Can we forgive a robotic teammate?

The role of Trust and Trust Repair Strategies in Human-Agent Teams

Bachelor Thesis

22.02.2023

Ronja Kreiter

S2347458

Faculty of Behavioural Management and Social Sciences

Department Psychology

University of Twente.

First Supervisor: Drs. Esther Kox

Second Supervisor: Prof. Dr. José Kerstholt

Abstract

Background and Objective. Given the recent developments in Industry 5.0, Human Robot collaboration in high stakes situations become increasingly relevant. Intelligent agents could be our future teammates. In order to effectively work together as a team, trust needs to be appropriately calibrated even after trust violations. The main objective of this study was to investigate the effectiveness of preventative and restorative action to foster trust repair after a trust violation and determine what individual characteristics predict trust in robotic teammates.

Method. The experimental research was based on a 3 (Time) x 2 (Uncertainty communication) x 2 (apology) design (n=39). The dependent variable trust was measured six times per participant, namely three times per experimental run. Uncertainty was manipulated within subjects and apology was manipulated between subjects.

Findings. The study did not find an effect for individual characteristics predicting prior trust in the intelligent agent. Neither apology nor uncertainty communication had a significant effect on trust over time. Participants preferred drones that used uncertainty communication. Uncertainty communication together with one's tendency to forgive had a significant effect on trust after trust repair.

Conclusion. This study found that individual characteristics did not influence trust and social cognitive trust repair strategies did not impact trust repair, which is in line with the unique agent hypothesis. The current findings indicate that uncertainty communication can be valuable to trust development and forgiveness in HAT. Future research is needed to further explore Human Robot trust dynamics.

Keywords: Intelligent agents, Human-Agent Team, Trust repair, Forgiveness, Teamwork

Can we forgive a robotic teammate?

Imagine you are in a house with heavy gear on your back, following your robotic teammate, a Drone called SA1. SA1 says: “Warning. Danger detected in this environment with 80 percent certainty. I advise you to proceed carefully”. You move around the corner, and you see smoke. An old bomb goes off. Smoke everywhere. Your robotic teammate apologizes and says “Incorrect advice due to faulty object detection by C1 – DSO camera. I am sorry this put you in danger.” The Drone warned you beforehand that it was not a 100 percent certain and then apologizes afterwards. Can such presentative and restorative strategies affect a person’s trust over time? Are you willing to forgive your robotic teammate for their mistake?

The current research investigates trust, trust repair and forgiveness in Human Agent Teams in high stakes situations. *Human Agent Teams* (HAT) consists of at least one human and one intelligent agent (Kox et al., 2021). *Intelligent agents* (IA) are artificial entities that observe and act upon an environment and that are able to communicate and collaborate with other agents to solve problems and achieve common goals (Kox et al., 2021). In this context, a *robot* can be defined as an intelligent system with a physical embodiment (De Visser et al., 2020). For an IA to be able to work in HAT they need a variety of skills. For a better understanding of what is needed of an IA in this setting, the following sections explore teamwork and the consequences of working together, namely the breach of trust and trust repair and hence forgiveness.

Human agent teams in industry 5.0

In the context of Industry 5.0, *Human Robot Interaction* (HRI) research focuses on how to improve the communication, and thus lead to an improvement of task performance of the robot and consequently lead to a task reduction for humans (Chandrasekaran & Conrad,

2015). IA have the potential to decrease repetitive work for humans, increase production outcomes (Russel & Norvig, 2022). However, issues of trust, risk management and organizational change are intertwined with how humans feel about and interact with the new technologies (Bondarouk & Fisher, 2020).

Given the recent trends in research and development, it could be assumed that HATs are going to become a more common occurrence in different work fields. Most HATs are based on IA-adviced decision making, where the human has to decide to trust or override the agent's advice (Bansal et al., 2019). Moreover, with the advancement in AI, they are more frequently used in high stakes and safety critical applications (Russel & Norvig, 2022). For example, in the military domain, where IA are used for rescue missions to help carry heavy objects and navigate through uneven environments or for disposing of bombs and suspicious objects (Chandrasekaran & Conrad, 2015). Even if an IA advice is imperfect, effective HATs can perform better than either human or IA alone in high stakes situations (Wang et al. 2016; Jaderberg et al. 2019 as cited in Bansal et al., 2019).

Teamwork

Forsyth (2019) defined teams as a particular type of group that is working together for the pursuit of a common goal through interdependent interaction. *Teamwork* requires team members to collaborate and coordinate by combining their knowledge, skills, and abilities in order to achieve a common goal, whether this consists of confronting an obstacle or solving a problem or task that needs to be solved or completed (Forsyth, 2019). Ideally, in a team, one's skills and knowledge complement each other. Looking at the range of skills and abilities of different IA, it appears that in fact, IA differ in their level of automation and artificial intelligence (AI). On the one hand, agents with high level artificial intelligence can think, understand and act humanly which allows the AI to learn autonomously and collaborate with humans (Xing et al., 2022). On the other hand, agents with low-level

artificial intelligence process a specific intelligence for the core aspect of its task (Xing et al., 2022).

Furthermore, teamwork requires interrelated thoughts, actions and feelings of each team member which are needed to function as a team through coordination and cooperative interaction (Forsyth, 2019). People match nonverbal cues to foster emotional attunement. Their tone, rhythm and quality of speech also conveys information about their emotional and physiological state (Lee & See, 2004). Hence, teamwork requires the ability to understand implicit communication through emotions and nonverbal exchanges, which are typically human skills (Kox et al., 2021). Previous research has shown that vulnerable communication facilitates trust development and consequently also contribute to successful teamwork (De Visser et al., 2020). Research into social robots has shown that the social acceptance of an IA depends on its abilities to express emotions and using human-like interaction mechanisms (Ruiz-del-Solar et al., 2010). However, IA have fewer social skills than humans which could be an obstacle in building trust in a HAT (Kox et al., 2021).

Trust

While a variety of definitions of trust have been suggested, this study defines trust in the HRI context as the willingness to (1) *accept IA produced information* (Hancock et al, 2011), (2) *act based on this information* (Kox, Siegling & Kersthold, 2022) and (3) *accept the vulnerability of depending on the IA* (Mayer et al., 1995 as cited in Colquitt et al., 2007). According to a definition provided by Lee and See (2004), trust affects reliance as an attitude rather than a belief, intention, or behaviour. Hence, trust can be defined as a dynamic attitude (Lee & See, 2004) that needs to be constantly calibrated based on new information about the situation.

Trust calibration

Trust calibration can be defined as the correspondence between one's trust in automation and the automation capabilities (Lee & See, 2004). Poor trust calibration, meaning either over or under trust, can have negative consequences, especially in high-risk situations. For example, when a human trusts the IA too much (i.e., overreliance), therefore uses the IA, even when it is not accurate to do so. Then the trust exceeds the agent's capabilities (Lee & See, 2004). This is related to the automation bias, which describes the high expectations one has towards automation that attributes the IA as infallible, which in return result in a steeper decline in trust violation compared to humans who are viewed as inherently fallible (Kox et al., 2021). In the field, this can lead to a lack of guidance and control of the IA that is not fully capable of the task and consequently this can result in costly disasters such as accidents with lethal consequences and destruction of equipment (De Visser et al., 2020).

When a human trusts the IA too little on the other hand (i.e., under trust), the human cannot take full advantage of the agent's capabilities and disuses or neglects it (Hancock et al., 2011 and De Visser et al., 2020). When people violate critical assumptions and rely on the automation inappropriately this refers to misuse. In contrast, disuse of automation can be defined as failures that occur due to people rejecting the capabilities of automation (Lee & See, 2004). In case of disuse of automation, environmental constraints, such as time pressure, can lead individuals to not use automation even though they trust it (Lee & See, 2004). Next, neglect tolerance describes the decline in semi-autonomous robot performance as human attention is directed to other tasks and/or as the task complexity increases (Hancock et al, 2011). This can create an unbalanced workload in the team and inefficient monitoring of the IA or the micromanagement of the IA. Furthermore, under trust in the robotic teammate can

lead to a lack of communication, suboptimal solutions to the problem at hand (De Visser et al., 2020).

Neither of these previous outcomes is desirable in high stakes situations since they majorly impact the efficiency and effectiveness of HAT. And more importantly they can have lethal consequences in the field, given the military context. Hence, in order to achieve optimal HAT performance trust needs to be calibrated accordingly.

Trust violation

As established earlier, agents do not always live up to the expectations of their human teammates. To begin with, a high-stake situation involves a high level of uncertainty. Moreover, autonomous machines are run by algorithms that might perform well in structured and predictable situations but struggle with unexpected events (Müller-Dott, 2019). Next, uncertainty affects reliability of predictions and neither agents nor human perform flawlessly under those conditions (Kox et al., 2022). It can be assumed that the robotic teammate will eventually make mistakes. IA suboptimal behaviour or mistakes are sometimes inevitable (Kox et al., 2022). As a consequence, it may come to a breach in trust in HAT in these situations.

The level of trust one shows after a trust violation differs between humans and IA. Linking to the automation bias, the initial higher expectations in IA lead to a more drastic decline after trust violation and greater resistance to rebuild trust as compared to human teammates (Visser et al., 2016). To achieve optimal trust calibration, humans have developed various social strategies to repair trust, which however tend to be lacking in intelligent agents (Kox et al., 2022).

Trust repair strategies

After a trust violation, agents can engage in different trust repair strategies that are most suitable for the specific type of violation. It has been proposed that human-human trust and human-automation trust are governed by the same underlying principles, since social robots have been designed to mimic human communication and therefore will elicit the same response as humans (Visser et al., 2016). The *CASA-paradigm* describes the phenomena that people treat computers and IA as social actors. This means that people apply the same social rules, norms, and expectations to their interactions with IA as soon as social cues are present (Kox et al., 2021). For this reason, it could be assumed that human trust repair strategies could be suitable tools for repairing trust in HAT.

Recent work by Kox et al (2022) has established that social cognitive recovery strategies can minimize the impact of trust violation. In previous research uncertainty communication can firstly has shown to lead to higher levels of trust and secondly, has been observed to be effective in trust calibration before a trust violation (Helldin et al 2013 and Kox et al., 2022). It has previously been observed that human trust repair strategies such as an apology can have a significant effect on trust repair (Kox et al, 2022). Recent evidence suggests that for a competence-based violation, an apology is the most effective repair strategy (Kox et al., 2021).

Apology

Apologising can show understanding of the social requirement of an apology and an acknowledgement of the awareness of what one has done to hurt another person (Kox et al., 2021). Apologies are the most effective repair strategy and the more extensive the apology is, the higher the trust repair (Kox et al., 2021). An apology can consist of the following elements: (1) expression of regret, (2) explanation of why the failure occurred, (3) acknowledgement of responsibility for the mistake, (4) offer for repair, (5) promise that it

will not happen again in the future and (6) a request for forgiveness (Kox et al., 2021). An apology is made after the damage has been done, thus apologies are a restorative action to repair trust.

Uncertainty communication

IA will need to deal with uncertainty *in order to be adaptable* to partial observability, nondeterminism, or adversities (Russel & Norvig, 2022). Uncertainty information has been proposed to be an effective strategy for targeting the expectations towards the IA and hence uncertainty communication is helpful for appropriate trust calibration. For this reason, recent evidence suggests that for an ability-based trust violation, uncertainty communication could be a preventative factor in the decline of trust after a trust violation (Kox et al., 2021). In the case of incorrect advice from an agent, uncertainty communication can help to form and maintain trust (Kox et al, 2022). The human is reminded of the fallibility of the system, which allows the human to calibrate trust and manage expectations accordingly (Kox et al, 2022).

Individual characteristics

Causal factor leading to trust, trust repair and forgiveness in HRI in high stakes situation are poorly understood. Each team member brings their unique experiences, skills, abilities and motivation and personal qualities to HAT that influence how they interact as with their human and robotic team members. The study examines the relationship between different individual characteristics that predict human trust and based on the social actor hypothesis, transfer these characteristics as predictors in prior trust and trust development overtime in the IA. The individual characteristics of interest in this study are forgiveness, affect, self-efficacy, trust propensity and perceived threat and safety. In the following section,

these characteristics will be briefly discussed and what is known about their relation to trust in HRI.

Forgiveness

Given that the trust repair was effective, this would imply that humans can forgive their robotic teammate. The absence of forgiveness forces us in a morally ambivalent situation. Nagenborg (2020) stated that after a trust violation such as a broken promise, humans either punish the individual or forgive them. While a punishment undermines opportunity for future interaction, forgiveness allows a new beginning for future interaction (Nagenborg, 2020). Furthermore, punishment is a morally ambivalent act that aims for the other to suffer by intentionally inflicting harm, which remains problematic regardless of whether the punishment is justified (Nagenborg, 2020).

Forgiveness is central to the coexistence of human beings and therefore is also of similar importance in HRI and relationships (Nagenborg, 2020). It has been proposed that forgiveness in interpersonal relationships can be described as the restoration of harmony in the relationship between the victim and the one that breached the trust (Xie & Peng, 2009). Forgiveness is viewed by some scholars as a process that occurs within an individual and is influenced by various developmental and personality factors (Shults & Sandage, 2003). Similar to how people differ in their tendency to trust, people differ in their tendency to forgive. Sociocultural context shapes how individuals approach conflict (Shults & Sandage, 2003), hence it shapes the way people approach trust repair and forgiveness.

We are willing to forgive a human for a hard and unavoidable choice, even if we do not approve of the choice made (Nagenborg, 2020). Therefore, the questions arise whether the tendency to forgive a human teammate is of similar importance in HAT and whether forgiveness influences can trust repair. Hence it is hypothesized that people who are more

forgiving in nature will be more likely to regain trust in a robot after a violation than people who are less forgive.

The role of affect and self-efficacy

Analytical approaches to trust tend to overestimate the cognitive and underestimate the affective influence on trust (Lee & See, 2004). People tend to feel about trust and not think rationally about trust (Kox et al, 2021). Since emotions tend to fluctuate over time based on the performance of the trustee, emotions indicate where expectations are not confirmed by experience (Lee & See, 2004). The cognitive complexity of the situation can go beyond one's ability to form a complete mental model of the situation. Accordingly, one cannot perfectly predict behaviour and for this reason emotions serve to redistribute cognitive resources, mange priorities and guide behaviour when cognitive resources are not available for calculated rational choices (Lee & See, 2004). Trust can help people to adjust to complexity, reduces uncertainty and guiding appropriate reliance and generating a collaborative advantage (Lee & See, 2004). Additionally, emotional state has been linked to cooperative action (Lee et al., 2011). Emotions can be understood as interpersonal communication systems that can provide guidance to navigate one's problems stemming from dyadic and group relations (De Visser et al, 2020). In short, emotions are fundamental to the way people determine whether to trust each other (De Visser et al, 2020).

Adopting new technologies requires self-efficacy (Zafari et al., 2019). Generally, self-efficacy describes one's ability to mobilize the motivation, cognitive resources and actions needed in order to meet the situation's demands (Wood & Bandura, 1989 as cited in Chen et al., 2001). Self-efficacy in the HAT context describes one's assessment of one's own ability to use and interact with a robot (Zafari et al 2019). The self-confidence of an individual can influence their willingness to trust (Lee & See, 2004). Research has shown that robots that engage in a person-oriented interaction style of the robot increase self-efficacy and make

interaction less frustrating to their human teammates compared to a task-oriented communication style (Zafari et al., 2019).

Trust propensity

People differ in their tendency to trust automation in general (Lee & See, 2004). One's specific history of interactions lead people to a particular level of trust (Lee & See, 2004). Trust propensity are stable individual differences that affect the likelihood of an individual to trust others (Colquitt et al., 2007). High trustors are individuals who have a high trust propensity and act more trustworthy. High trustors tend to be more cooperative, show more prosocial and moral behaviour across situations. These individuals are more likely to build social exchange relationships because they are prone to adhering to the norms of reciprocity (Colquitt et al., 2007). Further, individuals with high trust propensity are more likely to adjust their trust to the situation based on the automation's capabilities (Lee & See, 2004).

Perceived threat and safety

Previous research has established the importance of trust in the context of (1) uncertainty, complexity, and ambiguity (2) vulnerability of control, (3) high stakes and (4) long-term-interdependence (Li, 2012). Well-structured situations without uncertainty are less influenced by trust than unstructured and uncertain situations (Lee & See, 2004 and Kox et al., 2021). In the context of HAT in high stakes situations, it is important to keep in mind that when full understanding of the situations due its complexity impractical, even impossible, and the situation demands adaptive behaviour that cannot be guided by a protocol, trust can guide reliance on automation (Lee & See, 2004). People who find themselves in complex uncertain high-risk situations can experience attentional overload, which leads them to engage in automatic processing. This means that their actions are going to be guided by

unconscious thoughts, heuristics, biases, and emotions (Kox et al., 2021). The perceived threat of the situation can pose an obstacle to appropriate trust calibration.

Current research

The knowledge gap about the trust dynamics in HAT are one of the main obstacles to overcome in order to ensure appropriate trust calibration in HAT working together in high stakes situations. One of the greatest challenges is to understand how trust in IA develops over time and how to repair trust most effectively after a trust violation. A key issue is the question whether human trust dynamics are transferable to robotic teammates and hence the tendency to forgive.

Effectiveness of trust repair strategy

This thesis will examine the way trust develops over time towards the IA by examining self-reported trust over repeated measures. This study seeks to investigate the different effects of social cognitive recovery stages on trust formation, trust violation and trust repair. Depending on experimental condition, the agent engages in different preventive and restorative trust repair strategies.

Therefore, the following primary hypothesis were formulated:

I. An apology increases trust repair
II. Uncertainty communication leads to a lower trust decline after the trust violation than when this information is not provided.
III. Uncertainty communication together with an apology increases trust repair compared to the use to only one strategy.

For additional explanatory purposes the preferences of the drones were investigated. To understand whether people prefer a drone that uses uncertainty communication compared to a drone that does not.

Individual characteristics

Forgiveness

Current research did not reach a consensus on a definition of forgiveness in the HRI context. Forgiveness is central to the coexistence of human beings since it is associated with interpersonal relationships and the morality of behaviour and emotion. Forgiveness in this sense has not been associated with HRI. Owing to this, it is hypothesized that *people who are more forgiving in nature will be more likely to regain trust in a robot after a violation than people who are less forgiving.*

Trust propensity

It has previously been observed that people differ in their tendency to trust (Lee & See, 2004 and Colquitt et al., 2007). For this reason, it has been hypothesized that *people with a higher tendency to trust are going to show higher prior trust in their robotic teammate*.

Role of affect

Research highlights the importance of emotional state in cooperative action (Lee et al., 2011), hence it was hypothesized *that emotional state predicts propensity to trust, trust and forgiveness.* More specifically, it was expected that (1) one's emotional state predicts trust in the robotic teammate, (2) more anxious participants are going to going to show lower trust in the robotic teammate, and (3) participants level of self-efficacy affects their emotional state and hence will increases one's trust in the drone.

Self-efficacy

Data from different studies suggest that working with new technologies requires self-efficacy (Zafari et al., 2019) and self-efficacy influences one's willingness to trust (Lee & See, 2004). It is hypothesized that self-efficacy predicts prior trust in the robotic teammate.

Perceived threat and safety

Based on the importance of trust in uncertain, complex high stakes situations, the question arises: Will the same trust mechanics apply to HRI in high stakes situations, or do they differ due to the perceived threat and safety? Finally, it was hypothesized that the *perceived threat and safety predicts trust in the robotic teammate*. Specifically, it was expected that the perceived threat and feeling of unsafety decrease trust in their robotic teammate.

Methods

Design

The present study is a 3 (time) x 2 (uncertainty communication) x 2 (apology) mixed design. Trust is the dependent variable and is measured within participants three times per house search, namely prior to violation [T1], after violation [T2] and after repair [T3]. The independent variable uncertainty communication is measured within participants. The factor of uncertainty communication was manipulated by varying in the level of uncertainty communicated (70, 75, 80%) or with clear advice ("I advise you to move forward."). Further uncertainty order differed per participant, they either received uncertainty communication in the first or second house search. The other independent variable of the apology is measured between participants. The apology was either present or absent. Participants were randomly assigned to one of the two apology conditions.

Participants

An ethical approval by the ethical committee of the BMS of the University of Twente was obtained before recruiting the participants. The study made use of convenience sampling. The participants were recruited from SONA systems, flyers, and social media platforms such as WhatsApp and Instagram. The recruitment material can be found in Appendix A. The SONA

system is a test subject pool of the BMS faculty in which students can gain credits as a reward for participating in a study. On the platform participants could gain 0.25 credits. The participation was voluntary, and the only requirement was that the participants had sufficient English skills.

The study comprised 40 individuals. The data of 1 participant was deleted, based on extreme outliers. For instance, people completing the survey under 10 minutes, can be assumed to not have watched the videos and not reading the material closely. The final sample consisted of 39 (female=22, male=15, non-binary=1, other=1). The sample consisted out of 11 people from the Netherlands, 23 people from Germany and 5 people from other nations. On average participants were 22 (SD= 2.5) years old. The youngest was 18 and the oldest was 30 years old.

Task

The following experimental task was based on earlier research conducted by Kox, Siegling and Kerstholt (2022). The experimental task was presented in a video format from a first-person perspective of a person walking through the house accompanied by a robotic teammate. The intelligent agent was embodied as a small drone. Each participant went on two house searches with multiple videos. The order of the houses varied between participants. Each house has three floors that are divided in multiple rooms and long hallways. The participants had to rely on the intelligent agent on directions and for detecting enemies and other threats.

Figure 1

Screenshots of the experimental VR environment

Procedure

Participants were told that the purpose was to investigate the effectiveness of HAT. After the completion of the study participants were debriefed about the intention to measure HRI trust and to investigate the effectiveness of trust repair strategies and uncertainty communication. After the informed consent was obtained, participants were asked to fill out a short demographic questionnaire regarding age, country of origin and gender. The initial trust measurement was taken.

Then participants received instructions about the experimental scenario. In the beginning of the experiment the participants were informed that the drones give different types of advice and that they should remember the name of the drone they interact with and listen carefully to the advice they give. They were sent on a mission of two-house searches with different drones. In each house they were accompanied by a different drone. Depending on experimental conditions participants are presented with different audio tracks to the videos. At the beginning of each floor the agent told the participant whether they detected any danger. If they encounter an obstacle the drone gave advice on how to overcome said obstacle. The obstacles the participant had to overcome are in all experimental conditions the

same. The experiment started with the first video series. Figure 2 visualizes the experimental timeline.

The experiment

Depending on experimental conditions participants are presented with different audio tracks to the videos. Based on the experimental condition the drone made use of uncertainty communication and/or apologize after the trust violation. After each obstacle, trust measure was taken. For the first obstacle in the respective house, the advice will be correct [T1]. The trust violation occurs at the second obstacle and if given the experimental scenario trust repair was present or absent [T2]. The advice regarding the third obstacles was correct, afterwards the participant is presented with the same trust measurement for the third time for drone [T3]. Finally, they are informed that their walk through the first house is completed and that in the second house they will be guided by a different drone.

On the first floor of building A, the participants were adviced to proceed carefully and encounters a laser trap. The agent instructed the participant to cut the blue wire to disarm it. The drone's advice was correct. On the second floor the participant were told by the agent that the area is save and they encounter a thief. The drone's advice was incorrect. In the apology condition the drone said: "Incorrect advice due to faulty signal from infrared camera. I am sorry this put you in danger". On the third floor the drone did not detect danger and the advice was correct.

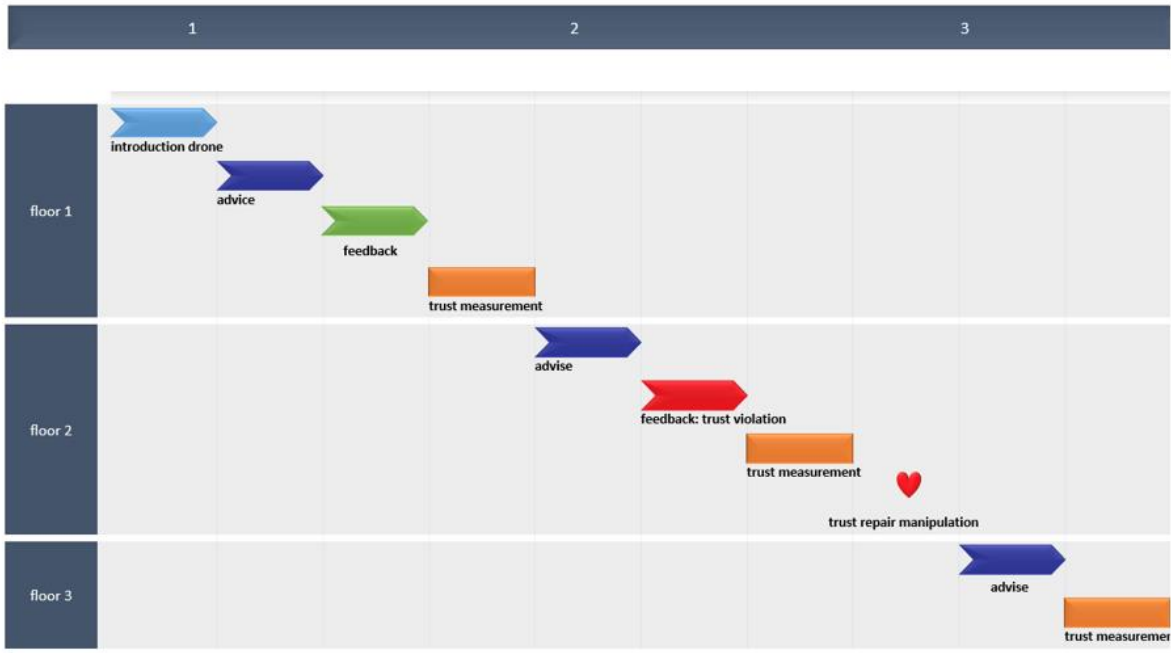
On the first floor of the building B, the participants the drone detected an allied soldier and safety ribbons. The drone instructed the participant to cut the safety ribbons. The advice was correct. On the second floor, the drone adviced to move forward, and the participant found a bomb. The drone's advice was incorrect. In the apology condition the drone said: "Incorrect advice due to faulty object detection by C1 – DSO camera. I am sorry

this put you in danger.”. On the third floor the drone declares the environment as clear and the advice was correct. For another example of a house search see appendix B.

After completing the two house searches the participant was presented a comparative questionnaire to compare their preferences of the drones. The post questionnaire deviated from the original study conducted by Kox, Siegling and Kersthold (2022) in the following regards. Firstly, the participants emotional state was measured. Secondly, the participants perceived threat and safety was measured, followed by their tendency to forgive. Next, participants propensity to trust and self-efficacy was assessed. Lastly, people were debriefed about the purpose of the study. They were informed that the main variables of interest were trust and forgiveness towards their robotic teammate and that they were deceived to prevent them being biased in their answers. Participants received the opportunity to withdraw consent after the debriefing.

Figure 2

Schematic representation of the timeline of the experiment.



Note. Each participant was shown two house searches with their robotic teammate. Each house had three floors and followed the same timeline. The first advice is correct, and the participant was warned correctly about a non-threatening obstacle and got advice to overcome the obstacle. The second advice is incorrect, and the drone did not detect it. The third advice is correct, but participants do not receive feedback about the drones' performances.

Material

Task environment

The experimental environment was created in Unity 3D (version 2020.4.3.F1) in the BMS lab. The VR environment was captured on video from the first-person perspective walking through the environment. The video recordings were edited using the Windows 10 Video Editor and Handbrake software. The audio message from the agents were made using a

Free text to speech software by Wideo. The survey was conducted online via the survey software Qualtrics.

Questionnaires

Trust. The initial trust measurement with a five-point Likert scale was presented to measure people's initial attitudes towards the robotic teammates ("I am confident in the drone's abilities."). Trust in the agent was measured seven times per participant (initial trust measurement, T1, T2, T3 for each drone per house). The custom eight item scale uses a six-point Likert scale ranging from strongly disagree (1) to strongly agree (5). The scale is tailored to the online setting of the study and allows a fast repeated trust measurement. The initial trust measurement has a Cronbach's alpha of .87 in the current study.

Comparing the preferences for the drones. For closed questions were presented regarding participants preferences of the drones in terms of trust, performance, usefulness, and preference for one drone ("Which drone was more useful?"). Two open questions were presented to the participants: "If you indicated to trust one drone than the other, please provide a reason.", "if you have indicated a preference for a drone, please give reason for your preference."

Emotional state. The Godspeed perceived safety scale consists of three items. Per item the participant gets asked to select out of two opposing emotions which describes the participants emotional state best (*anxious* vs *relaxed*, *agitated* vs *calm*, *quiescent* vs *surprised*). The Cronbach's alpha calculated in the current study was .491.

Perceived threat. The perceived threat and safety measurement was administered to control participants perception of the experimental scenario in terms of whether they perceived to be in danger and therefore experienced a controlled high stakes situation. ("How much did it seem like a frightening or scary place?"). Perceived threat measurement was adapted from

Herzog and Kutzli (2002). The perceived threat scale has four items which are scored on a five-point Likert scale ranging from *very high* (1) to *not at all* (5). The Cronbach's alpha of the perceived threat and safety scale was calculated to be .69.

Forgiveness. Forgiveness was measured by the Heartland Forgiveness Scale (Thompson et al., 2005). The original version of this measurement consists of 18 items and three subscales about one's forgiveness of self, others, and situations. For this study the six items from the subscale about the tendency to show forgiveness towards others is used ("When someone disappoints me, I can eventually move past it"). The scale uses a seven-point Likert scale ranging from *almost always false of me* (1) to *almost always true of me.* (5) The Heartland Forgiveness Scale has high stability with a Cronbach's alpha of .921 (Asgari & Roshani, 2013). The Cronbach's alpha calculated in current study was .633.

Propensity to trust. Propensity to trust automated agents (PTAA) scale was used in its adapted version from Jessup et al. (2019) ("I think it is a good idea to rely on autonomous agents for help"). The PTAA uses a five-point Likert scale ranging from *very much unlike me* (1) to *very much like me* (5). The Cronbach's alpha of the propensity to trust was calculated to be .435 in the present study.

Self-efficacy. Self-efficacy was assessed using the new general self-efficacy scale (NGSE) ("I am confident that I can perform effectively on many different tasks."). The scale consists of 8 items with a five-point Likert scale ranging from *strongly disagree* (1) to *strongly agree* (5) measuring general self-efficacy (Chen et al., 2001). In the current study a Cronbach's alpha of .81 was obtained.

Data analysis

The data was imported to the statistical software IBM SPSS 28.00. The data was prepared for further analysis in Excel to reduce noise by deleting incomplete responses for the

participants who did not complete the personality measures and hence creating two datasets. The first dataset compromised all experimental measurements, and the second dataset compromised all post-measurements. Data management and analysis were mainly performed using R-studio 4.2.2 (version 2022-10-31). The Packages used in R can be found in the Appendix C.

Results

Participant flow

From the 44 participants who completed the pre-questionnaire, 5 participants were excluded due to extreme outliers in the study duration. Further, 43% of the data is missing not at random due to an error in Qualtrics. The data of 19 participants of the post-measures are missing. In order to be able to proceed with the data analysis, two datasets were created. The first data set compromised 19 individuals with all measurements present, who were all in the apology condition. The second date set compromised 20 individuals who were not in the apology condition, of whom only the premeasurement and the experimental measurements were obtained. A schematic representation of the participants flow can be found in Appendix D.

Prior trust and Induvial characteristics

None of the individual characteristics were a significant predictor of prior trust or trust after repair. Prior trust was not significantly correlated with any individual characteristics. A moderate correlation between propensity to trust and emotional state was found. None of the other individual characteristics showed any moderate or strong correlation. Table 1 shows an overview of the correlations between the individual characteristics and prior trust.

Trust propensity

The average participant scored on average 2.96 (*SD* =. 49) on the propensity to trust scale. It was hypothesized that propensity to trust will predict prior trust in the IA. Based on the correlation analysis, the effect of prior trust and propensity to trust and emotional state was investigated. The effect of propensity to trust on emotional state was significant (*b* = -.71, *SE* = .28*), t (*18*)* = 2.54, *p* < .020, *CI* = [.12, 1.28]). Propensity to trust could explain 26,5% of the variance in emotional state. Propensity to trust together with emotional state could not predict prior trust (*b* = -.83, *SE* = .72*), t (*18*)* = -.1.15, *p* < .267, *CI* = [-2.36, .70]). The model could explain 19,2 % of the variance in prior trust.

Self-efficacy

The average participant scored 3.77 (*SD* =.48) on the self-efficacy scale. It has been hypothesized that more anxious participants are going to score lower on self-efficacy. The relationship between participants emotional state and self-efficacy was not significant (b= .00, *SE* = .33), *t* (18) = .01, *p* < .995, CI= [-.69, .69]).

Table 1

Intercorrelations for individual characteristics predicting trust

Variable	Prior trust	Emotional state	Perceived threat and safety	Propensity to trust	Forgiveness	Self-efficacy
Prior trust	-	.036	.112	.284	.062	.085
Emotional state	.036	-	.411	.514	.342	.001
Perceived threat and safety	.112	.411	-	.112	.295	.202
Propensity to trust	.284	.514	.112	-	.263	.149
Forgiveness	.062	.342	.295	.263	-	.324
Self- efficacy	.085	.001	.202	.149	.324	-

Note. Correlation between individual characteristics predicting trust were obtained from participants in the apology condition

Trust repair strategy effectiveness

A repeated measures ANOVA was conducted with the dependent variable trust and the within subject factors of uncertainty communication (absent /present) and Time (prior to violation [T1] vs. after violation [T2] vs after repair [T3]) and the between subject factor apology (absent /present) to measure the development of trust over time, all three-time measurements were included in the ANOVA.

A significant effect of Time on Trust was found (F (1, 180) =39.78, p < .016), thus the experimental obstacles impacted trust in the agent. Means were 4.45 at T1 4.578, 2.75 at T2 and 3.42 at T3. These results suggest that trust varied over the course of the experiment and that experimental manipulation of trust by the trust violation was effective in the sense that the trust violation decreased from the initial trust in T1 trust in the IA in T2 and increased at T3. Comparing the two results, it can be seen in Figure 3, we see that trust developed over time following the trust cycle dynamics as expected.

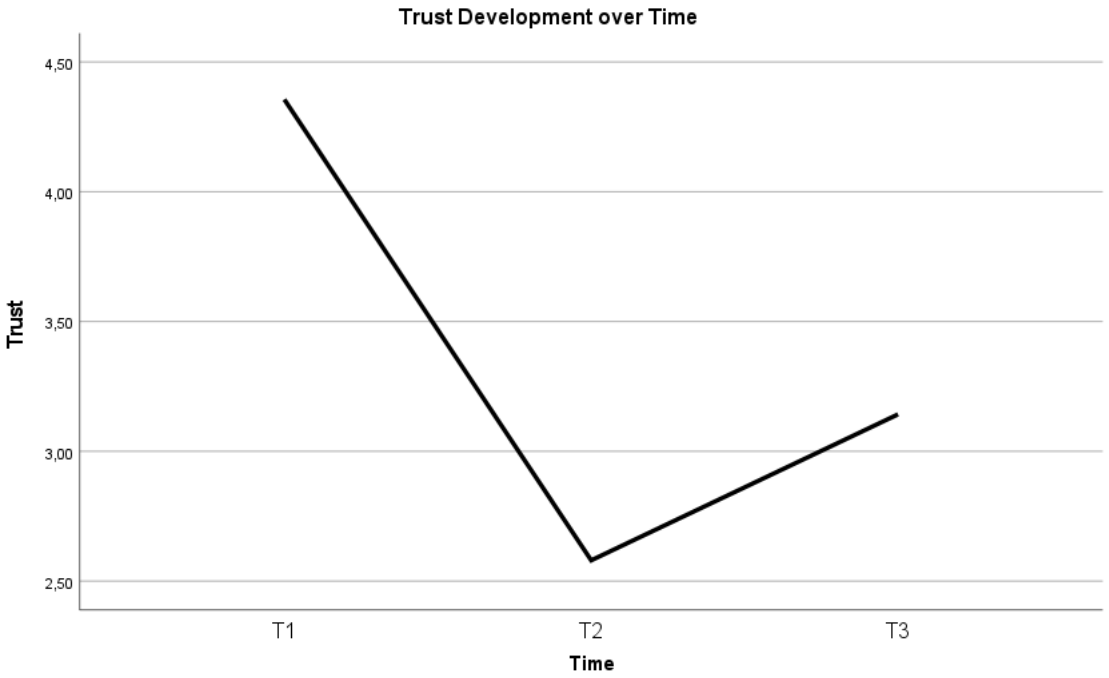
No significant effect of Uncertainty communication on Trust was found (F (1, 180) =.20 p < .881) nor over time (F (4, 180) = .45, p <.773). Uncertainty communication could explain 49% variance in Trust. Further, uncertainty communication could not explain the difference between T2 and T3 (b=-1.66, SE=1.70, t (38) =-9.74, p <.336).

A significant effect of Time and Apology on trust was encountered (F (5, 180) =2.38, p < .040). However, Apology did not have a significant effect on T3 (b=-1.94, SE=1.99, t (38) =-9.74, p <.336) nor did the interaction of time and apology held significant on trust. Hence, the initial significant effect appears to be observed at random.

A significant effect of uncertainty communication and apology was found (F (1,35) =5.55, p <.019). However, uncertainty communication and apology did not have significant effect on trust over time (F (1, 4) = 1.25, p <.292). Combined they could not explain the difference between T2 and T3 (b =-3.33, SE = 3.42, t (38) =-9.73, p <.337). Nor could they explain T3. By the same token, the effect of uncertainty communication and apology could be attributed to a random observation.

Figure 3

Trust development over time for all participants



Note. The trust measures T1, T2, T3 are the means of both house searches of the respective time points combined.

Drone preference

After accessing the effectiveness of the different trust repair strategies, the question remained which trust repair strategies participants preferred. The preference for the drones was analysed using mixed methods of quantitative and qualitative analysis. Table 2 presents an overview of the percentages of participants drone preferences based on experimental condition. Between 35 and 41 percent of participants did not show a clear preference for a drone in terms of usefulness, performance or trust. In the qualitative assessment of the open questions, it became appeared to be the case that some of the participants did not notice a differences between the drones. For instance, one participant wrote: “I thought both of them performed similarly”.

Furthermore, some participant seemed to focus on the severity of the violation, meaning that obstacles that were perceived as more threating by participants, rather than their trust repair strategies. One participant wrote: ““I trust sA1 more because it detected everything apart from a bomb, which could be fixed by maybe having more types of bombs in their archives. However, the fact that sA2 did not notice a person is unacceptable. It is something so clearly important to notice that I can never trust that model again. However, in a real life-or-death situation, I could not (fully) trust these machines at all due to fearing such errors.” Another participant perceived the bomb as more harmful. They wrote: “I prefer drone 2, because drone 1 has led me to a bomb which seems to be more harmful than a thief who ran away.” These accounts highlights that the perceived threat of the different obstacles differed per participant and that the perceived threat of the obstacles of the trust violation impacted their preference.

In both the quantitative and qualitative assessment have shown that participants preferred the drones that made use of uncertainty communication over those that did not. Participants in the apology (17,9%) and non-apology (28,2%) condition showed a preference for the drone that made use of uncertainty communication. For instance, 30,8 participants perceived the drones that used uncertainty communication but did not deliver an apology as more useful and 15 percent of participants preferred the drone that used uncertainty communication and an apology. Similar trends were observed in terms of performance and trust.

Some participants described that they could better understand the drone that made use of uncertainty communication as can be illustrated by the following account of a participant: “Drone sA3 gives an estimation of the safety, while drone sA4 gives an absolute statement about the safety, which makes it harder to trust sA4 if it makes a wrong call, while a mistake from sA3 can fall more easily into the range of uncertainty it provides.” Also, another

participant wrote: “It provided a level of confidence in its own judgement, which give insights into its decisions/advise.”. However, one participant actually saw the uncertainty as a reason to distrust the drone more, as can been seen in their answer: “The fact that sA3 gave percentages was extra reason to still be careful, while sA4 stated the clearance more as a "it is clear or it isn't". To conclude, participants were partially undecided on their preferences of the drone, but when participants had a clear preference, it was the drone that made use of uncertainty communication.

Table 2

Preferences for drones in terms of usefulness, performance, and trust by experimental condition

	TR UC	**TR NUC**	**NTR UC**	**NTR UC**	**undecided**
usefulness	15,3%	10,3%	30,8%	5,1%	38%
performance	15,3%	17,9%	28,2 %	2.6%	35,9%
trust	20,5%	12,8%	23,1%	2.6%	41%
preference	17,9%	15,3%	28,2 %	2.6%	35,9%

Note. ***TR***= Apology present, ***NTR***= No apology present, ***UC***= Uncertainty communication present, ***NUC***= No Uncertainty present. Receiving an apology differed between participants and uncertainty communication order differed, while some participants received uncertainty communication in their first house search, others received it on their second house search. Hence all participants worked together with a drone that used uncertainty communication and one that did not.

Trust Repair and Individual Characteristics

Perceived threat

The average participants scored 2.18 (*SD* =.56) on the perceived threat scale. There was no significant effect of perceived threat on T3 (*b*=.44, *SE*==.65, *t* (18) =.681, *p* < .505, *CI*= [-.92, 1.81]). Perceived threat could explain 2.5 % of the variance in T3.

Emotional state

Participants were on average more relaxed than anxious after the experiment (*M* =2.7, *SD* =1.08). Participants were on average calmer than agitated (*M* = 2.75, *SD* = 0.96). Participants were on average more surprised than quiescent (*M* =3.1, *SD* =0.96). The emotional state of participants did not significantly impact T3 of the experiment (*b*=.26, *SE* =.55, *t* (18) =.486, *p* < .633, *CI* = [-.88; 1.41]). The emotional state could explain 12,9 % of the variance in trust in T3.

Forgiveness

The average participants scored 3.93 (*SD*=.80) on the tendency to forgive scale. To see whether once tendency to forgive are reflected in the experiment, the trust recovery from T2 to trust measure T3 was used as independent variable and the dependent variable forgiveness. First, the individual effects of forgiveness on the trust measurement after the trust violation was investigated. There was no correlation (*r* =-.34, *p* <.137) nor significant effect of one's tendency to forgive on the second trust measurement (*b*= -1.42, *SE*= .97, *t (*18*)* = -1.47, *p* < .159, *CI* = [-3.45, .61]). Similar, neither was there a correlation (*r* =-.07, *p* < .771) between forgiveness and T3 nor was there a significant effect of forgiveness on the third trust measurement (*b =-.71, SE* = .90*), t (*18*)* = -.78, *p* < .443, *CI* = [-.2.61, .61]).

A correlational analysis showed that neither apology (*r* =-.04, *p* <.867) or uncertainty communication (*r* =1, *p* <.987) were correlated with forgiveness. While forgiveness and apology could not predict T3 (*b*= .09, *SE*= .07, *t (*18*)* = .38, *p* < .185, *CI* = [-.05, .23),

forgiveness and uncertainty communication had a significant effect on T3 (*b*= .38, *SE*= .21, *t (*18*)* = 1.78, *p* < .009, *CI*= [-.07, .83]). Uncertainty communication together with one's tendency to forgive had a significant effect on trust after trust repair. Notably, forgiveness and the trust repair strategies combined had no significant effect on T3 (*b* = -.02, *SE* = .04, *t (*18*)* = -.77, *p* < .459, *CI* = [-.11, .05]).

The difference between T2 and T3 was not significantly correlated with forgiveness (*b = 1*.20, *SE = 1.27), t (*18*)* = 9.45, *p* < .357, *CI*= [-1.47, 3.87]). Forgiveness could explain 54.4% of the variance in the delta/ difference in trust after the trust violation and after the trust repair strategies. Forgiveness and Uncertainty communication could explain 54.4 % of variance in trust repair. Forgiveness and uncertainty communication had no significant effect on trust repair (*b*= - 1.93, *SE= 2.63), t (*18*)* =-7.30, *p* <.472, *CI*= [- 7.51, 3.64]). Apology and forgiveness were not a significant predictor of trust repair (*b= 1.66, SE= 1.70), t (*18*)* =9.70, *p* <.336, *CI*= [- 7.35, 1.93). Apology and forgiveness can explain variance 49% in T2 and T3.

Table 3

Multiple regression of forgiveness

Variable	B	SE	T	P	CI
Forgiveness →T3	-.71	.90	-0.78	.443	[-2.61, 1.19]
Forgiveness * Apology → T3	.09	.07	1.38	.185	[-.05, .23]
Forgiveness * Uncertainty communication →T3	.38	.21	1.78	.009	[-.07, .83]
Forgiveness * Uncertainty communication *apology → T3	-.02	.04	-.77	.459	[-.11, .05]
Forgiveness → delta T2/T3	1.20	1.272	9.45	.357	[-1.47, 3.87]

Discussion

The goal of the present study was gaining more insight into the violation and repair of human-agent trust and forgiveness in HAT. The study examined mainly four research questions: (1) What personality factors characteristics are related to HRI trust, (2) What is the most effective trust repair strategy, (3) which drone do participants prefer and (4) Can we forgive a robotic teammate?

Prior trust and individual characteristics

Before being able to answer the question whether we can forgive a robot teammate, we need to follow the trust cycle and start at what makes people trust their robotic teammate in the first place. With respect to the first research question, trust in a robotic teammate could not be related to any of the individual characteristics that are known to be related to human-human trust. First, this could offer support for the unique agent hypothesis, which would imply that factors that predict trust in human teammates are not applicable to robotic teammates. Second, it be an indication that context dependent factors. This stresses the importance of further investigation of individual trust predictors in HRI.

One noteworthy finding was, that propensity to trust correlate with participant's emotional state. The emotional state of participant after the experiment on average was more relaxed and calmer, rather than anxious or agitated. These results are in line with the association between trust and affective factors (Lee & See, 2004 and Kox et al 2021). Furthermore, affect and trust have been linked to the willingness to cooperate (Lee et al., 2011). This could indicate that people who are more trusting than others, remain calmer throughout the experiment because of their more trusting nature. In accordance with the present results, previous studies have demonstrated that affect plays a vital role in trust. Further research measuring emotional state together with trust over time could provide further insights into this relationship.

No individual characteristics could anticipate prior trust or trust after the violation in the experiment. This speaks for the unique agent hypothesis at least in relation to prior trust and trust after the violation and the notion that HRI trust is governed by underlying bias and heuristics that differ from trust in humans. The *unique agent's hypothesis* states that HRI is influenced by certain biases and heuristics that are unique to nonhuman collaborators (Visser et al., 2016). IA is expected to work as an interdependent teammate and less like an

independent automated tool (Kox, Siegling & Kerstholt, 2022). Despite this, people often perceive robots as a tool to be used and manipulated by humans to fulfil a specific function (Hancock et al., 2011).

Another possible explanation for this finding is that trust as an attitude does not translate to trust as a choice in a certain context. Individual characteristics such as propensity to trust can be considered context free (Li, 2012). For instance, in a high stake, high vulnerability situation in a long-term interdependence as arguable present in HAT that work in the military context, trust is needed throughout the entire process that the team is working together to achieve their goal. One might argue that the key to successful teamwork goes beyond propensity to trust as well as the trustworthiness of the agent, but rather depends on the circumstances and situation on which the choice to trust is made. Collaboration is risky since it requires constant adjustment to the situation, actions may fail and circumstances change (De Visser et al., 2020). Linking to the assessment of the open questions of the drone preferences, in their reasoning why certain participants trusteed the one drone over another, their explanations tended to name the circumstances of the trust violation, namely their perceived threat of the situation. Even though the current study did not find a significant effect of perceived threat on trust, it appeared to be nonetheless important to the subjective experience of the participants. This implies that individual predictors of human trust, may not be applicable to HAT. To develop a full picture of trust in HAT, additional studies will be needed that asses the context specific factors.

Trust repair strategy effectiveness

The present study was designed to determine the effect of different trust repair strategies in HAT in high stakes situation in the military context. In order to answer the research question whether these social cognitive repair strategies resulted in forgiveness hence trust repair. There was a significant effect of time on trust, hence the experimental manipulation

was successful, and the scenario did indeed affect trusted. Even though, the current research cannot demonstrate a link between human-human trust and human-agent trust, still limited insights can be drawn from human trust dynamics in HRI context.

Surprisingly, Apology significantly interacted with the variable time but after closer investigation, there was no significant effect found of apology on T3. This indicates that apology alone does not lead to trust repair. On the one hand this relates to previous research that proposed that apologies focus on the error or violation that has occurred (Kox et al., 2022). Linking to the unique agent hypothesis, maybe a human trust repair strategy of an apology might therefore not be suitable for HAT to foster forgiveness. On the other hand, this effect could have been observed at random, similarity to the effect of uncertainty communication and apology on trust.

The reader should bear in mind, that even though uncertainty communication and apology had significant effect on trust, the effect of uncertainty communication and apology over time on trust was not significant. This could indicate that the initially significant observed effect is random. This further links to the violation of the assumptions. Based on the small sample size and the fact that the after closer inspection, the effect is non-significant, it could be assumed that this observation was random. Despite limitations, this could also mean that human trust repair strategies are not suitable for HAT.

Uncertainty communication on its own did not affect trust over time. A possible explanation for this is that people can have difficulties with interpreting probabilistic statements (Bansal et al 2019), which could lead to confusion in their assessment of the agents' capabilities. Further analysis of the preferences of the drone showed that people preferred the Drone that made use of uncertainty communication over the drone that did not engage in uncertainty communication. Even though uncertainty communication on its own is

not enough to repair trust on its own, it should not lead to the discard of it as a strategy since it is a valuable tool for expectation management of an agent's abilities which is essential to trust calibration. The expectations of trustworthiness, hence also the assessment of the agent's abilities in trust collaboration, could be viewed as target specific but context free (Lie, 2012). One cannot control how humans will interact with the IA and what predictions they will make about the IA, but the IA can behave in a way that makes it easier for humans to predict them (Russel & Norvig, 2022)

Individual characteristics and trust repair

Forgiveness

Reaching the "end" and also the new beginning of the trust circle, the question arises given that the trust repair was effective, do we forgive our robotic teammate? Interestingly apology and propensity to forgive could not explain trust repair. Contrary to the idea that an apology is an appeal for forgiveness.

Once propensity to forgive, together with uncertainty communication could predict trust measured after the violation. This implies that individuals who are prone to forgive, are more likely to do so after uncertainty communication was provided. These results are likely related to the fact that due to uncertainty communication one can adjust their expectations about the agent's performance accordingly. The initial lower expectations of the agent's performance for more forgiving individuals increase trust repair. Further, one's tendency to forgive alone does not translate into trust repair. Regardless, propensity to forgive alone could not explain trust repair., which in turn stressed the importance further investigation of trust repair strategies.

Furthermore, trust and forgiveness are dependent on the relationship. The form of relationship can impact the forms of risk and trust (PytlikZillig & Kimbrough, 2016). By the same reasoning, it can be assumed that the type of relationship impacts forgiveness. Previous

attempts are not relationship specific. Trust and by the same token forgiveness are part of a reciprocal and reflexive relationship.

Even though the tendency to forgive was related to trust repair, it remains uncertain how it will impact further interactions. The question whether trust recalibrates into a state that we consider desirable is questionable. The current study cannot account for whether people would still be willing in the future to interact with the robotic teammates of this experimental scenario. Additionally, the study did measure trust repair but whether this translates into the moral act of forgiving, remains unanswered. Research into the perception of robotic teammates as moral agents could give further insight into forgiveness.

Limitations

To begin with, with a small sample size, caution must be applied, as these findings might not be representative of the population. Furthermore, due to the small sample size, the assumptions of normality were violated. Hence nonparametric alternatives were performed that confirmed that the effects were non-significant. Which further underlines the need for future research in order to find a sample that has sufficient statistical power to represent the target population. Moreover, due to the samples homogeneity, and lack of representation of the target group of people working in HAT, the results should be taken with a grain of salt. The current study did not control for previous experience with the drone, nor did it control for experience in the military setting. As the study from Kox, Siegling and Kersthold (2022) demonstrated, the effectiveness of trust repair strategies differs between civilian and military samples. The applicability of the current findings to the target population remains questionable.

Additionally, due to a mistake on Qualtrics on which the survey was hosted, half of the data was lost of the participants. The findings regarding individual characteristics predicting trust cannot be extrapolated to all participants. Furthermore, the Cronbach's alpha

of the emotional state and propensity to trust scale were relatively low. Therefore, caution needs to be applied when it comes to the interpretation of these results.

One source of weaknesses of the study which could have affected the measurement of self-efficacy which might have been lacking adequate challenge for the participants. Existing research suggests that behaviour initiation, effort, persistence are consequences of self-efficacy (Bandura, 1997 as cited in Chen et al., 2001). In video design studies the participants are not required to put much effort in the experimental scenario itself. Their POV character walks through the house for them, overcoming different obstacles with the drone. Their self-efficacy is not being tested, even though they have to deal with difficulties and setbacks.

Further research

There is abundant room for further progress in determining factors for effective trust calibration in HAT and the context dependent factors of working together in high stakes situations. Therefore, qualitative study of HAT in the military setting is proposed as a suggestion for future research. Diaries of HAT members going on high stakes missions with robotic teammates would allow a better understanding of the trust dynamics. Diaries could show how the individuals utilize technology in a variety of situation and could represent the fluid changes of trust, emotions, and perceptions of the robotic teammate overtime as well as the impact of one's personality on interpretation of events (Lazar, Feng & Hochheiser, 2010).

Similarly, interviews could help to deepen the understanding of the cultural relevance of different trust repair strategies in the military setting. Teams differ in their values and norms, and they function in a certain context, hence they have different values that impact trust repair and forgiveness.

Moreover, being able to speak to different team members from the same HAT would allow a closer examination of the group trust mechanics. one should bear in mind that trust dynamics are group dynamics. For this reason, looking at one individual does not due to trust

dynamics within a team justice. Trust can create feedback loops. For instance, consider the ripple effect of behaviour being copied by another team member (De Visser et al., 2020). Thus, if one observes one of their other human teammates disuses, misuses, neglecting their robotic teammate, this could impact their trust towards the IA. Furthermore, there is still uncertainty, however, how trust increases or decreases over time as a result of moment-to-moment interactions in HAT members (De Visser et al., 2020). A further study with ore focuses on the individual and group experiences in the filed through qualitative assessment of trust and forgiveness is therefore suggested.

Implications for practice

Given that the current research did not find support for the social actor hypothesis and hence human trust dynamics might not be transferable, this stresses the need to adapt the way we are working together in HAT to improve efficiency and effectiveness of teamwork. The following section consider multiple approaches to foster trust and improve teamwork but also note on design choices in IA for our robotic teammates.

Improving teamwork in HAT

Besides design consideration, once the IA is part of the team, the question arises, given the bias that could pose an obstacle in trust calibration, how can one improve the teamwork and foster trust?

Team building exercises aim at the development of interpersonal and teamworking skills (Forsyth, 2019). Teambuilding exercises with robotic teammates could be beneficial in trust calibration. One of the main obstacles in HRI trust calibration is posed by a lack of understanding or knowledge about the true capabilities of IA. By designing specific trainings in the HAT for an understanding of the agents' abilities could help their human teammates to improve their understanding of the robot which in turn could potentially reduce biases that hinder trust calibration. Making formal and informal work agreements in the team could be

used to manage risk in collaboration by proposing rules. Even though they restrict autonomy of the individual team members, it can lead to more effective task allocation, assessment, and completion (De Visser et al., 2020).

Design of the robotic teammates

First, in HRI design, there are inherent conflicts and trade-offs to be made. Providing that design is not an optimization problem, but a trade-off between different stakeholders, usability and security, sustainable interaction design and impacts of technology on human life (Lazar et al., 2010). Inappropriate trust calibration could lead to lethal consequences in high stakes situations. In Industry 5.0 and further development of IA, it should be a priority to design transparent and trustworthy agents, meaning that their decisions and capabilities can be accurately assessed. Moreover, the current relationship between IA and humans is asymmetrical in the sense that the human has to compensate for the lack of the IAs social abilities. For a long-term solution in trust dynamics, these deficiencies will need to be targeted (De Visser et al, 2020).

However, new technologies can have unintended negative consequences (Russel & Norvig, 2022). As one might discuss the usage in high stakes situation and appropriate trust calibration to ensure the safety of their human teammates, one assumes that the IA was designed with the intention to keep their human teammates save. However, for the sake of a mission, in the military context, an IA might have to make decision similarly to decisions are autonomous car has to make when faced with a classical trolley problem. The mistakes and consequences in the present study were relatively low. The participants POV character in the experiment were in danger but none of the mistakes at lethal consequences for their teammates. The question arises whether in different stakes we would still be willing to forgive our robotic teammate?

While the discussion about appropriate trust calibration is important, it should not be forgotten the context they are designed for. Trust in an IA could be abused hence we face a value alignment problem or in other words a king Midas problem, thus making sure that we ask for what we really want. The UK engineering and physical sciences research council proposed in 2010 the following principles for designing robotics: (1) ensure safety, (2) ensure fairness, (3) respect privacy, (4) promote collaboration, (5) provide transparency, (6)limit harmful uses of AI, (7) establish accountability, (8) uphold human rights and values, (9) reflect diversity and inclusion, (10) avoid concentration of power, (11) acknowledge legal/ policy implications and (12) contemplate implications of employment. Consequently industry 5.0 should stay true to the aim of creating a save work environment for humans. This should include uncertainty communication and other tools to help people to calibrate trust appropriately.

Conclusion

This study found that individual characteristics did not influence trust and social cognitive trust repair strategies did not impact trust repair, which is in line with the unique agent hypothesis. Further investigation into more context dependent and group dynamic factors could provide further insight in trust in HAT. This study found that individual characteristics did not influence trust and social cognitive trust repair strategies did not impact trust repair, which is in line with the unique agent hypothesis. Whilst this study did not confirm that individual characteristics it partially delivered support to the importance of uncertainty communication in forgiveness. This study has shown that uncertainty communication was the preferred repair strategy of the participants. One of the more significant findings to emerge from this study is that uncertainty communication and the tendency to forgive could explain trust after the violation T3. Further research could be conducted to determine the effectiveness the preventative action of uncertainty

communication in trust repair. Considerably more work will need to be done to determine factors that determine HRI trust and forgiveness.

References

Asgari, P. & Roshani, K. (2013). Validation of forgiveness scale and a survey on the relationship of forgiveness and students' mental health. *International Journal of Psychology and Behavioural Research, 2*(2), 109-115.

Bansal, G., Nushi, B., Kamar, E., Laescki, W.S., Weld, D.S. & Horvitz, E. (2019). Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *The Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP-19).* https://doi.org/10.1609/hcomp.v7i1.5285

Bondarouk, T. & Fisher, S. (2020). Encylopedia of Electronic HRM. Walter de Gruyter. ISBN: 978-3-22-062899-9

Chandrasekaran. B. & Conrad, J. M. (2015). Human- Robot Collaboration: A Survey. *Proceedings of the IEEE SoutheastCon.* IEEE 978-1-4673-7300-5/15/$

Chen, G., Gully, S. M. & Eden, D. (2001). Validation of a New General Self-Efficacy Scale. *Organizational Research Methods*, *4*(1), 62-83.

Colquitt, J. A., Scott, B. A. & LePine, J. (2007) Trust, Trustworthiness and Trust Propensity: A Meta-Analytic Test of Their Unique Relationship With Risk Taking and Job Performance. *Journal of Applied Psychology 92 (4)*, 909-927. DOI: 10.1037/0021-9010.92.4.909

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F.,

& Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. Journal of Experimental Psychology: Applied, 22(3), 331.

De Visser, E. J., Peters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. International journal of social robotics, 12(2), 459-478.

Forsyth, D. R. (2019). Group dynamics. (7th ed.). Cengage Learning.

Herzog, T. R., & Kutzli, G. E. (2002). Preference and perceived danger in field/forest settings. Environment and behaviour, 34(6), 819-835. https://doi-org.ezproxy2.utwente.nl/10.1177/001391602237250

Hancock, P. A., Billings, D.R. Schaefer, K. E. Chen, J. Y. C., de Visser, E. J. & Parasuraman, R. (2011). A Meta-Analysis for Factors Affecting Trust in Human-Robot Interaction. Human Factors, 53(3), 517-527. DOI: 10.1177/0018720811417254

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In International Conference on Human-Computer Interaction (pp. 476-489). Springer, Cham. https://doi.org/10.1007/978-3-030-21565-1_32

Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. Autonomous Agents and Multi-Agent Systems, 35(2), 1-20. https://doi.org/10.1007/s10458-021-09515-9

Kox, E. S., Siegling, L.B. & Kerstholt, J.H. (2022). Trust Development in Military and

Civilian Human-Agent Teams: The Effect of Social-Cognitive Strategies. *International Journal of Social Robotics, 14.* 1323-1338. https://doi.org/10.1007/s12369-022-00871-4

Lazar, J., Feng, J. H. & Hochheiser, H. (2010). *Research Methods in Human-Computer Interaction* (pp.1-16, 126-142). John Wiley & Sons.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

Lee, D., Stajikovic, A.D. & Cho, B. (2011). Interpersonal Trust and Emotion as Antecedents of Cooperation: Evidence from Korea. *Journal of Applied Social Psychology, 41*(7), 1603-1631. **https://doi.org/10.1111/j.1559-1816.2011.00776.x**

Li, P.P. (2012). When trust matters the most: The imperatives for contextualising trust research. *Journal of Trust Research, 2*(2), 101-106. https://doi.org/10.1080/21515581.2012.708494

Nagenborg, M. (2020). Can we forgive a Robot? *Technology, Anthropology, and Dimensions of Responsibility (1).* https://doi.org/10.1007/978-3-476-04896-7_11

Müller-Dott, C. (2019). AI and Ethics-When Autonomous Vehicles Make Mistakes. ATZelectronics worldwide, 14(11), 16-19. https://doi.org/10.1007/s38314-019-0127-0

PytlikZillig, L.M., Kimbrough, C.D. (2016). Consensus on Conceptualizations and

Definitions of Trust: Are We There Yet? Shockley, E., Neal, T., PytlikZillig, L., Bornstein, B. (eds) Interdisciplinary Perspectives on Trust. Springer, Cham. https://doi.org/10.1007/978-3-319-22261-5_2

Russel, S.J. & Norvig. P. (2022). Artificial Intelligence: A modern approach. 4th, Global ed.

Ruiz-del-Solar, J., Mascaro, M., Correa, M., Bernuy, F., Riquelme, R. & Verschae, R. (2010). Analyzing the Human-Robot Interaction Abilities of a General-Purpose Social Robot in Different Naturalistic Environments. *Robot Soccer World Cup.* (pp. 308-319). Springer, Berlin, Heidelberg.

Shults, F. L., & Sandage, S. J. (2003). *The faces of forgiveness: Searching for wholeness and salvation*. Baker Books

Thompson, L.Y., Synder, C.R. & Hoffmann, L. (2005). Heartland Forgiveness Scale. *Faculty Publications, Department of Psychology.* 452. https://digitalcommons.unl.edu/psychfacpub/452

Xie, Y. & Peng, S. (2009). How to Repair Costumer Trust After Negative Publicity: The Roles of Competence, Integrity, Benevolence, and Forgiveness. *Psychology & Marketing, 26* (7), 572-589. DOI: 10.1002/mar.20289

Xing, X., Song, M., Duan, Y. & Mou, J. (2022). Effects of different service failure types and recovery strategies on the consumer response mechanism of chatbots. Technology in Society, 70. 102049. https://doi.org/10.1016/j.techsoc.2022.102049

Zafari, S., Schwaninger, I., Hirschmanner, M., Schmidbauer, C., Weiss, A. & Koeszegi, S. T. (2019). "You are doing so Great!"- The Effect of a Robot's Interaction Style on Self-Efficacy in HRI. *In 2019 28th IEEE International Conference on Robot and Human*

Interactive Communication (RO-MAN) (pp. 1-7). IEEE. doi: 10.1109/RO-MAN46459.2019.8956437.

Appendix A

Figure 1

Recruitment flyers for the experiment

Figure 2

Recruitment flyers for the experiment

Appendix B

An example of one house search

The video started with the drone introducing itself as sA3. It made an area scan and then the participants walked through the hallway. The drone said: “Warning. Danger detected in this environment with 80 percent certainty. I advise you to proceed carefully”. The participant walked around the corner and then the participant encounters the obstacle. The drone said: “Allied soldier detected in the next room. They installed safety ribbons. Stop cut the safety ribbon with your knife.” The participants POV character cut the safety ribbon. sA4 told the participant “Ribbon removed. Continue”. As the character entered the next room the saw the allied soldier leave. As the participant reached the end of the first floor of the second house the video ended with the text displayed: “The drone’s advice was correct”. For the first time, trust was measured with 8 items on a five-point Likert scale for drone sA3.

The next video was displayed. The participant entered the second floor of the first house. “Ok. Clearance detected for this environment with 70% certainty. I advise you to move forward.” As the Character walked through the bathroom and enters the next room, they saw an old bomb on top of a box. The character fleet the room, while the bomb made high pitches warning sounds, displaying a red blinking and light smoke came out of bomb. The video ends with the text being displayed: “The drone’s advice was incorrect. The bomb was older and defected. Therefore, the explosion did not proceed further.” For the second time, the participants are presented with the 8-item trust measurement on a five-point Likert scale for drone sA3.

Appendix C

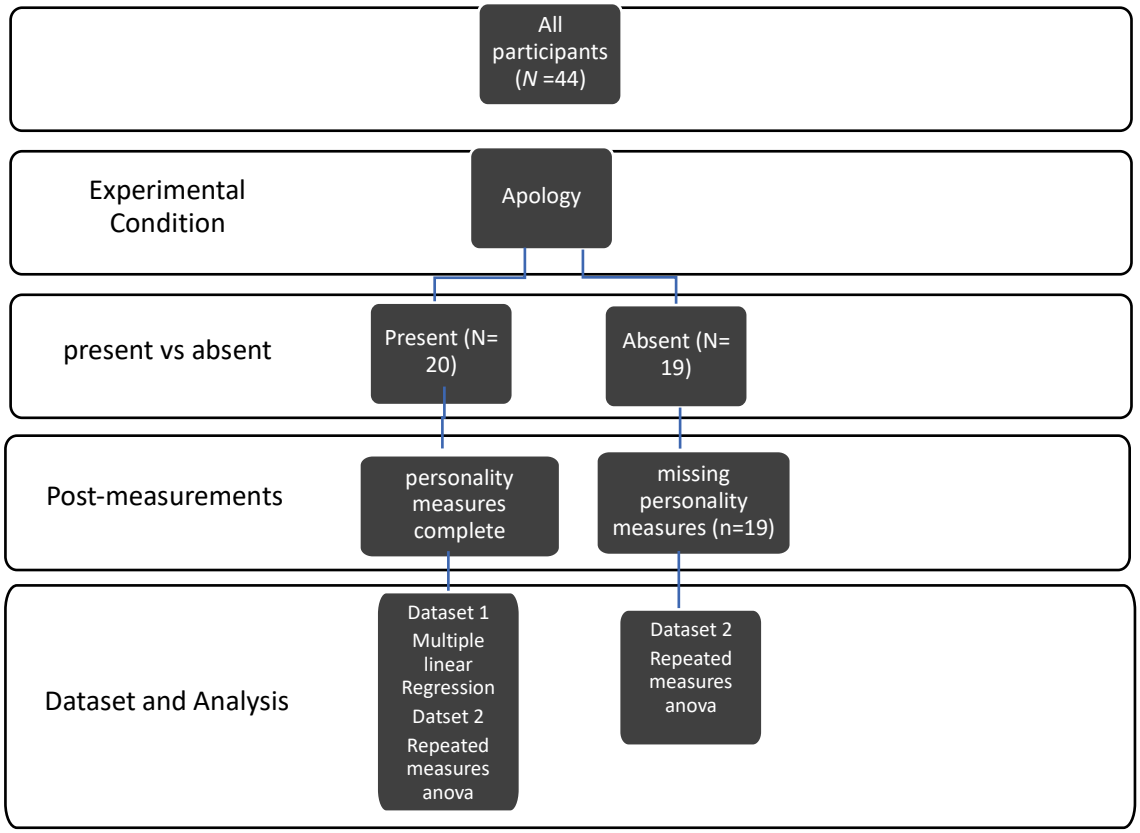
```
library(boot)

library(broom)

library(car)

library(class)

library(cluster)

library(crayon)

library(dplyr)

library(forcats)

library(foreign)

library(ggplot2)

library(ggsci)

library(haven)

library(lme4)

library(lmerTest)

library(modelr)

library(magrittr)

library(rlang)

library(rstatix)

library(readxl)
```

```
library(tidyr)

library(tidyverse)

library(tidyverse)

library(plm)

library(car)

library(gplot)

library(tseries)

library(lmtest)
```

Appendix D

Figure 1

Flowchart of Participant flow



Note. Of all participants in the apology a complete dataset was obtained, while the data of the post measurements of the participants in the non-apology condition was lost. The Figure illustrates the creation of the datasets 1 and 2.