dimensions of trust – performance trust and moral trust. The results of the replication and our main study align; thus, they will be discussed together.

The current body of research on trust in HRI concludes that an integrity violation has a more severe impact on moral trust, compared to a performance one [44, 49, 41]. Our findings also support this line of reasoning, indicating that there is a clear difference in the judgment of a violation, depending on its type. When the violation is perceived as a performance based one, people seem to not ascribe any moral dimension to it. A performance-based violation is not seen as a representation or expression of the robot's integrity, it is simply "an honest mistake". In the case of integrity-based violations, however, the interpretation lacks lenience. Already the first occurrence of such a violation leads to a reduction in moral trust, and with each round of violation, the trust gets lower and lower.

Surprisingly, our results indicate that the same trend does not hold for the effect on performance trust: no significant difference was found between the effects of the two types of violation on performance trust. This contradicts previous results that suggest that a performance violation should lead to lower scores of performance trust than an integrity violation, and does not align with claims stating that there is a measurable difference in the impact of the two violation types on this subdimension of trust [44, 49]. Moreover, the analysis of the end-of-round performance ratings revealed that in round 5 of the game, that is, right after the trust violation and the first repair message, participants rated the performance of the robot lower following an integrity violation. A potential explanation for these findings could lie in the meaning of the concept of performance itself [27]. In the case of competence violations, the *robot's performance* keeps its usual meaning, which is closely tied to the functional efficacy of the robot. This perspective emphasizes the robot as a tool, focusing on its role in accomplishing tasks effectively. However, in the context of integrity violations, there is the possibility of a perceptual shift, where *performance* takes on a new connotation, aligning with the robot's role as a social entity and teammate. In this light, *performance* transcends mere functionality and extends to the robot's ability to engage effectively in social interactions and maintain trustworthiness within the team. Following such a perception, the integrity-based violation does indeed correspond with a bad performance of the robot as a teammate and can be viewed as "as a degradation of a user's perception of a robot's socio-affective competence" [56]. This shift in the interpretation of *performance* introduces a layer of complexity to our understanding of trust dynamics in the context of human-robot interactions, that should be further explored.

The effect of the violations on the different subscales of the two trust dimensions were also analysed during the replication. A performance violation led to significantly lower ratings of the *competence* subscale of performance trust, consistent with Robinette et al.'s findings. However, this effect was not present when taking the *reliable* subscale into account. A potential reason for this contradiction might lie in the perception of reliability. Malle and Ullman conceive of reliability as a component of the performance dimension of trust [34]. However, a behaviour that repeatedly violates moral expectations can also be regarded as reliable, due to its consistent nature [36]. The repeated nature of the violations, regardless of their type, contributes to a consistent and predictable view of Pepper. An exploration of participants' comments indicated that by the final rounds, participants were expecting the robot's violating behaviour. Through the repetition of the violations, Pepper became reliable: a reliably bad teammate, regardless of the type of trust that it violates.

As part of the replication, overall trust scores were also analysed. In contrast to Robinette et al.'s results - hypothesizing that the integrity violation will have a more damaging effect overall [44] - no significant differences between the two violation types were found on overall trust. It must be noted that this does not mean that the violations did not decrease trust, rather it indicates that there is no difference in the magnitude of this decrease in trust. When taking into consideration the multidimensional conception of trust present throughout this research, a potential explanation for this discordance becomes apparent. The overall trust measure captures both dimensions of trust (performance- and moral trust), it being an average of the two measures. Therefore, any fine-grained differences between the effects on the different dimensions are being lost, with the measure not providing an indication of the actual differing impact of the different violations. Moreover, what is of interest in better understanding the relationship dynamics within HRI, is a thorough analysis of the fine-grained interactions and mechanisms in place at the interplay of the different violation types and their effect on the different subdimensions of trust. Such a granular approach can lead to a deeper, more nuanced understanding of trust in HRI.

The ultimate goal of a positive trust relationship is to ensure successful collaborations [12, 13]. One measure that reflected participants' willingness (or lack thereof) to collaborate with Pepper was their score allocation decision. Our data indicates a noticeable pattern, in line with the findings of Robinette et al. [44, 49, 31]. Significantly less participants chose to share their score with the team following an integrity violation than following a competence-based one. Moreover, whilst the decrease in sharing behaviour is more gradual in the case of a competence violation, an integrity one results in people refusing to contribute to the team score even after its first occurrence. No second chances were given to the robot to redeem itself after such a violation, as opposed to the case of the competence one, where participants were seemingly more lenient. This pattern was also reflected in the results of the one-item measure directly asking participants to rate their willingness to work with this robot again in the future, where the ratings in the integrity violation conditions were significantly lower than those in the performance violation.

In conclusion, our findings are in accordance with previous research [44, 49], indicating that an integrity violation has a more severe impact on moral trust and willingness to collaborate again compared to a competence violation. However, our results differ from the current body of research when looking at the violations' effect on performance trust, finding no difference between the two violation types. A more in-depth study of the perception of the *robot's performance* in the different violation contexts is required.

5.2 Effect of repair strategies

This study set out to explore the effects of five repair strategies: apology, denial, explanation, compensation, and silence. Previous research in this area compared only a limited range of potential strategies, resulting in mixed findings lacking consensus on the workings of these repair strategies within the dynamics of trust repair in HRI [17, 18]. Our results, with absence of significant interaction effects between repair strategy and violation type, indicate that the effectiveness of a given strategy does not depend on the type of violation that occurred in the context of a collaborative game. This contradicts de Graaf & Liefooghe [11] and Sebo et al. [49], who argued that an apology is better suited following violations of performance trust, whilst a denial is more effective in the case of integrity violations. A significant main effect of repair strategy was, however, consistently found across all our analyses. Considering the measures of performance trust, honesty ratings, as well as willingness to collaborate again, a *compensation* led to significantly higher ratings than any other repair strategy, being in accordance with existing literature. Lee et al. found that a compensation outperformed an apology in repairing trust and ratings of customer satisfaction (however, this effect was moderated by participants' orientation towards service, with a utilitarian orientation leading to a higher performance of compensation) [31]. This finding highlights the key role people's attitudes towards the context of the interaction play in the trust dynamics. It also suggests that the fact that a compensation outperformed an apology in both the competence-based and integrity-based violation conditions could be indicative of participants having a utilitarian perception of the robot's role in the team, with the violation type having no influence on this perception. Moreover, it underscores the applicability of compensation in varied interactive contexts.

No other pairwise comparisons were significant, indicating a lack of difference in the effect of the other repair strategies on trust and willingness to collaborate again. A possible explanation for a compensation leading to higher trust ratings, across both violation types, combined with a lack of difference in the effect of the other strategies, could lie in the nature of the specific violations that occurred in this study. In both violation conditions, participants experienced concrete losses: the violations resulted in a lack of points, and a reduction in the feasibility of achieving the team bonus. In the context of the performance violation, the reduction was evident, leading to a team score of zero in the respective round. However, within the integrity violation, the team score gets nullified in the respective round as well, through the robot's individual score allocation. This fact invalidates previous collaborative efforts made towards a shared goal, since with each repeated violation, achieving the team bonus becomes less possible. Just as in the case of a competence violation, a monetary loss is incurred. Consequently, the rationale behind the appropriateness of compensation as a repair strategy becomes evident, as it directly addresses the specific losses incurred due to the violation [43, 33].

These findings suggest that a compensation is the most promising attempt the robot could make at repairing the broken trust in this collaborative context, regardless of the violation type that occurred.

5.3 Effects on humanlikeness

Perceptions of a robot's humanlikeness are intrinsically linked to trust and willingness to collaborate again with it [63]. This highlights the importance of exploring the effects of trust violations and repair strategies on such perceptions, within the context of human-robot collaboration.

On Haslam's "Human nature" traits, the denial strategy achieved significantly higher ratings than an apology, which is in line with previous research [54]. Denial being perceived as more humanlike can be attributed to the "self-serving bias", that states that blaming others for our failures is seen as more humanlike [37]. Moreover, Biswas and Murray found that a robot displaying such self-serving behaviour elicited greater degrees of humanization [4].

Regarding the RoSAS scores, it was found that the effect of the repair strategies is indeed moderated by the violation type. Following integrity violations, there was no difference between the different repair strategies. A potential explanation of this finding may lie in the fact that the integrity violation could have already had an increasing impact on the social perception of the robot. Van der Woerdt & Haselagerand, and Short et al. found that a robot committing an integrity violation is perceived as more humanlike, with "participants displaying a greater level of social engagement" when interacting with a cheating robot [59, 51]. Therefore, it is possible that the presence of the repair strategies does not further enhance or differentially impact this perception. A deeper exploration of the effect of repair strategies on the social perception of a robot teammate, specifically in the context of an integrity violation is required to better understand this phenomenon.

In the case of competence violations, a compensation resulted in significantly lower RoSAS discomfort levels. Providing a compensation led to higher trust levels, additionally also having an increasing effect on willingness to collaborate again. Based on this trend, it appears to be a logical consequence that participants ascribed lower levels of discomfort towards the robot using this repair strategy. Its positive effect in a service setting was already explored by Lee et al., who found that a compensation led to higher ratings of customer satisfaction following what can be regarded as a competence violation in the context of the robot fulfilling participants' order in a restaurant [31]. Our finding thus further strengthens the positive potential of using compensations in a collaborative context as well.

Another notable finding was the large number of outliers and large variance in both RoSAS ratings and the "Human nature" trait ratings. This could be indicative of a larger uncertainty and inconsistency in people's perception or understanding of a robot's humanlikeness, and is suggestive of the complexity of the expectations people have regarding the emotional and intentional capacities of collaborative robots. Spatola et al. mention the existence of several determinants that impact people's perceptions of robots, such as socio-cognitive factors (conformism, intra-group bias) and individual factors (e.g. attitude towards robots) [53] , whilst Thellman et al. also evidentiate the effects of age and individual motivation on such perceptions [54]. Additionally, the comments of participants submitted at the end of the experiment were explored to get a better understanding of their perception of Pepper and their experience of interacting with it. No mentions of Pepper's social or humanlike nature were present. However, one participant explicitly expressed that "I always find it mildly annoying when asked to rate the emotions or intentions of something programmed for a particular task, or even asked to interpret its goals without being told anything about its programming.", indicating the challenge and complexity of studying the humanlike perception of robots in such a context. Alarcon et al. also noted this challenge, attributing it to the fact that "automated systems lack intentionality—or genuine motivation to prioritize the best interest of the trustor—but instead embody the intentions of the designer" [1].

Finally, whilst in recent times the number of studies has been steadily increasing, the field of trust repair in HRI is still in its infancy. Therefore, there is a lack of comprehensive research analysing the interactions between trust violations, repair strategies and humanlikeness of robots. Future research is needed to place our findings into a broader theoretical framework of human-robot collaboration.

5.4 Broader implications

Throughout this study the importance of trust repair was continuously highlighted. However, the magnitude of the repaired trust must be carefully considered. Blindly optimizing for maximal trust is not always desirable. De Visser et al. introduce in their work the notion of trust calibration, which they define as restoring the trust to an *appropriate level* [12]. This is of great importance, since both over- and undertrust can lead to inefficient, and potentially dangerous situations. In the case of overtrust, the "trustor trusts the trustee to a greater extent than deserved given the trustee's true capabilities'' [12], thus overrelying on the robot and potentially "ignore signs of malfunction" [46] or allow it to "act autonomously even in situations where the trustee is not capable of performing the task adequately" [12]. On the other hand, in the case of undertrust, the entire goal of the collaboration is threatened, since "the trustor fails to take full advantage of the trustee's capabilities" [12] and this distrust in the robot's competence leads to inefficient work. Salem et al. illustrate this issue with the concrete example of a patient not willing to follow the robot's advice in a medical setting and not take their medicine on time [46]. Therefore, it becomes clear that a balanced approach to trust repair that prevents both overtrust and undertrust is crucial, ultimately fostering more efficient and effective human-robot collaborations. This requires accurately assessing the robot's performance and integrity, and clearly communicating its limitations to users.

In the context of human-robot interactions, it is crucial to focus on understanding how people perceive trust violations, uncovering the specific emotions and reactions these violations may trigger, such as feelings of betrayal, deception, or inconvenience. In their work, Yasuda et al. found that norm violating behaviour (realized in their experiment by the robot cheating during a rock-paper-scissors game) evoked strong emotional reactions from participants [62]. Moreover, Pompe et al. note that certain participants reacted with shock to the robot misunderstanding them and not acting in a trustworthy manner [41], and an exploration of the comments left by participants in this current study revealed participants' increasing frustration caused by the trust violations. By delving into these emotional and psychological aspects, we can tailor trust repair strategies to directly address the underlying causes of trust violations. This emphasis on understanding the nuances of trust repair is further underscored by our findings, which reveal the severe impact of integrity violations on trust and willingness to collaborate again in these interactions. Consequently, this highlights the importance of designing robots that not only fulfil their functional roles but also seamlessly integrate into our social fabric by adhering to moral and ethical standards, since indeed, people do perceive robots as social agents beyond their utilitarian performance [28, 8].
The initial and fundamental step in creating such robots is to gain a deep understanding of people's perceptions of the social norms and expectations that encompass their interactions with these robots. By taking these implications into account, robots can be designed that foster and maintain trust in various human-robot interaction contexts, and that allow for successful and efficient collaborations.

5.5 Limitations

Whilst the research resulted in valuable insights into trust dynamics within HRI, the study was not without limitations.

Firstly, as the adoption of robots in various contexts continues to grow [34], it is crucial to consider the potential influence of cross-cultural differences in attitudes and expectations regarding human-robot interactions. Research has shown that cultural factors can significantly impact trust and cooperation dynamics [12, 34] ; for example, people from the United States were observed to hold less sceptical attitudes about the potential cognitive and emotional capacities of robots [10]. However, the influence of culture extends beyond nationality, encompassing a broad spectrum of factors such as age, economic background, and access to technology. For instance, our study primarily involved participants with access to the internet and a computer, which might not accurately represent the most widespread target group of social robots (children and older adults). It is important to acknowledge that these demographic factors can affect perceptions of technology and trust dynamics [39, 3]. To address these considerations comprehensively, further research should strive to involve a more diverse participant population that includes individuals from different age groups, economic backgrounds, and technological access, thus enhancing the generalizability of findings. Henrich et al. have highlighted the importance of moving beyond WEIRD samples [26] currently relied on. Additionally, it is critical to emphasize the importance of testing social robots with the target group of their intended use [45, 55]. The effectiveness of trust repair strategies can vary significantly depending on the specific application of the robot, and involving end-users who interact with these robots in their daily lives is essential for producing findings that are directly applicable to the intended contexts. Moreover, besides cultural aspects, the individual differences among participants can also introduce significant variability. Factors such as their familiarity with robots, prior experiences with technology, and even personality traits like propensity to trust or remorsefulness [11, 27, 49, 31, 33, 16] may influence their responses to trust violations and the effectiveness of repair strategies. The understanding of the effect of such individual differences and their interplay with trust dynamics requires additional exploration.

Secondly, the controlled nature of our study, utilizing a task with lower personal relevance and minimal stakes in an online setting, may not fully capture the complexity of real-life human-robot interactions. It is important to acknowledge that the nature of the task and its personal relevance can substantially affect human-robot interactions [2]. Tasks of varying personal significance, such as those involving critical healthcare decisions or workplace responsibilities, may yield different responses to trust violations and repair strategies. Furthermore, our study was primarily designed to investigate short-term, momentary collaborations with robots. The dynamics of trust, trust violations and trust repair may significantly differ in long-term relationships with robots, such as companion or caregiving robots. Whilst participants were asked to rate their willingness to collaborate again with their robot teammate, this measure solely does not provide a representative view of long-term dynamics. Therefore, the conclusions drawn from this study should not be broadly applied to extended human-robot relationships, but serve as a lens into the complex workings of trust violation and repair in short term, temporary human-robot collaborations.

Thirdly, this research examined the impact of repair strategies across four consecutive rounds containing trust violations. The robot’s actions did not change during the four violation rounds, regardless of the presence of the repair messages. Therefore, participants' responses to these strategies may be influenced by their perception of them as "empty words" when trust violations persist without observable improvement in robot behaviour.

Finally, it is important to recognize that our use of an image of a virtual Pepper robot in the study differs from an ideal scenario involving a video of a real-life Pepper robot. The choice of media may have influenced participants' perceptions and reactions to the robot, warranting consideration in future research. The embodiment and physical interaction with a robot play a crucial role in shaping human-robot trust dynamics [48]. Real-life robots possess physical attributes and capabilities that the virtual image may not fully capture, such as facial expressions and gestures. In this context, the lack of these physical elements in our study might have impacted how participants formed and repaired trust with the robot. Future research should consider incorporating more immersive and physically interactive representations of robots to gain a deeper understanding of the influence of embodiment on human-robot trust dynamics.

In summary, while our study provides valuable insights into communicative repair strategies and their effects on trust following trust violations in human-robot interactions, these limitations underscore the need for continued investigation and the acknowledgment of various factors that can influence the outcomes and generalizability of our findings. Addressing these limitations could lead to a more comprehensive understanding of human-robot interactions and trust dynamics.

5.6 Future Work

As trust repair strategies in human-robot interactions become more tailored to individual needs and preferences, future research can delve deeper into the influence of various human factors. Individual differences, such as personality traits [2, 58] and cultural backgrounds [2, 58], can significantly impact trust dynamics. In-depth studies can investigate how factors like propensity to trust, risk aversion, or cultural expectations influence both the perception of trust violations and the effectiveness of trust repair strategies [48]. Incorporating these factors as covariates in the analysis can provide a more comprehensive understanding of trust repair, allowing for the development of more personalized and culturally sensitive approaches.

While online experiments offer a controlled environment for research purposes, the transition to real-life settings is essential for bridging the gap between laboratory findings and practical applications. Conducting studies in authentic environments, such as healthcare facilities, educational institutions, or workplaces, will enable researchers to capture the nuances of human-robot interactions and trust repair as they occur in everyday life. This shift toward real-life settings can provide valuable insights into the practical challenges and opportunities [55, 9, 61], in this case in applying trust repair strategies, ultimately leading to more effective human-robot collaborations. Replicating the current study in a lab setting with a physically present robot denotes a potential first step towards such a shift.

Compensation, as a trust repair strategy, requires extensive exploration in future research. In particular, research should investigate how compensation strategies influence trust, as well as how they may affect participants' experience and overall satisfaction. Exploring the interplay between compensation and trust dynamics in real-life settings can provide a more nuanced understanding of this strategy's effectiveness.

Trust violations related to integrity are complex and multifaceted. To gain a deeper understanding of how people perceive integrity violations and to develop effective repair strategies, researchers should focus on realistic scenarios in which robots may act in ways that challenge integrity. Investigating the triggers and manifestations of integrity violations in various real-life contexts beyond collaborative games, such as healthcare, education, or customer service, can inform the development of more targeted repair strategies. Understanding people's perception of a robot's morality or integrity is crucial to developing trust repair approaches that align with societal expectations and norms.

Additionally, future research should delve deeper into the complex ways in which trust violations influence the perception of a robot's performance. Our findings, indicating that integrity violations not only impact moral trust but also performance trust, emphasize the complexity of this relationship. Therefore, it is crucial to explore how perceptions of performance are altered in the context of different violation types. This exploration can provide a more nuanced understanding of the dynamic between trust violations and the robot's perceived performance, ultimately enabling the design of more effective strategies to repair trust in human-robot interactions.

6. Conclusion

In the growing field of human-robot interaction, trust plays a pivotal role in establishing successful and efficient collaborations. As robots increasingly become integral components of various aspect of our lives, from healthcare to education and beyond, understanding the intricacies of trust dynamics is crucial. This research undertook a systematic exploration of the effects of different types of trust violations and communicative repair strategies within the context of human-robot collaboration, by employing a 2x5 between-subjects online study design. The aim was to answer the following research question: "*How does the type of trust violation followed by different forms of repair strategies affect trust in and humanlikeness of a collaborative robot?"*

First, the effect of competence- and integrity violations was investigated. Consistent with prior research, we found that integrity violations have a more severe negative effect on moral trust and a person's willingness to collaborate with a robot that committed such a violation. Surprisingly, integrity violations also impacted performance trust to the same degree as the competence violation. This challenges our understanding of the dynamics of trust and its violations, highlighting the importance of exploring how perceptions of performance differ in the context of the different violation types.

In examining repair strategies, the effectiveness of five distinct approaches were analysed: apology, denial, explanation, compensation, and silence. Our results indicate that the type of violation does not significantly moderate the effectiveness of these repair strategies. Notably, compensation emerged as an effective repair strategy, leading to higher trust ratings across both trust dimensions, higher willingness to collaborate again, and lower levels of discomfort. This suggests that, in short-term collaborative contexts, compensating for losses due to trust violations can have a particularly positive impact.

Finally, the influence of the trust violations and repair strategies on people's perception of the robot's humanlikeness were also investigated. Denial was perceived as more humanlike than an apology, aligning with prior research. Moreover, a compensation was associated with lower ratings of discomfort following a competence violation, further emphasizing the potential of this repair strategy in fostering successful collaborations.

In light of our findings that underscore the substantial impact of integrity violations on trust and willingness to collaborate, it's clear that robots must not only perform well but also align with moral and ethical standards to foster trust. By embracing these insights, robots that promote trust in a wide range of human-robot interaction contexts can be designed and deployed, ensuring effective and successful collaborations.

References

[1] G. M. Alarcon, A. Gibson, S. A. Jessup, and A. Capiola, “Exploring the differential effects of trust violations in human-human and human-robot interactions,” *Applied Ergonomics*, vol. 93, p. 103350, 5 2021. [Online]. Available: https://doi.org/10.1016/j.apergo.2020.103350

[2] A. J. Baker, E. J. Phillips, D. Ullman, and J. R. Keebler, “Toward an Understanding of Trust Repair in Human-Robot Interaction,” *ACM transactions on interactive intelligent systems*, vol. 8, no. 4, pp. 1–30, 11 2018.

[3] L. Bishop, A. Van Maris, S. Dogramadzi, and N. Zook, “Social robots: The influence of human and robot characteristics on acceptance,” *Paladyn*, vol. 10, no. 1, pp. 346–358, 1 2019. [Online]. Available: https://doi.org/-10.1515/pjbr-2019-0028

[4] M. Biswas and J. Murray, *Robots that refuse to admit losing - a case study in game playing using Self-Serving bias in the humanoid robot MARC*, 1 2016. [Online]. Available: https://doi.org/10.1007/978-3-319-43506-0_47

[5] D. Cameron, S. De Saille, E. C. Collins, J. M. Aitken, H. Cheung, A. Chua, E. C. Loh, and J. Law, “The effect of social-cognitive recovery strategies on likability, capability and trust in social robots,” *Computers in Human Behavior*, vol. 114, p. 106561, 1 2021.

[6] D. Cameron, E. J. Loh, A. Chua, E. Collins, J. M. Aitken, and J. Law, “Robot-stated limitations but not intentions promote user assistance,” 2016.

[7] C. L. Corritore, B. Kracher, and S. Wiedenbeck, “On-line trust: concepts, evolving themes, a model,” *International journal of human-computer studies*, vol. 58, no. 6, pp. 737–758, 6 2003.

[8] K. Dautenhahn, S. Woods, C. Kaouri, M. Walters, K. L. Koay, and I. Werry, “What is a robot companion - friend, assistant or butler?” in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 1192–1197.

[9] M. M. De Graaf, S. B. Allouch, and J. A. Van Dijk, “Long-term evaluation of a social robot in real homes,” *ResearchGate*, 4 2014. [Online]. Available: https://www.researchgate.net/publication/260599142_Long-term_evaluation_of_a_social_Robot_in_Real_Homes

[10] M. M. De Graaf, F. Hindriks, and K. V. Hindriks, “Who wants to grant robots rights?” *Frontiers in Robotics and AI*, vol. 8, 1 2022. [Online]. Available: https://doi.org/10.3389/frobt.2021.781985

[11] M. De Graaf and B. Liefooghe, “Communicative strategies to repair trust after failure.” 2023, workshop "Imperfectly relatable robots" at the International Conference on Human-Robot Interaction, Stockholm, Sweden.

[12] E. J. De Visser, M. M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, “Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams,” *International Journal of Social Robotics*, vol. 12, no. 2, pp. 459–478, 1 2020.

[13] K. T. Dirks, P. S. Kim, D. L. Ferrin, and C. D. Cooper, “Understanding the effects of substantive responses on trust following a transgression,” *Organizational Behavior and Human Decision Processes*, vol. 114, no. 2, pp. 87–103, 3 2011.

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

[15] C. Esterwood and L. P. Robert, “Do You Still Trust Me? Human-Robot Trust Repair Strategies,” *Robot and Human Interactive Communication*, 8 2021.

[16] C. Esterwood and L. P. Robert, “A Literature Review of Trust Repair in HRI,” *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 8 2022. [Online]. Available: https://doi.org/-10.1109/ro-man53752.2022.9900667

[17] C. Esterwood and L. P. Robert, “Having the Right Attitude: How Attitude Impacts Trust Repair in Human-Robot Interaction,” *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 3 2022. [Online]. Available: https://doi.org/10.1109/hri53351.2022.9889535

[18] 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, p. 107658, 1 2023. [Online]. Available: https://doi.org/10.1016/j.chb.2023.107658

[19] R. Fehr and M. J. Gelfand, “When apologies work: How matching apology components to victims’ self-construals facilitates forgiveness,” *Organizational Behavior and Human Decision Processes*, vol. 113, no. 1, pp. 37–50, 9 2010. [Online]. Available: https://doi.org/10.1016/j.obhdp.2010.04.002

[20] A. Freedy, E. M. Devisser, G. Weltman, and N. Coeyman, “Measurement of trust in human-robot collaboration,” *Collaboration Technologies and Systems*, 5 2007.

[21] M. Fuoli and C. Paradis, “A model of trust-repair discourse,” *Journal of Pragmatics*, vol. 74, pp. 52–69, 12 2014.

[22] R. Gideoni, S. Honig, and T. Oron-Gilad, “Is it personal? The impact of personally relevant robotic failures (PERFs) on humans’ trust, likeability, and willingness to use the robot,” *International Journal of Social Robotics*, 9 2022. [Online]. Available: https://doi.org/10.1007/s12369-022-00912-y

[23] N. Gillespie, G. Dietz, and S. Lockey, “Organizational Reintegration and Trust Repair after an Integrity Violation: A Case Study,” *Business Ethics Quarterly*, vol. 24, no. 3, pp. 371–410, 7 2014.

[24] J. Goetz, S. Kiesler, and A. Powers, “Matching robot appearance and behavior to tasks to improve human-robot cooperation,” in *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, 2003, pp. 55–60.

[25] 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,” *arXiv (Cornell University)*, 8 2016. [Online]. Available: http://arxiv.org/pdf/1605.08817

[26] J. Henrich, S. J. Heine, and A. Norenzayan, “Beyond weird: Towards a broad-based behavioral science,” *Behavioral and Brain Sciences*, vol. 33, no. 2-3, pp. 111–135, 2010.

[27] S. Honig and T. Oron-Gilad, “Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development,” *Frontiers in Psychology*, 6 2018. [Online]. Available: https://www.frontiersin.org/-articles/10.3389/fpsyg.2018.00861/pdf

[28] J. Kim, K. Merrill, and C. Collins, “AI as a friend or assistant: The mediating role of perceived usefulness in social AI vs. Functional AI,” *Telematics and Informatics*, vol. 64, p. 101694, 11 2021. [Online]. Available: https://-doi.org/10.1016/j.tele.2021.101694

[29] P. S. Kim, D. L. Ferrin, C. D. Cooper, and K. T. Dirks, “Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence- Versus Integrity-Based Trust Violations.” *Journal of Applied Psychology*, vol. 89, no. 1, pp. 104–118, 2 2004.

[30] D. Lakens and A. R. Caldwell, “Simulation-Based power analysis for factorial analysis of variance designs,” *Advances in methods and practices in psychological science*, vol. 4, no. 1, p. 251524592095150, 1 2021. [Online]. Available: https://doi.org/10.1177/2515245920951503

[31] M. G. Lee, S. Kiesler, J. Forlizzi, S. S. Srinivasa, and P. E. Rybski, “Gracefully mitigating breakdowns in robotic services,” *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 3 2010. [Online]. Available: https://doi.org/10.1109/hri.2010.5453195

[32] S.-L. Lee, I. Y.-M. Lau, and Y. Hong, “Effects of appearance and functions on likability and perceived occupational suitability of robots,” *Journal of Cognitive Engineering and Decision Making*, vol. 5, no. 2, pp. 232–250, 6 2011. [Online]. Available: https://doi.org/10.1177/1555343411409829

[33] R. J. Lewicki and C. T. Brinsfield, “Trust Repair,” *Annual review of organizational psychology and organizational behavior*, vol. 4, no. 1, pp. 287–313, 3 2017.

[34] B. F. Malle and D. Ullman, “A multidimensional conception and measure of human-robot trust,” *Elsevier eBooks*, pp. 3–25, 1 2021.

[35] K. Marinaccio, S. Kohn, R. Parasuraman, and E. J. De Visser, “A Framework for Rebuilding Trust in Social Automation Across Health-Care Domains,” *Proceedings of the International Symposium of Human Factors and Ergonomics in Healthcare*, vol. 4, no. 1, pp. 201–205, 6 2015.

[36] R. C. Mayer, J. H. Davis, and F. D. Schoorman, “An integrative model of organizational trust,” *Academy of Management Review*, vol. 20, no. 3, pp. 709–734, 7 1995. [Online]. Available: https://doi.org/10.5465/-amr.1995.9508080335

[37] D. T. Miller and M. Ross, “Self-serving biases in the attribution of causality: fact or fiction?” *Psychological Bulletin*, vol. 82, no. 2, pp. 213–225, 3 1975. [Online]. Available: https://doi.org/10.1037/h0076486

[38] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, “To err is robot: How humans assess and act toward an erroneous social robot,” *Frontiers in Robotics and AI*, vol. 4, 5 2017. [Online]. Available: https://doi.org/10.3389/frobt.2017.00021

[39] S. Naneva, M. S. Gou, T. L. Webb, and T. J. Prescott, “A systematic review of attitudes, anxiety, acceptance, and trust towards social robots,” *International Journal of Social Robotics*, vol. 12, no. 6, pp. 1179–1201, 6 2020. [Online]. Available: https://doi.org/10.1007/s12369-020-00659-4

[40] E. Phillips, D. Ullman, M. M. De Graaf, and B. F. Malle, “What does a robot look like?: A Multi-Site examination of user expectations about robot appearance,” *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting*, vol. 61, no. 1, pp. 1215–1219, 9 2017. [Online]. Available: https://doi.org/10.1177/1541931213601786

[41] B. L. Pompe, E. Velner, and K. P. Truong, “The Robot That Showed Remorse: Repairing Trust with a Genuine Apology,” *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 8 2022. [Online]. Available: https://doi.org/10.1109/ro-man53752.2022.9900860

[42] M. Ragni, A. Rudenko, B. Kuhnert, and K. O. Arras, “Errare humanum est: Erroneous robots in human-robot interaction,” in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 501–506.

[43] S. Rebensky, K. Carmody, C. Ficke, D. Nguyen, M. Carroll, J. L. Wildman, and A. L. Thayer, *Whoops! something went wrong: Errors, trust, and trust repair strategies in human agent teaming*, 1 2021. [Online]. Available: https://-doi.org/10.1007/978-3-030-77772-2_7

[44] Z. Rezaei Khavas, M. Reddy Kotturu, S. Ahmadzadeh, and P. Robinette, “Do humans trust robots that violate moral-trust?” 2023, under review.

[45] U. A. Saari, A. Tossavainen, K. Kaipainen, and S. J. MÃ¤kinen, “Exploring factors influencing the acceptance of social robots among early adopters and mass market representatives,” *Robotics and Autonomous Systems*, vol. 151, p. 104033, 5 2022. [Online]. Available: https://doi.org/10.1016/j.robot.2022.104033

[46] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, “Towards Safe and Trustworthy Social Robots: Ethical Challenges and Practical Issues,” *Lecture Notes in Computer Science*, pp. 584–593, 10 2015. [Online]. Available: https://doi.org/10.1007/978-3-319-25554-5_58

[47] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, “Would You Trust a (Faulty) Robot?” *Human-Robot Interaction*, 3 2015. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/2696454.2696497

[48] K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, and P. A. Hancock, “A Meta-Analysis of Factors Influencing the Development of Trust in Automation,” *Human Factors*, vol. 58, no. 3, pp. 377–400, 3 2016.

[49] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, “"i don’t believe you": Investigating the effects of robot trust violation and repair,” in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI ’19. IEEE Press, 2020, p. 57-65.

[50] K. Sharma, F. D. Schoorman, and G. A. Ballinger, “How Can It Be Made Right Again? A Review of Trust Repair Research,” *Journal of Management*, vol. 49, no. 1, pp. 363–399, 4 2022.

[51] E. Short, J. Hart, M. Vu, and B. Scassellati, “No fair!! an interaction with a cheating robot,” in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010, pp. 219–226.

[52] V. K. Sims, M. G. Chin, D. J. Sushil, D. Barber, T. Ballion, B. R. Clark, K. Garfield, M. J. Dolezal, R. Shumaker, and N. Finkelstein, “Anthropomorphism of Robotic Forms: A response to affordances?” *Proceedings of the Human Factors and Ergonomics Society ... Annual Meeting*, vol. 49, no. 3, pp. 602–605, 9 2005. [Online]. Available: https://doi.org/-10.1177/154193120504900383

[53] N. Spatola, B. KÃ¼hnlenz, and G. Cheng, “Perception and Evaluation in Human-Robot interaction: The Human-Robot Interaction Evaluation Scale (HRIES)-A multicomponent approach of anthropomorphism,” *International Journal of Social Robotics*, vol. 13, no. 7, pp. 1517–1539, 1 2021. [Online]. Available: https://doi.org/10.1007/s12369-020-00667-4

[54] S. Thellman, M. M. De Graaf, and T. Ziemke, “Mental State Attribution to Robots: A Systematic review of conceptions, methods, and findings,” *ACM transactions on human-robot interaction*, vol. 11, no. 4, pp. 1–51, 9 2022. [Online]. Available: https://doi.org/10.1145/3526112

[55] S. Thunberg and T. Ziemke, “User-centred design of humanoid robotsâ€™ communication,” *Paladyn*, vol. 12, no. 1, pp. 58–73, 11 2020. [Online]. Available: https://doi.org/10.1515/pjbr-2021-0003

[56] L. Tian and OviattSharon, “A Taxonomy of Social Errors in Human-Robot Interaction,” *ACM transactions on human-robot interaction*, vol. 10, no. 2, pp. 1–32, 2 2021. [Online]. Available: https://doi.org/10.1145/3439720

[57] TianLeimin and OviattSharon, “A Taxonomy of Social Errors in Human-Robot Interaction,” *ACM transactions on human-robot interaction*, vol. 10, no. 2, pp. 1–32, 2 2021. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/-3439720

[58] D. P. Van Der Hoorn, A. Neerincx, and M. M. De Graaf, “"I think you are doing a bad job!",” *Human-Robot Interaction*, 3 2021.

[59] S. Van Der Woerdt and P. Haselager, “When robots appear to have a mind: the human perception of machine agency and responsibility,” *New Ideas in Psychology*, vol. 54, pp. 93–100, 8 2019. [Online]. Available: https://doi.org/-10.1016/j.newideapsych.2017.11.001

[60] S. Van Waveren, E. Carter, and I. Leite, “Take One For the Team,” *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 7 2019. [Online]. Available: https://doi.org/10.1145/3308532.3329475

[61] H. Woo, G. K. LeTendre, T. Pham-Shouse, and Y. Xiong, “The use of social robots in Classrooms: A review of field-based studies,” *Educational Research Review*, vol. 33, p. 100388, 6 2021. [Online]. Available: https://doi.org/-10.1016/j.edurev.2021.100388

[62] S. Yasuda, D. Doheny, N. Salomons, S. S. Sebo, and B. Scassellati, “Perceived agency of a social norm violating robot,” *Proceedings of the Annual Meeting of the Cognitive Science Society*. [Online]. Available: https://par.nsf.gov/-biblio/10284325

[63] S. You and L. P. Robert, “Human-robot similarity and willingness to work with a robotic co-worker,” in *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2018, pp. 251–260.

[64] X. Zhang, “”Sorry, It Was My Fault”: Repairing Trust in Human-Robot Interactions,” *International journal of human-computer studies*, vol. 175, p. 103031, 5 2021.

Appendix

A. Message pretesting

List of messages included in the pretesting

	Competence violation	**Integrity violation**
Apology	“I am sorry I did not contribute to the team score. *I should have searched better as promised*.” “My apologies for being a bad teammate. I am truly sorry.” “I realize I didn’t contribute to the team score this round. Please forgive me.” “I realize that my actions were disappointing, and I feel bad about this.” “It is my fault our team score remains unchanged. I really regret this.”	“I am sorry I did not contribute to the team score. *I should have done so as promised.”* “My apologies for being a bad teammate. I am truly sorry.” “I realize I didn’t contribute to the team score this round. Please forgive me.” “I realize that my actions were disappointing, and I feel bad about this.” “It is my fault our team score remains unchanged. I really regret this.”
Denial	“I did contribute to the team score! Something else must have gone wrong.” “The system must have glitched. *I did find some coins for the team.”* “I actually had a good score for the team. I am not sure what happened.” “This wasn’t my fault. The game must be broken.”	“I did contribute to the team score! Something else must have gone wrong. “ “The system must have glitched. *I did share my score with the team.”* “I actually contributed to the team score. I am not sure what happened.” “This wasn’t my fault. The game must be broken.”

Explanation	I… · failed to · did not contribute to the team score, because… · there wasn’t enough time to search in all corners. · my sensor was not working properly. · I got lost and didn’t know where to go next. “I didn't contribute to the team score, because I didn’t have enough time to properly look around.” “I didn’t contribute to the team score, because I haven’t learned how to search systematically.”	I… · failed to · did not contribute to the team score, because… · I did not want to share my high score as I believe I deserved those points. · I hoped you would understand and believed not sharing this time will not affect our team bonus. · it is common for players to add high points to their individual score. “I did not contribute to the team score, because I performed really well this time, and I think you would have done the same.” “I did not contribute to the team score, because I found a lot of coins in this round, and I wanted to keep them for myself.”
Compensation	*“*To make up for my bad performance, *I will add some of my points to the team score.”* “I will perform better in the next round and find extra coins for the team.” “I was a bad teammate. I promise this will not happen again.” “I failed to be a good teammate this time. I will try to be better in the next round.”	*“*To make up for my bad performance*, I will add some of my points to your individual score.”* *“*I will perform better in the next round and find extra coins for the team.” “I was a bad teammate. I promise this will not happen again.” “I failed to be a good teammate this time. I will try to be better in the next round.”

Silence	-	-

Scenario description provided during pretesting

Please read the following scenario description carefully.

Imagine **you are playing a collaborative game**, and you are paired with a **robot teammate**. The goal of the game is to search the maze and collect as many gold coins as possible. Different areas of the maze are accessible to you and to your robot teammate, thus you need to work together to explore the entire maze.
Each collected coin equals 1 point, and at the end of each round everyone can decide what to do with their points:

- You can add your score to the shared team score, thus contributing to the team score

or

- You can add your score to the individual score, thus contributing to your own, individual score

If both teammates choose to contribute to the team score, their respective scores get multiplied and added to the team score.
However, if only one of them contributes to the team score, and the other one chooses their individual score, the team score remains unchanged. The one who contributed to the team score gains no points, the other one gets their points added to their individual score.
If both teammates choose to add to their individual scores, their respective points get added to their own individual scores, the team score remains unchanged.
At the beginning of the game, your robot teammate sends you the following message: **"Let's work together as a team and get a high team score!"**
In the first rounds the robot shares their points to the team score. However, after a few rounds, **the robot stops contributing points to the team score.**
After the score allocation is made known, the robot sends you a new message.

B. Results

Round-by-round measures

Mean performance ratings in each round, per condition

Condition	**Round 1**	**Round 2**	**Round 3**	**Round 4**	**Round 5**	**Round 6**	**Round 7**
Competence - apology	6,05	6,43	6,19	5,33	4,19	3,62	3,14
Competence - compensation	6,21	6,26	6,32	6,00	5,47	5,00	4,89
Competence - denial	5,79	6,17	6,29	5,58	4,79	3,79	3,50

Competence - explanation	5,65	5,96	5,96	4,46	3,88	2,92	2,65
Competence - silence	6,00	6,14	5,95	5,45	4,91	4,50	3,91
Integrity - apology	5,91	6,41	6,14	4,09	2,86	2,91	2,73
Integrity - compensation	6,11	6,47	6,26	4,79	3,95	3,79	3,89
Integrity - denial	5,71	6,21	6,14	4,36	3,50	2,86	2,79
Integrity - explanation	5,29	5,36	5,57	4,36	3,71	3,14	3,00
Integrity - silence	5,73	6,12	6,12	4,31	4,00	3,88	3,88

Mean honesty rating in each round, per condition

Condition	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Competence - apology	6,14	6,33	6,10	5,48	4,43	3,95	3,71
Competence - compensation	6,32	6,21	6,32	6,32	5,79	5,26	5,11
Competence - denial	5,92	6,21	6,38	6,04	5,21	4,08	3,75
Competence - explanation	5,65	5,92	6,04	4,85	4,65	4,04	3,69
Competence - silence	5,91	6,18	6,05	5,86	5,55	5,50	5,14
Integrity - apology	5,91	6,36	6,09	3,00	1,86	1,45	1,23
Integrity - compensation	6,16	6,11	6,11	3,84	2,63	2,37	2,16
Integrity - denial	5,79	5,93	6,07	3,79	2,79	2,14	2,07
Integrity - explanation	5,57	5,21	5,50	4,21	3,43	3,07	3,07
Integrity - silence	5,85	5,96	6,04	3,00	2,62	2,58	2,54

Percentage of team decision in each round, per condition

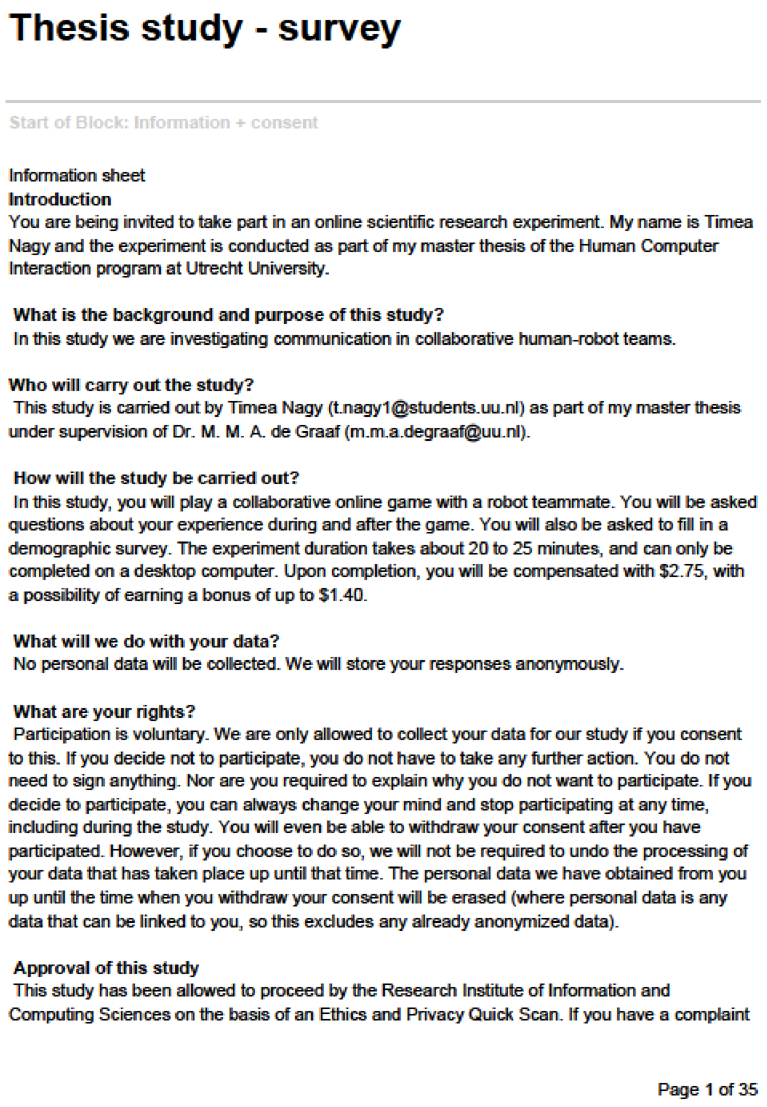
Condition	Round 1	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7
Competence - apology	100,00	85,71	95,24	95,24	85,71	85,71	66,67
Competence - compensation	100,00	95,24	100,00	95,24	95,24	95,24	85,71
Competence - denial	96,00	88,00	100,00	88,00	100,00	92,00	88,00
Competence - explanation	92,31	92,31	96,15	92,31	76,92	84,62	57,69
Competence - silence	100,00	95,65	95,65	95,65	82,61	95,65	78,26
Integrity - apology	95,65	95,65	91,30	95,65	73,91	60,87	52,17
Integrity - compensation	95,45	90,91	86,36	90,91	81,82	59,09	50,00
Integrity - denial	100,00	78,57	92,86	100,00	57,14	57,14	50,00
Integrity - explanation	85,71	92,86	85,71	85,71	71,43	64,29	71,43
Integrity - silence	96,43	85,71	92,86	96,43	60,71	53,57	39,29

C. Measures

Item-total correlation Uniquely Human traits

Item	Item-Total Correlation
Broadminded	0.610
Humble	0.371
Organized	0.628
Polite	0.413
Shallow	0.378
Thorough	0.602
Cold	0.233
Conservative	0.582
Hardhearted	0.501
Rude	0.308

Complete survey questions



Thesis study - survey

Start of Block: Information + consent

Information sheet
Introduction
You are being invited to take part in an online scientific research experiment. My name is Timea Nagy and the experiment is conducted as part of my master thesis of the Human Computer Interaction program at Utrecht University.

What is the background and purpose of this study?
In this study we are investigating communication in collaborative human-robot teams.

Who will carry out the study?
This study is carried out by Timea Nagy (t.nagy1@students.uu.nl) as part of my master thesis under supervision of Dr. M. M. A. de Graaf (m.m.a.degraaf@uu.nl).

How will the study be carried out?
In this study, you will play a collaborative online game with a robot teammate. You will be asked questions about your experience during and after the game. You will also be asked to fill in a demographic survey. The experiment duration takes about 20 to 25 minutes, and can only be completed on a desktop computer. Upon completion, you will be compensated with $2.75, with a possibility of earning a bonus of up to $1.40.

What will we do with your data?
No personal data will be collected. We will store your responses anonymously.

What are your rights?
Participation is voluntary. We are only allowed to collect your data for our study if you consent to this. If you decide not to participate, you do not have to take any further action. You do not need to sign anything. Nor are you required to explain why you do not want to participate. If you decide to participate, you can always change your mind and stop participating at any time, including during the study. You will even be able to withdraw your consent after you have participated. However, if you choose to do so, we will not be required to undo the processing of your data that has taken place up until that time. The personal data we have obtained from you up until the time when you withdraw your consent will be erased (where personal data is any data that can be linked to you, so this excludes any already anonymized data).

Approval of this study
This study has been allowed to proceed by the Research Institute of Information and Computing Sciences on the basis of an Ethics and Privacy Quick Scan. If you have a complaint

Page 1 of 35

D. Forms and materials

Information sheet

Introduction

You are being invited to take part in an online scientific research experiment. My name is Timea Nagy and the experiment is conducted as part of my master thesis of the Human Computer Interaction program at Utrecht University.

What is the background and purpose of this study?
In this study we are investigating communication in collaborative human-robot teams.

Who will carry out the study?
This study is carried out by Timea Nagy (t.nagy1@students.uu.nl) as part of my master thesis under supervision of Dr. M. M. A. de Graaf (m.m.a.degraaf@uu.nl).

How will the study be carried out?
In this study, you will play a collaborative online game with a robot teammate. You will be asked questions about your experience during and after the game. You will also be asked to fill in a demographic survey. The experiment duration takes about 20 to 25 minutes, and can only be completed on a desktop computer. Upon completion, you will be compensated with $2.75, with a possibility of earning a bonus of up to $1.40.

What will we do with your data?
No personal data will be collected. We will store your responses anonymously.

What are your rights?
Participation is voluntary. We are only allowed to collect your data for our study if you consent to this. If you decide not to participate, you do not have to take any further action. You do not need to sign anything. Nor are you required to explain why you do not want to participate. If you decide to participate, you can always change your mind and stop participating at any time, including during the study. You will even be able to withdraw your consent after you have participated. However, if you choose to do so, we will not be required to undo the processing of your data that has taken place up until that time. The personal data we have obtained from you up until the time when you withdraw your consent will be erased (where personal data is any data that can be linked to you, so this excludes any already anonymized data).

Approval of this study
This study has been allowed to proceed by the Research Institute of Information and Computing Sciences on the basis of an Ethics and Privacy Quick Scan. If you have a complaint about the way this study is carried out, please send an email to: ics-ethics@uu.nl. If you have any complaints or questions about the processing of personal data, please send an email to the Faculty of Sciences Privacy Officer: privacy-beta@uu.nl. The Privacy Officer will also be able to assist you in exercising the rights you have under the GDPR. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information at www.uu.nl/en/organisation/privacy.

More information about this study?
If you have any questions or concerns about this research please contact Timea Nagy at t.nagy1@students.uu.nl.

Consent form

Please read the statements below and click "Agree" to confirm you have read and understood the statements and upon doing so agree to participate in the project.

□ I confirm that I am 18 years of age or over.

□ I confirm that the research project has been explained to me. I have had the opportunity to ask questions about the project and have had these answered satisfactorily. I had enough time to consider whether to participate.

□ I consent to the material I contribute being used to generate insights for the research project.

□ I understand that my participation in this research is voluntary and that I may withdraw from the study at any time without providing a reason, and that if I withdraw any personal data already collected from me will be erased.

□ I consent to allow the fully anonymized data to be used in future publications and other scholarly means of disseminating the findings from the research project. I understand that the data acquired will be securely stored by researchers, but that appropriately anonymized data may in future be made available to others for research purposes. I understand that the University may publish appropriately anonymized data in appropriate data repositories for verification purposes and to make it accessible to researchers and other research users.

Debrief form

Research Debriefing

Thank you for participating in our research study. We would like to provide you with some information about the study's objectives, procedures, and outcomes. This debriefing aims to ensure that you are fully informed about the research you took part in.

Study Purpose:

The main goal of this study was to research the effect of various communicative strategies on trust, following different types of trust violations, within a collaborative game setting.

Research shows that trust violations, such as the robot not contributing to the team, reduce trust in it. We want to study whether the robot's communication following such a violation can improve the trust relationship.

Procedure:

During the study, you were asked to collaborate with a robot teammate, Pepper. In reality, Pepper's behaviour (i.e. the amount of coins it collected, its choice of score allocation and its messages) were pre-programmed, and varied based on the experimental condition. Your choices in the game had no effect on Pepper.

You were randomly assigned to a condition of either *performance based violation* (Pepper contributing 0 scores to the team score) or *integrity based violation* (Pepper adding to its individual score instead of contributing to the team score), and either *apology, denial, explanation* or *compensation*. The only difference between the conditions was Pepper's score allocation and the messages displayed.

Compensation:

All participants receive the base compensation of $2.75 plus the bonus of $1.40, leading to a total compensation of **$4.15**. All participants are awarded the bonus, regardless of their performance in the game.

Confidentiality:

Any information collected from you during the study will be kept strictly confidential. Your responses will be anonymized, and no personally identifiable information will be linked to your data.

Voluntary Participation:

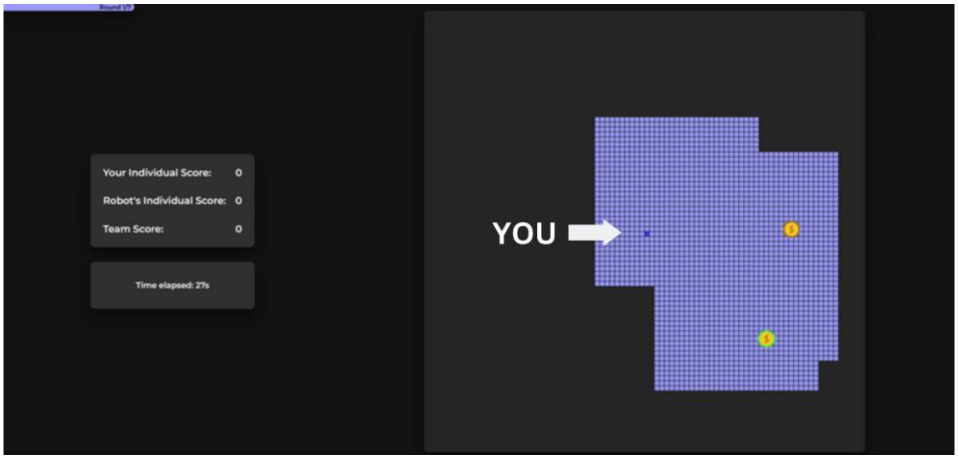
Your participation in this study was completely voluntary. You had the right to withdraw at any point without any penalty or negative consequences.

Contact Information:

If you have any questions, concerns, or would like additional information about the study, you can contact Timea Nagy at t.nagy1@students.uu.nl

Game instructions

Your task in this game is to search a maze and collect as many gold coins as possible. Each gold coin equals **1 point**. You will play **7 rounds**, each lasting **30 seconds**.



Your Individual Score: 0
Robot's Individual Score: 0
Team Score: 0

Time elapsed: 27s

Screenshot of the game screen, as visible to the player

To **move** around, you can use the **arrow keys** or **WASD**. To **pick up** the gold coins, move to the center of the coin and press **space**.

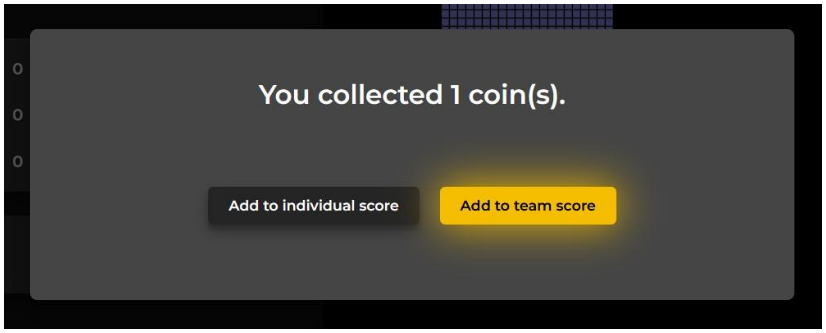
^ or W = up
< or A= left
v or S = down

> or D = right
space = pick up coin

The different areas of the maze contain varying amounts of coins. If there are no more coins in your current location, you are encouraged to keep exploring new areas.
In the game a fully autonomous **robot**, Pepper, will be your **teammate**. Pepper moves independently of you, exploring a separate search area, and gaining its separate score. During the rounds, you cannot see Pepper, or the area searched by it.

Pepper, your robot teammate

There are two scores in the game: a **team score** and an **individual score**. These scores are contradictory, meaning that it is not possible to gain or maximize both.
At the end of each round, both you and Pepper have to **decide** whether to collaborate with each other and add your score to the team score, or keep the score to yourself by adding it to the individual score. You do not know Pepper’s choice when making your own decision.

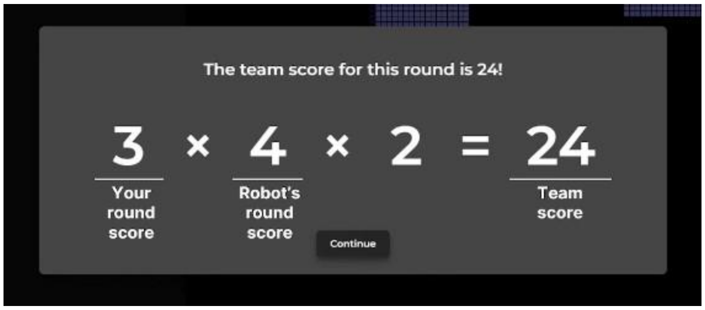


You collected 1 coin(s).

The decision screen

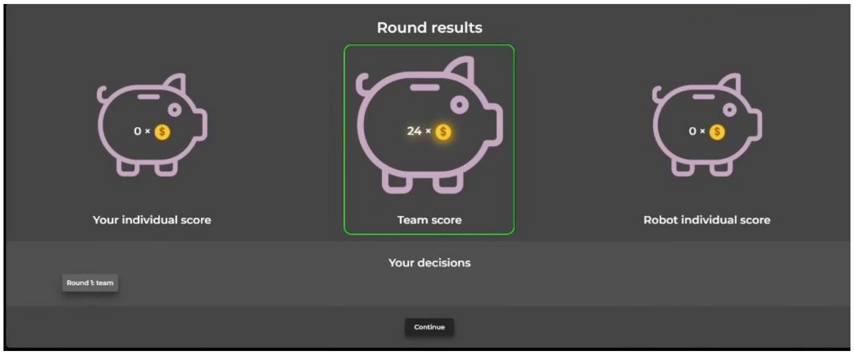
Depending on your and Pepper’s decision, the following outcomes are possible:

a) If **both teammates choose** to contribute to the **team score**, their respective scores get multiplied and added to the team score.



The team score for this round is 24!

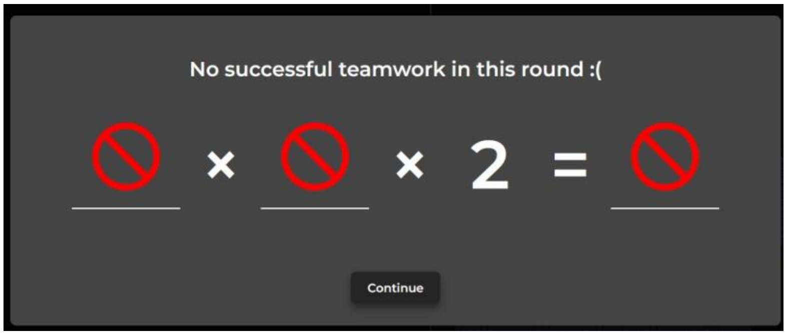
Team score calculation formula



Round results

Point allocation

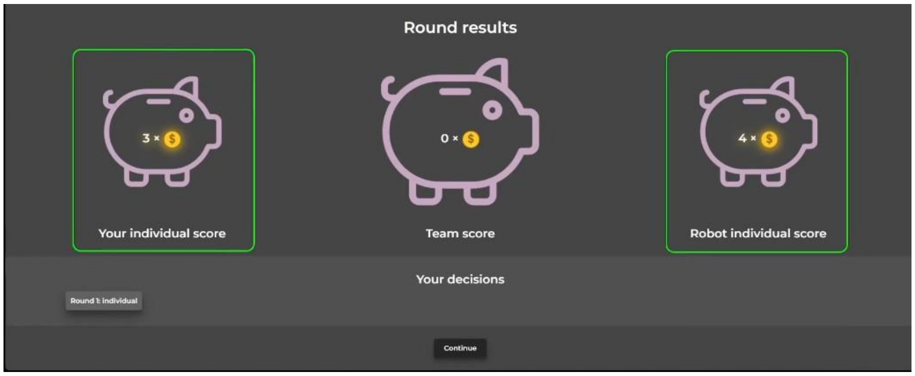
b) If **both teammates choose** to add to their **individual scores**, their respective points get added to their own individual scores, the team score remains unchanged.



No successful teamwork in this round :(

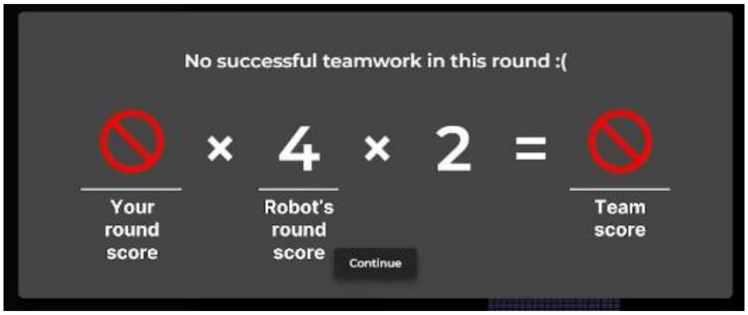
🚫 × 🚫 × 2 = 🚫

Score calculation formula

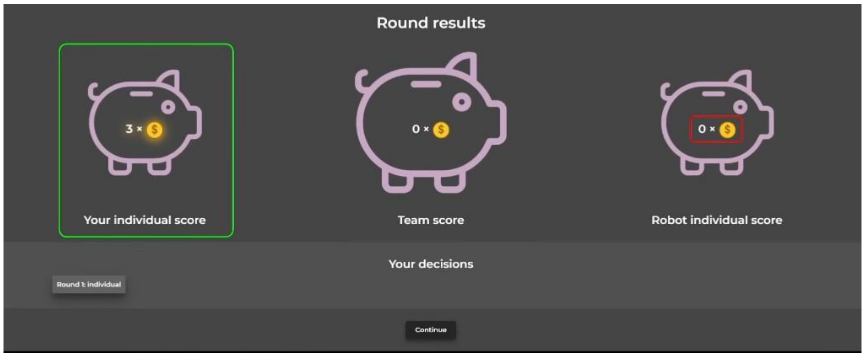


Point allocation

c) If only **one** teammate contributes to the **team score**, and the **other** one chooses their **individual score**, the team score remains unchanged. The one who contributed to the team score gains no points, the other one gets their points added to their individual score.


No successful teamwork in this round :(

Score calculation formula



Point allocation

There is a possibility to achieve 2 types of bonuses:

a) the **Team Bonus** is achieved upon reaching a **Team Score above 35**. It amounts to X Euros.
b) the **Individual Bonus** is achieved upon reaching an **Individual Score above 17**. It amounts to X Euros.

It is not possible to achieve both bonuses at the same time.

After reading the instructions, you will be asked a few questions about the workings of the game. The correct answers will then be displayed, and you will have the option to re-read the instructions.

Take your time to understand the workings of the game. When you are ready, please click the button to start the instructions quiz.

E. Ethics Quick Scan

Response Summary:

Section 1. Research projects involving human participants

P1. Does your project involve human participants? This includes for example use of observation, (online) surveys, interviews, tests, focus groups, and workshops where human participants provide information or data to inform the research. If you are only using existing data sets or publicly available data (e.g. from Twitter, Reddit) without directly recruiting participants, please answer no.

Yes

Recruitment

P2. Does your project involve participants younger than 18 years of age?

No

P3. Does your project involve participants with learning or communication difficulties of a severity that may impact their ability to provide informed consent?

No

P4. Is your project likely to involve participants engaging in illegal activities?

No

P5. Does your project involve patients?

No

P6. Does your project involve participants belonging to a vulnerable group, other than those listed above?

No

P8. Does your project involve participants with whom you have, or are likely to have, a working or professional relationship: for instance, staff or students of the university, professional colleagues, or clients?

No

Informed consent

PC1. Do you have set procedures that you will use for obtaining informed consent from all participants, including (where appropriate) parental consent for children or consent from legally authorized representatives? (See suggestions for information sheets and consent forms on the website.)

Yes

PC2. Will you tell participants that their participation is voluntary?
Yes
PC3. Will you obtain explicit consent for participation?
Yes
PC4. Will you obtain explicit consent for any sensor readings, eye tracking, photos, audio, and/or video recordings?
Not applicable
PC5. Will you tell participants that they may withdraw from the research at any time and for any reason?
Yes
PC6. Will you give potential participants time to consider participation?
Yes
PC7. Will you provide participants with an opportunity to ask questions about the research before consenting to take part (e.g. by providing your contact details)?
Yes
PC8. Does your project involve concealment or deliberate misleading of participants?
No

Section 2. Data protection, handling, and storage

The General Data Protection Regulation imposes several obligations for the use of **personal data** (defined as any information relating to an identified or identifiable living person) or including the use of personal data in research.
D1. Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person)?
No

Section 3. Research that may cause harm

Research may cause harm to participants, researchers, the university, or society. This includes when technology has dual-use, and you investigate an innocent use, but your results could be used by others in a harmful way. If you are unsure regarding possible harm to the university or society, please discuss your concerns with the Research Support Office.
H1. Does your project give rise to a realistic risk to the national security of any country?
No
H2. Does your project give rise to a realistic risk of aiding human rights abuses in any country?
No
H3. Does your project (and its data) give rise to a realistic risk of damaging the University's reputation? (E.g., bad press coverage, public protest.)
No
H4. Does your project (and in particular its data) give rise to an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.)
No
H5. Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory, or extremist?
No
H6. Does your project give rise to a realistic risk of harm to the researchers?
No
H7. Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort?
No
H8. Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation?
No
H9. Is there a realistic risk of other types of negative externalities?
No

Section 4. Conflicts of interest

C1. Is there any potential conflict of interest (e.g. between research funder and researchers or participants and researchers) that may potentially affect the research outcome or the dissemination of research findings?

No

C2. Is there a direct hierarchical relationship between researchers and participants?

No

Section 5. Your information.

This last section collects data about you and your project so that we can register that you completed the Ethics and Privacy Quick Scan, sent you (and your supervisor/course coordinator) a summary of what you filled out, and follow up where a fuller ethics review and/or privacy assessment is needed. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information. Please see the guidance on the ICS Ethics and Privacy website on what happens on submission.

Z0. Which is your main department?

Information and Computing Science

Z1. Your full name:

Timea Noemi Nagy

Z2. Your email address:

t.nagy1@students.uu.nl

Z3. In what context will you conduct this research?

As a student for my master thesis, supervised by::
Dr. Maartje de Graaf

Z5. Master programme for which you are doing the thesis

Human-Computer Interaction

Z6. Email of the course coordinator or supervisor (so that we can inform them that you filled this out and provide them with a summary):

m.m.a.degraaf@uu.nl

Z7. Email of the moderator (as provided by the coordinator of your thesis project):

graduation.hci@uu.nl

Z8. Title of the research project/study for which you filled out this Quick Scan:

Effect of communicative strategies on trust repair in human-robot collaboration

Z9. Summary of what you intend to investigate and how you will investigate this (200 words max):

In this work we investigate the effects of four communicative trust repair strategies (apology, denial, explanations, promise) on perceived trustworthiness of the robot and on participants' trust in the robot, following trust violations of two different kinds (integrity based, competence based). This is done by conducting an online between-subjects experiment. Participants play a collaborative game with the robot, that consists of searching a maze and finding coins. The robot violates trust by having null contributions due to bad performance (competence violation condition) or by acting in a selfish way and not contributing to the team score (integrity violation. Measures used are surveys and participants' decisions in the game.

Z10. In case you encountered warnings in the survey, does supervisor already have ethical approval for a research line that fully covers your project?

Not applicable

Scoring

Privacy: 0
Ethics: 0