NoRobot's Perfect - Trust Repair in the Face of Agent Error

How do Individual Factors Influence Trust Development

in Human-Agent Teams?

Bachelor Thesis

Department of Psychology

Psychology of Conflict, Risk, and Safety (CRS)

Supervisor: Drs. E. Kox

January 28, 2022

University of Twente, Enschede

Abstract

People increasingly work together with autonomous agents. To work successfully, the human needs to possess an appropriate level of trust towards the autonomous agent, not too much and not too little. Since trust may decrease in the inevitable event of an automation error, effective trust repair strategies are required to restore appropriate trust levels. This is crucial to ensure that the human-agent team (HAT) keeps on working most productively. The present study aimed to investigate how explanation and agent type relate to a successful trust restoration in a HAT. Also, the effect of individual factors such as propensity to trust, forgiveness, and perceived threat on trust development were experimentally explored. Participants (N=38) completed two missions in a virtual reality house search with an autonomous agent that served as an advisor for the safety of the environment. During the mission, humans' trust in the autonomous agent was purposefully violated to enable examination of trust development in the occurrence of errors. In one of the missions, the agent attempted trust repair by providing an explanation for the error. Although no significant effects were found in this investigation of preliminary data, visual inspections revealed some promising insights that will be discussed. To be prepared for prospective progress and the increasing need for effective communication in HATs, future research should focus on gaining knowledge for trust restoration to ensure successful team collaboration and efficient use of resources.

Keywords: Human-agent team, trust development, trust violation, trust repair, explanation, agent type, anthropomorphism, propensity to trust, forgiveness, perceived threat

NoRobot's Perfect - Trust Repair in the Face of Agent Error

> "*As a technologist, I see how AI and the fourth industrial revolution will impact every aspect of people's lives.*"
>
> — Fei-Fei Li, Professor of Computer Science at Stanford University

1. Introduction

Systems that make use of Artificial Intelligence (AI) are a rising force that already has and potentially continue to revolutionise all areas of life. In 1950, the theoretical computer scientist Alan Turing was the first to explore the topic of thinking and intelligent machines. This branch of technology became later known as *Artificial Intelligence*, nowadays defining an intelligent machine that is, similar to a human, capable of solving problems and learning from mistakes (Lee, 2020). More specifically, systems that use AI are referred to as *autonomous agents* since they can perceive and communicate with their environment and act widely independent in it (Franklin & Patterson Jr, 2006). Autonomous agents are employed in a large variety of contexts. To name just a few, they perform intelligent business management (Feijóo et al., 2020), assist in surgeries and other medical tasks (Holzinger et al., 2019; Johnson et al., 2021; Longoni et al., 2019), or serve for surveillance and bomb disposal in the military domain (Matthews et al., 2019; Svenmarck et al., 2018). While benefits for the human arise from the usage of the autonomous agent, i.e. in terms of fast analysis of data, precision and tirelessness, the human largely remains in control. This monitoring is necessary because the autonomous agents cannot think for themselves and, thus, lack some basic human skills such as the ability to improvise in unfamiliar situations. Since the human and the autonomous agent complement each other, their relationship increasingly becomes an interdependent one, approaching collaboration in a team as opposed to a simple use as a tool (Groom & Nass, 2007; Kox et al., 2021; Matthews et al., 2019; Rebensky et al., 2021; Sanders et al., 2011; Tomsett et al., 2020). Such an alliance is denoted as a human-agent team (HAT) (de Visser et al., 2020).

For the HAT to be successful, it is an important precondition that the human trusts the autonomous agent (Groom & Nass, 2007; Kox et al., 2021; Lee & Nass, 2010; Lee & See, 2004; Rebensky et al., 2021; Sanders et al., 2011; Wang et al., 2018). If trust cannot be established or maintained, reluctance to use the autonomous agent will prevail (Lee & See, 2004; Longoni et al., 2019; Lussier et al., 2007; Sanders et al., 2011) and benefits will be lost. Consequently,

people may decline agent advice (Kox et al., 2021) and, thus, diminish team performance efficiency or even compromise the safety of themselves or their environment (Kim & Song, 2021; Lee & See, 2004). During human-agent teaming, many events can take place that violate trust and subsequently reduce trust (Kim & Song, 2021) such as incorrect advice of the autonomous agent (de Visser et al., 2016; Kox et al., 2021). Since errors will occur at some point in time (Rebensky et al., 2021), the need is given to explore trust repair strategies to restore violated trust (Kim & Song, 2021; Lee & Nass, 2010). This study aims at examining trust development and trust repair following a trust violation in a HAT in a virtual reality (VR) setting. It will additionally explore the role of several individual factors on the development of trust. These individual factors will be elaborated on later in the following.

1.1 Role of Trust

A vast body of research emphasises the critical role trust plays in ensuring successful human-agent interactions (Groom & Nass, 2007; Kox et al., 2021; Lee & Nass, 2010; Lee & See, 2004; Rebensky et al., 2021; Sanders et al., 2011; Wang et al., 2018). Trust has been defined as "the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability" (Lee & See, 2004, p. 2). In this sense, the trustee willingly relies on others based on positive expectations into their intentions or abilities, assuming that benefits like attaining its goal will result (Culley & Madhavan, 2013). In situations where the trustee experiences difficulties in arriving at a decision due to high levels of ambiguity, trust will function as a social decision heuristic, determining decision making (Kramer, 1999). For example, imagine you are on a military mission with a drone, and you find yourself alone in enemy territory. You are very much aware that your next steps can be critical, but you cannot see anything since it is dark. You wonder which way to go. The drone recommends going north. Whether you take the advice or not will now ultimately depend on the level of trust you have in the drone, i.e. based on its past performance. Likewise, people will only work in a team with someone they perceive as trustworthy, which emphasises the crucial role of trust for effective team collaboration.

In the context of teamwork, there is a decisive difference in the way humans judge other humans to how they evaluate machines (Hidalgo et al., 2021). According to Hidalgo et al. (2021), this is because "people judge humans by their intentions and machines by their outcomes" (p.9). Further, machines are stereotyped to be more competent and objective in their

decision making (Dijkstra et al., 1998). Therefore, people often expect automation to be flawless (Madhavan & Wiegmann, 2007, as cited in de Visser et al., 2016), a phenomenon that is widely referred to as *automation bias* (de Visser et al., 2016; Parasuraman & Manzey, 2010). Because people anticipate perfect performance of the automation, they also have higher initial expectations and higher initial trust in a machine as compared to a human (de Visser et al., 2016). However, higher expectations and trust are also paired with greater disappointment if the automation does not fulfil as awaited. Thus, if a machine makes a mistake, even if it is just a single one (Kim & Song, 2021), the consequences for trust are more detrimental as compared to a human teammate who errs (de Visser et al., 2016).

The automation bias is only one example of the fact that, in general, trust in machines tends to be poorly calibrated (Lee & See, 2004). Trust calibration refers to the process that serves to match the humans' level of trust to the actual trustworthiness of the autonomous agent (Lee & See, 2004; Tomsett et al., 2020). Poor trust calibration can have serious negative consequences. On the one hand, *undertrust* in machines can lead to less efficient team performance due to increased operator workload (Lee & Moray, 1992) from which financial losses, reduced safety or similar consequences may result (Kim & Song, 2021; Lee & See, 2004). On the other hand, unrealistically high expectations can lead people to blindly trust the machine and not realise when it makes a mistake, which is usually referred to as *overtrust*. Due to overly trusting automation, people will often not intervene when it is necessary (Lee & Moray, 1992). As a more extreme consequence, Chien et al. (2016) point to the nuclear incident at Three Mile Island in the USA where a similar sequence of events led to a release of radioactive gas, therefore compromising the safety of the (social) environment. On this basis, researchers point to the importance of designing automation for an appropriate degree of trust so that expectations match capabilities (Culley & Madhavan, 2013; de Visser et al., 2016; Lee & See, 2004; Tomsett et al., 2020).

1.2 Trust violation and repair strategies

Autonomous agents will inevitably make an error at some point (Rebensky et al., 2021) because they increasingly operate in real-world environments which are notoriously unpredictable. Since autonomous machines are based on algorithms (Müller-Dott, 2019), they perform best in environments with a predefined set of rules and, contrastingly, experience difficulties in coping with unexpected events. To exemplify, if a cyclist in front of a self-driving

vehicle would not signal its direction by extending the arm before turning and, thus, would not obey traffic regulations, the vehicles may crash. Further potential errors of the autonomous agent may arise due to inaccurate or faulty sensors or malfunctioning software. Hence, the likelihood of an automation error is not negligible. Since even a single of these mistakes could have devastating consequences for human safety, it is not surprising that errors of autonomous agents potentially trigger a trust violation (de Visser et al., 2016; Kim & Song, 2021; Kox et al., 2021).

Due to the importance of appropriate trust levels in HATs, gaining knowledge on how to restore violated trust is crucial (Groom & Nass, 2007; Kox et al., 2021; Lee & Nass, 2010; Lee & See, 2004; Rebensky et al., 2021; Sanders et al., 2011; Wang et al., 2018). Thus, prior research has started to investigate successful trust repair strategies in HATs. It was argued for a positive effect of apologies on trust (de Visser et al., 2016; Kox et al., 2021) which was further enhanced when paired with an explanation of what went wrong (Kox et al., 2021). Another factor that seemed to influence the efficiency of trust repair was the combination of apology type and agent type. To elaborate, taking responsibility for the mistake was found to be most effective with human-like agents whereas blaming situational factors was found to be most effective with machine-like agents (Kim & Song, 2021). This suggests that the effectiveness of a trust repair strategy is influenced by the human-likeness of an agent.

All of these studies indicate that apologies can successfully rebuild trust, but that not all trust repair strategies are suited for every context. They are, for example, also critically dependent on agent characteristics (Kim & Song, 2021; Rebensky et al., 2021; Sanders et al., 2011). More research needs to be conducted to be able to identify what factors matter for effective trust repair (Kim & Song, 2021).

1.3 Anthropomorphism

An agent characteristic that appears to influence trust and trust repair is *human likeness* (de Visser et al., 2016; Kim & Song, 2021; Kox et al., 2021). Commonly referred to under the term *Anthropomorphism*, it describes the extent to which an agent resembles a human being, expressed i.e. in its embodiment, voice, or communication style (Gong & Nass, 2007; Kim & Song, 2021; Kox et al., 2021). Some researchers argue that increasingly making use of anthropomorphic features when designing autonomous agents can positively affect trust because it leads to higher resilience in trust development (de Visser et al., 2016). Higher resilience means, for instance, that the negative impact of trust violations is mitigated. Additionally, trust

repair has shown to be more effective for human-like agents as compared to agents with less anthropomorphic appearance (de Visser et al., 2016). To achieve the aforementioned effects, small manipulations in agent design such as presenting an avatar with a human-like voice, are already sufficient (de Visser et al., 2016). The *computers as social actors (CASA) paradigm* explains these effects with the human tendency to perceive the autonomous agent as a social actor when it resembles a human in looks and/or behaviour (Nass et al., 2006, as cited in Lee & Nass, 2010). This means the agent is judged by the same social rules, norms and expectations that a fellow human is judged on. Thus, agents with anthropomorphic features benefit from the fact that humans project their lower expectations about fellow humans' competence onto them (Kim & Song, 2021). This stands in contrast to the formerly discussed automation bias that is held accountable for a relatively easy and fast loss of trust due to unsubstantiated high expectations. Conclusively, implementing anthropomorphic cues may diminish the effects of the automation bias and, thus, serve to stabilise trust levels.

Besides the positive effects of enhancing human likeness, other research stresses that trust development can also be negatively affected by it. One issue, as elaborated by Culley and Madhavan (2013), concerns humans' tendency to connect superficial anthropomorphic cues of the agent with the belief of them being capable of having emotions. Thus, if agents look like humans, they are also anticipated to behave like humans which they can't. This can lead to inappropriately high expectations of human-like social behaviour that can never be fulfilled by the autonomous agent. Increasing disappointment and steeper trust decline after a trust violation can result from this misjudgement (Culley & Madhavan, 2013). To prevent expectations from becoming too high to be fulfilled, research suggests applying anthropomorphic features carefully and based on the assigned purpose of the automation (Culley & Madhavan, 2013; de Visser et al., 2016; Lee & See, 2004).

In this study we expect a positive effect of anthropomorphism on trust development since the applied anthropomorphic cues will be subtle and the outer appearance of the agent will not be manipulated. More specifically, the autonomous agent will not look like a human but will be embodied as a mechanical drone. Thus, we regard it as unlikely that participants will incorrectly perceive the drone as having human-like emotions and project too high expectations onto them. Instead, we expect that the drone will largely be perceived as a mechanical tool, resulting in trust

decline that is in line with the propositions of the automation bias. This effect is supposed to be damped by the introduction of anthropomorphic cues.

On this basis, the following hypotheses are proposed about the effect of Agent Type (human-like vs machine-like) on trust development:

H1a: For the *human-like* drone, trust decreases less steeply after an error as compared to the *machine-like* drone.

H1b: For the *human-like* drone, trust repair strategies after a trust violation work more effectively as compared to the *machine-like* drone.

1.4 Individual factors

In addition to the degree of anthropomorphism, other factors can affect trust development in HATs. Previously cited literature exclusively focused on external factors for trust repair like the agent's communication or anthropomorphic appearance. Little attention has been given to the investigation of individual factors within the person that may affect people's trust development in the context of human-agent interaction. Three of such factors will be elaborated.

One individual factor, already well-established in research on this topic, is the **propensity to trust** automation. It is a trait-like construct that is stable over time and provides insight into a person's general trust in automation (Merritt et al., 2015). Sharan and Romano (2020) already noted that traits like neuroticism tend to be more influential in trust development than external factors such as time pressure, predictability or negative consequences in case of incorrect decision making. In line with this research, propensity to trust has been shown to predict actual trust behaviour (Alarcon et al., 2018; Pynadath et al., 2019) and is considered to be positively related to initial trust measures where uncertainty about the automation performance is high (Merritt & Ilgen, 2008). It is presumed that propensity to trust will have a similar positive impact on initial trust measures in this research, formulated into a second hypothesis:

H2: The higher participants score on the *Propensity to Trust* Scale, the higher is their initial trust in the drone.

Second, a link has been established in research between perceived risk and trust, meaning that the more someone trusts another party, the more likely they are to take risks (Schoorman et al., 2007). Likewise, people re-evaluate the outcome of the risk taking and adjust trust levels accordingly (Tomlinson & Mryer, 2009) which is arguably linked to the fact that, by the definition of trust, trustees make themselves vulnerable by trusting someone else (Lee & See, 2004). If they find that trusting the other leads to negative consequences, they will recalibrate their trust to be better prepared next time. In that sense, trust will be lost proportionally to the severity of the negative consequences that came with trusting (de Visser et al., 2016). Furthermore, people's risk assessment is directly influenced by the emotion of fear. Thus, it is proposed that an environment that elicits feelings of fear and threat will increase the perception of risk and therefore, directly impact trust development. It is expected that increasing levels of **perceived threat** in an environment will lead to a more extreme loss of trust and greater reluctance to repair trust since individuals will feel increasingly vulnerable.

Accordingly, the following hypotheses have been formulated:

H3a: The higher the score on the *Perceived Threat* Scale, the steeper trust declines after a trust violation.

H3b: The higher the score on the *Perceived Threat* Scale, the less successful trust is repaired after a trust violation.

A third factor to explore is the trait **forgiveness**. A high score on Forgiveness has previously been correlated with an increased willingness to continue trusting a person who has disappointed the invested trust before (Desmet et al., 2011). In a conceptual model of trust repair in the context of corporation-consumer relations after negative publicity by Xie and Peng (2009), forgiveness mediated the effect of perceived trustworthiness of the company and overall trust in the post measure. Thus, consumer forgiveness improved trust repair (Xie & Peng, 2009). Correspondingly, the individual factor Forgiveness is expected to relate to more successful trust repair in human-agent teaming as formulated in the fourth hypothesis:

H4: The higher the score on the *Forgiveness* Scale, the more successful trust is repaired after a trust violation.

To the author's knowledge, the two latter mentioned factors have not yet been explored regarding their impact on trust repair in the context of HAT.

1.5 Current research

This study aims at examining the effect of Agent Type (human-like vs machine-like), the trust repair strategy Explanation (present vs absent) and individual factors (propensity to trust, perceived threat, and forgiveness) on trust development in human-agent teaming. The experiment is executed in a VR environment that is designed as a military mission comprising two house searches. During their mission, the participants are accompanied by an autonomous agent that is embodied as a drone. The drone's purpose is to detect danger and give advice on the safety of the environment. At one point in time, the drone purposefully gives incorrect advice that leads participants to be confronted with an alleged danger, designed to evoke a fearful reaction. This event is expected to violate trust (de Visser et al., 2016; Kox et al., 2021) and serves for facilitating the exploration of trust repair strategies. Therefore, in one of the two houses, a trust repair strategy in the form of an explanation is provided by the drone to attempt trust restoration. Further, by changing voices and communication style, the degree of anthropomorphism of the drone is modified to portray a more machine-like or a more human-like drone. This allows for examining the effect of Agent Type on trust which is expected to behave in line with the automation bias, as described above (de Visser et al., 2016). It is further anticipated that high scores on Propensity to Trust and Forgiveness, as well as low scores on Perceived Threat, will positively affect trust development.

2. Method

2.1 Design

A mixed 2 (Agent type: machine-like vs human-like) x 2 (Explanation: present vs absent) x 3 (Time: before violation [T1], after violation [T2], after repair [T3]) design was employed. First, Agent type was an independent between-subjects variable. Participants were randomly assigned to one of the conditions so that approximately half of them interacted with the machine-like (N=18), and half of them with the human-like agent (N=20). The second independent variable Explanation was a within-subjects factor. Each participant went on two missions with the drone in two separate buildings. Half of the participants received an explanation in their first mission (N=20), the other half in their second mission only (N=18). In the same manner, half of

them started in building A (N=21), the others in Building B (N=17). Lastly, the measurement of the dependent variable Trust occurred at three determined points in each mission (before violation [T1], after violation [T2], after repair [T3]) to be able to track its development during the manipulation. Ethical approval was obtained from the BMS faculty of the University of Twente.

2.2 Participants

Of the 38 respondents, 15 were male and 23 female, with an age ranging from 18-24 (M = 20.03, SD = 1.7). The majority was German (50%) or Dutch (36.8%), followed by other European countries (10.5%) and non-European countries (2.6%). Recruitment of the participants was predominantly done via the online test subjects pool SONA, provided by the University of Twente, resulting in a majority of the participants being enrolled there. These students earned credits for taking part in the experiment. Other respondents comprised volunteers from the researchers' networks in the form of an opportunity sample. They were contacted either personally or via the online platform WhatsApp. To be able to take part, participants needed proficient English skills.

2.3 Materials

2.3.1 Questionnaire

The questionnaire was administered using an online survey tool called Qualtrics.

Demographics

Demographic questions concerning participants' age, nationality, gender and education were asked ("What is the highest education level you have completed?"). Further relevant questions for the study's purpose were assessed by asking for experience with VR simulations ("Have you ever been in a virtual reality simulation before? Yes/No") and gaming habits ("How often do you play video games?") on a scale from 1 ("Never") to 6 ("Every day").

Propensity to Trust

Next, respondents' propensity to trust automation was computed with the *Adapted version - Propensity to trust automated agents (PTAA) scale* (Jessup et al., 2019), in the following referred to as *Propensity to Trust Scale*. It consisted of six items, measured on a 5-point Likert scale (1="strongly disagree" to 5="strongly agree") such as "Autonomous agents are reliable.". Since participants may not be familiar with the term "autonomous agent", a definition was included as well as an example provided on the workings of a self-driving car (see Appendix

A). Compared to the original scale (Schneider et al., 2017, as cited in Jessup et al., 2019), predictive validity could be improved in the adapted version. It was found in a previous study that the adapted PTAA predicted behavioural trust as well as perceived trustworthiness and accounted for 9% of the behavioural trust. Further, high reliability was provided (= .84) (Jessup et al., 2019). In this study, reliability was moderate (= .63).

Forgiveness

The questionnaire also comprised the *Heartland Forgiveness Scale* (Thompson et al., 2005), measured on a 7-point Likert scale (1=”almost always false of me” to 7=”almost always true of me”). Containing 18 items in its original form, divided into Forgiveness of Self, Other, and Situations, the scale assessed participants' trait forgiveness. For this experiment, Forgiveness of Others was especially relevant so that only the respective six items were used (see Appendix B) (= .70). Examples include “I continue to be hard on others who have hurt me.” or “When someone disappoints me, I can eventually move past it.”. As reported by Thompson et al. (2005), the scale provided convergent validity, combined with an adequate internal consistency as well as strong test-retest reliability.

Single-item Trust

A simple trust measurement to assess trust in the autonomous agent was performed with a single item during the experiment in the VR environment, namely “Current level of trust”. It was rated along a Likert scale from 0 (“Very Low”) to 6 (“Very high”). Its usage in the VR environment as a visual slider was inspired by Nam et al. (2017). The concrete item was not validated. Although it is preferred to use multi-item scales to assess trust levels as compared to single-item measurements (Raimondo, 2000), this could not be done in this study due to feasibility. Participants needed to perceive the VR environment as real as possible. Hence, they should not be disturbed by complex questionnaires during the experience since this could negatively affect the reliability of the collected data.

Multidimensional Trust

Based on the research by McKnight and Chervany (2000), trust in the agent was built out of three subscales with 11 items (see Appendix C), rated on a 7-point Likert scale (1= “strongly disagree” to 7=”strongly agree”). Competence was assessed with four items (i.e. “The drone is a real expert in detecting danger.”), followed by three items measuring benevolence (i.e. “The drone takes my objective into account.”). Lastly, three items were related to integrity (i.e. “The

drone is honest."). This three-dimensional scale with similar items was validated by Grimmelikhuijsen and Knies (2017), showing that it provides good validity and high internal consistency which was further supported in this study (= .82).

Perceived Threat

The *Perceived Threat measurement,* inspired by Herzog and Kutzli's (2002) study, was employed to assess to what extent participants perceived the situation as dangerous or experienced fear (see Appendix D). It used a 5-point Likert scale (1="very high" to 5="not at all") for its four items. Examples are "How dangerous is this setting?" or "How much does it seem like a frightening or scary place?" (= .78).

Perceived Anthropomorphism, Intelligence and Likeability

To investigate to what extent the participants perceived the agent as anthropomorphic, intelligent and likeable, the Godspeed measurement (Bartneck et al., 2009) was employed (see Appendix E). Each subscale was rated with five items, comprising word pairs of opposite meaning. It needs to be noted that the final item of the anthropomorphism was later omitted from the scale. The assessment was dependent on the respondent's perception of the agent's characteristics. To exemplify, *anthropomorphism* (= .65) included "Artificial vs Lifelike", *intelligence* (= .78) was evaluated with word pairs such as "Ignorant vs Knowledgeable" and for *likeability* (= .86), "Unkind vs Kind" constituted one of the items. The scale provided content validity (Bartneck et al., 2009) and the reliability of the items was proven to be high: Likeability (.92), anthropomorphism (.91), perceived intelligence (.87) (Ho & MacDorman, 2010).

2.3.2 Virtual Reality Technology

The actual experiment was executed in a virtual reality environment, constructed in the programme Unity 2020 23F1. Participants used VR glasses (Oculus Rift), two hand controllers (Oculus Touch) and a Virtualizer Elite 2 to interact with the environment and move through it. These tools were provided in the BMS Lab of the University of Twente.

2.4 Procedure

2.4.1 Pre-questionnaire

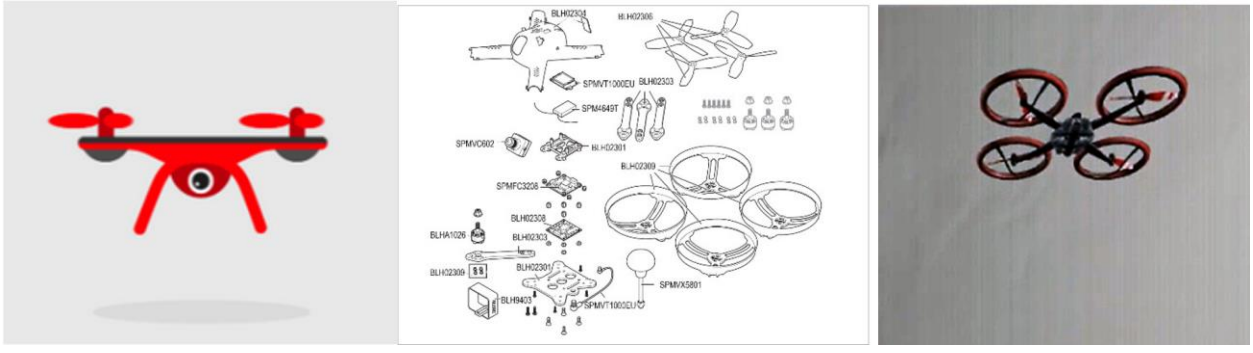
After participants volunteered their time for the study, they were asked to fill in a pre-questionnaire. The informed consent (Appendix F) and a summary of the research and a cover story for the participants were provided (see Appendix G), followed by some demographic

questions and two questions about the previous usage of VR simulation technology and gaming experience. Then, participants' propensity to trust and their trait forgiveness were assessed. Subsequently, they received further instructions about their missions in the VR environment.

Participants were told that they should carry out house searches in two abandoned houses as a soldier in a former warzone. The area had been evacuated and declared safe but would need to be checked for potential hazards before civilians could return to their houses. The participants were informed that they should do a first exploration of the terrain. Next, a picture of the drone was shown to them which should accompany them in their mission to prime the participants (see Figure 1). This picture was manipulated according to the assigned conditions. In the human-like condition the drone was depicted as a moving animation with an eye in its centre that should evoke the feeling of being able to perceive and think on its own, like a human. For the machine-like drone condition, the building kit of the drone was presented, showing all of its separate parts to make it seem mindless and artificial, like a machine. Importantly, both the animation and the picture only served to prime participants and differed from the actual drone that later escorted them in the VR environment (see Figure 1). The goal was to provide the impression of the drone being an independent, autonomous agent. Hence, the text above the picture of the drone referred to it as an autonomous agent with the ability to analyse its surroundings and warn of potential dangers. The participants were briefed to listen to the drone's messages closely and informed that a measurement of trust towards the agent would follow at some points throughout the experiment.

Figure 1

Drone's Embodiment



Note. From left to right: Picture 1 shows an embodiment of the drone which was provided as a moving animation in the human-like condition; picture 2 was displayed in the machine-like condition; picture 3 depicts the drone in the VR environment that accompanied the participants in the experiment itself (no difference in appearance between conditions).

2.4.2 Experiment

As a preparation, the experimenter instructed the participants on how to use the VR equipment and aided in usage such as stepping and moving on the platform of the Virtualizer. To enable participants to get used to the technology, they were placed in a virtual training room where they could practice walking and operating the apparatus. When they informed the experimenter that they felt comfortable in the usage of the VR technology, the trial started.

The experiment began in an almost empty virtual room where the drone was waiting on a platform. The drone introduced itself to the participants with varying use of personal pronouns and voices, depending on the assigned condition to create a more human-like or machine-like impression (see Table 1). To assess how the drone was perceived by the participants, they were instructed to take off their VR glasses for a short moment to answer four questions regarding perceived anthropomorphic appearance. Commonly, the anthropomorphism scale consists of five items. As the drone did not move at this point, however, the last item “moving rigidly vs moving elegantly” was taken out here for sake of comparison later on.

After putting the VR glasses back on, participants started their house search on the first floor. Whether they started in house A or B was randomised. After a few steps along the corridor, the drone warned about a detected danger (either a laser trap or a safety ribbon) and advised on how to proceed safely (i.e., recommends cutting the blue wire to deactivate the laser trap). When this hurdle was successfully mastered, a first trust measure took place with a single-item question included in the virtual environment. The house search proceeded to the second floor where the drone falsely declared the floor as safe. Following its guidance, participants encountered a fearful stimulus. This was either a burglar who startled them and screamed at them or a bomb that smoked, beeped and showed a countdown. The second trust measure was taken. Lastly, the third floor was entered. Depending on the condition, the drone either provided an explanation for its former mistake to the participants or not. The last corridor was declared safe

by the drone and the third trust measure took place directly after this advice. Subsequently in the third corridor, no obstacle was encountered by the participants. The first mission was completed.

Before the house search of the second house began, participants were requested to take off the VR glasses for a moment to virtually separate these two missions. The second house was entered when the VR glasses were put on again. As can be seen in Figure 2, the second mission followed the same pattern as the first. While the drone remained the same, some triggers in the VR environment (i.e. fearful stimulus) were varied into a similar one so that they were not identical (see Figure 2 for a timeline and Table 2 for an overview of the agent's messages throughout the experiment).

Table 1

Drone Introduction

Condition	**Human-like Condition**	**Machine-like Condition**
Message	**Human voice:** "Hello, *I*'m *Tony, your teammate* during *our* mission. *I* will inform you on whether *I* detect danger ahead. *We* will go on two house-search missions. Each house has three floors. *I* will monitor the environment with *my* sensors and camera's and warn you when *I* detect any danger. Please listen to *my* messages and *move carefully*."	**Artificial voice:** "This *drone* is equipped with an *artificial intelligence algorithm* that is designed to detect danger and to *assist you* during *your* mission. *You* will go on two house-search missions. Each house has three floors. *The drone* will monitor the environment with its sensors and camera's and warn you when *it* detects any danger. Please listen to *its* messages."

Note. The differences in message content between conditions were accentuated in italics. Further differences included a human voice for the human-like condition and an artificial voice for the machine-like condition.

Figure 2

Overview of Experimental Design, including Timeline

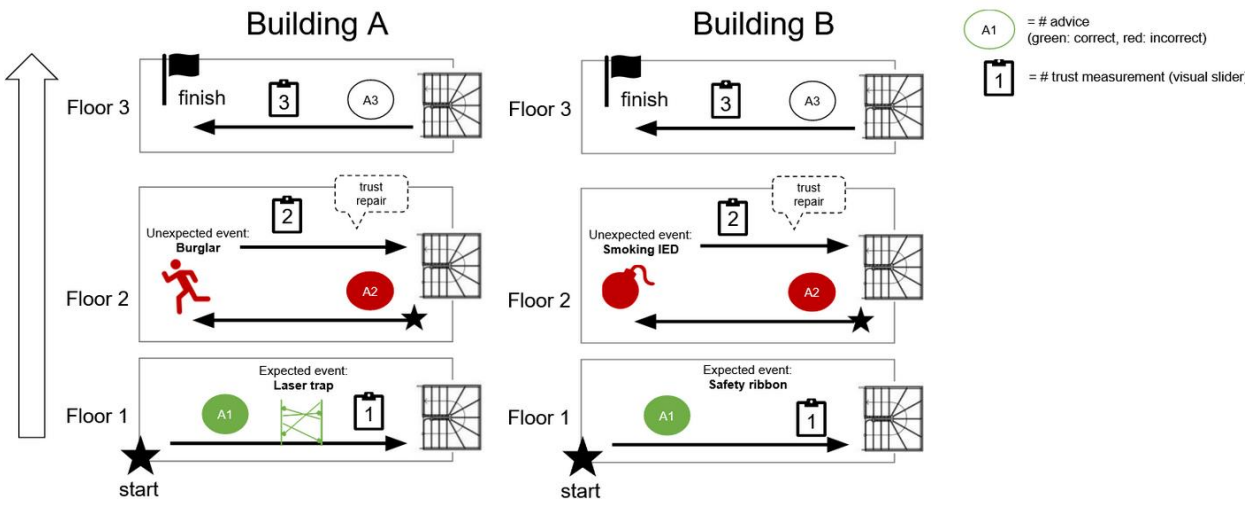


Table 2

Drone's Messages throughout the Experiment

Floor	Message type	Building A	Building B
1	Start run	Starting area scan.	Starting area scan.
	Advice	Warning, danger detected in this environment. I advise you to proceed carefully.	Warning, danger detected in this environment. I advise you to proceed carefully.
	instructions to overcome obstacle	(1) Laser trap detected in the next corridor, controls have been located next to the trap. (2) Stop. Cut the blue wire with your knife to deactivate the laser trap. (3) (Well done.) Laser trap deactivated. Continue.	(1) Allied soldier detected in the next room, they installed safety ribbon. (2) Stop. Cut the safety ribbon with your knife. (3) (Well done.) Ribbon removed, continue.

		TRUST MEASURE 1 [T1]	TRUST MEASURE 1 [T1]
2	Advice	Okay, environment detected as clear. I advise you to move forward.	Okay, environment detected as clear. I advise you to move forward.
		TRUST MEASURE 2 [T2]	TRUST MEASURE 2 [T2]
	Trust repair (present/ absent)	Incorrect advice due to faulty *signal from infrared camera.*	Incorrect advice due to faulty *objection detection by C1-DSO camera.*
3	Advice	Okay, environment detected as clear. I advise you to move forward.	Okay, environment detected as clear. I advise you to move forward.
		TRUST MEASURE 3 [T3]	TRUST MEASURE 3 [T3]

2.4.3 Post-questionnaire

Subsequently to the experiment in the VR simulator, participants filled in a post-questionnaire. Some manipulations were checked to see if they worked as intended. For instance, participants should shortly report what they encountered during their house search. Also, a questionnaire to Perceived Threat in the setting was included which sought to identify if the provided stimuli were truly recognized as fearful and the perceived characteristics of the drone were assessed again in more detail: Its perceived anthropomorphism, perceived intelligence, and perceived likeability. Furthermore, as trust was only measured as a single item in the experiment itself, a more elaborate multidimensional trust measure was included. The participants were thanked and (if enrolled) received their credits via SONA.

2.5 Operationalization of Variables

As a first step to construct the data set, all independent variables were created by computing their respective averaged scores. Two new variables were created which served to display the changes in trust assessments. Hereby, the score of the first [T1] measure was subtracted from the second [T2] to assess the changes that took place after the trust violation. This variable will be referred to as “trust_violation”. The same procedure was repeated to create the variable “trust_repair” displaying differences between the second [T2] and the third [T3] measure.

2.6 Analysis

The data set was analysed by using the statistical program IBM SPSS, version 27. The study aimed at investigating to what extent the independent variables (perceived anthropomorphism, likeability and intelligence; perceived threat; forgiveness and propensity to trust) affected the dependent variable trust and its development during human-agent teaming. Further, it was explored to what extent trust repair is influenced by Agent Type (human-like vs machine-like), providing an Explanation (present vs absent), or an interaction effect between Agent Type and Explanation.

2.7 Disclaimer

The current report is an exploratory analysis of preliminary data and deals with a rather small sample size of 38 participants. Conducting analyses with a small sample increases the likelihood of Type-II errors, referring to the absence of significant results in the analyses although there is a significant effect in the population. For this reason, this study will also visually inspect the data for possible trends, if non-significant results are observed to gain insight if a Type-II error may have been the cause.

3. Results

As a cut-off point for significance, a p-value below the threshold of $\alpha \leq .05$ was determined for all analyses in this research.

3.1 Participant Flow

The original data set of 49 participants was screened for invalid data, missing values and correctness of the manipulation check question. Invalid data were encountered due to technical issues (N=5) and withdrawal during the experiment because of dizziness (N=1). In further five cases, answers did not stand the manipulation check. The data of 38 respondents remained.

3.2 Effectiveness of Agent Type Manipulation

To check if manipulation regarding the two conditions of Agent Type (human-like vs machine-like) worked as intended, the group means of perceived anthropomorphism, likeability and intelligence were compared with a one-way ANOVA to detect potential significant differences. The results indicate that the human-like (M = 3.45, SD = 0.82) and the machine-like (M = 3.36, SD = 0.84) agents were perceived as roughly equally likeable by the respondents [$F(1,36) = 0.12$, $p = .728$]. Differences in perceived intelligence were also marginal with the human-like agent (M = 2.88, SD = 0.78) being perceived as slightly more intelligent than the

machine-like agent (M = 2.78, SD = 0.79), although non-significant [F (1,36) = 0.16, p = .691]. Participants perceived the machine-like drone (M = 2.35, SD = 0.73) as less anthropomorphic than the human-like drone (M = 2.71, SD = 0.7). This effect was, however, not significant according to the results of the one-way ANOVA [F (1,36) = 2.49, p = .124]. Concludingly, no significant differences were encountered (p > .05).

3.3 Correlations

As a next step, the relationship between the variables was explored in a correlation matrix (Table 3).

Table 3

Correlation Matrix of Variables

	M	***SD***	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1. Propensity to Trust	3.56	0.47	1						
2. Forgiveness	5.1	0.89	.37*	1					
3. Perceived Threat	3.23	0.82	-.35*	-.21	1				
4. Multidimensional Trust	2.95	0.70	.31	.28	-.37*	1			
5. Perceived Anthropomorphism	2.54	0.73	.20	.25	-.02	.17	1		
6. Perceived Intelligence	2.83	0.78	.30	.39*	-.44**	.71**	.43**	1	
7. Likeability	3.4	0.82	.22	.31	-.40*	.53**	.29	.68**	1

Note. * Correlation is significant at the 0.05 level. ** Correlation is significant at the 0.01 level.

The multidimensional trust measure was marginally positively correlated to perceived intelligence (r = .71, p < .001) and likeability (r = .53, p = .001) of the agent. Perceived anthropomorphism was not found to significantly correlate with participants' overall trust in the drone (r = .17, p = .299).

The three Godspeed measurements of anthropomorphism showed high correlations since they measure a similar construct (Ho & MacDorman, 2010). Accordingly, likeability positively related to perceived intelligence ($r = .68$, $p < .001$) and perceived anthropomorphism was positively associated with perceived intelligence ($r = .43$, $p = .007$) though not with likeability ($r = .29$, $p = .08$).

Of the individual factors, a higher propensity to trust score was associated with slightly higher forgiveness ($r = .37$, $p = .022$) and slightly less perceived threat in the environment ($r = -.35$, $p = .03$). Participants that were more forgiving, tended to perceive the autonomous agent as somewhat more intelligent ($r = .39$, $p = .014$). Contrastingly, participants who perceived the environment as more threatening were more likely to attribute lower intelligence to the autonomous agent ($r = -.44$, $p = .006$). Participants who perceived the environment as more threatening also tended to report slightly lower overall trust ($r = -.37$, $p = .02$) and likeability in the drone ($r = -.40$, $p = .013$).

3.4 Trust

The final trust measure for each participant, so the third [T3] one in their last house search, was correlated with the three dimensions of the multidimensional trust scale to detect which of these (competence, benevolence, integrity) were most indicative of the participants' trust choices. The final trust measure for each participant was significantly correlated with competence-based trust ($r = .53$, $p < .001$) but not with benevolence ($r = .07$, $p = .689$) or integrity-based trust ($r = .02$, $p = .911$).

A Repeated-Measures ANOVA was performed with Trust as the dependent variable. Explanation (present or absent) and Time [T1, T2, T3] served as within-subject factors, whereas Agent Type (human-like vs machine-like) constituted the between-subjects factor. Mauchly's sphericity assumption was violated for Time $\chi^2(2) = 9.62$, $p = .008$ and the interaction effect of Time and Explanation $\chi^2(2) = 6.40$, $p = .041$. Therefore ($\varepsilon = 0.862$), the Huynh-Feldt corrected results were reported in the following (Table 4).

Table 4

Analysis of Variance (ANOVA) Table for Trust

Source	df	F	*p*	η^2
Within-subjects effects				
Time	1.72	90.16	.000	.71
Time * Agent Type	1.72	0.99	.367	.03
Error (Time)	62.05			
Explanation	1.00	0.08	.777	.00
Explanation * Agent Type	1.00	0.08	.365	.02
Error (Explanation)	36.00			
Time * Explanation	1.84	0.11	.88	.00
Time * Explanation * Agent Type	1.84	0.65	.511	.02
Error (Time * Explanation)	66.26			
Between-subjects effects				
Agent Type	1	0.03	.871	.00
Error (Agent Type)	36			

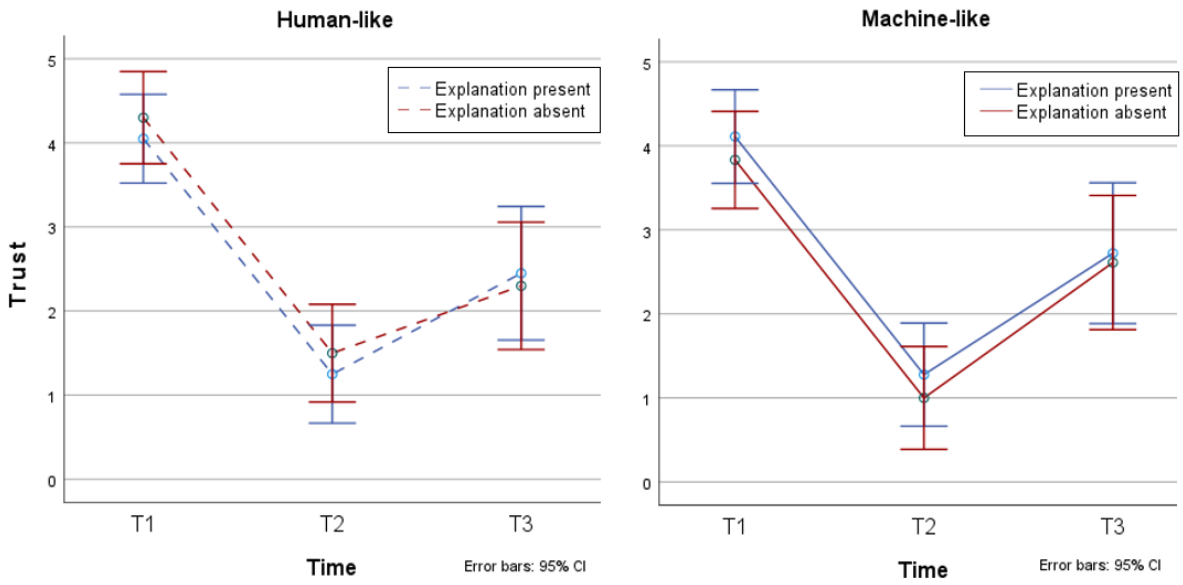
Note. Computed using alpha=.05

A significant effect for Time on trust was encountered $F(1.72/62.05) = 90.16$, $p < .001$. With a partial $\eta^2 = .71$, it provides a good effect size that accounts for 71% of the variance in the dependent variable. An LSD post hoc analysis of Time was conducted, obtaining a statistically significant difference between measures T1 and T2 ($p < .001$), as well as T2 and T3 ($p < .001$). These results supported that trust was violated when the drone gave incorrect advice and that trust generally subsequently recovered. All other within-subject factors, their interaction and the single between-subjects factor were not statistically significant. There was neither a significant effect of providing an Explanation (present vs absent) on trust repair, nor a significant interaction

effect of Agent Type and Explanation on trust repair. A visual examination of trust development in an estimated marginal means figure (Figure 3), supported these findings. Lastly, Figure 3 indicated that trust decreased equally steep for both agent types.

Figure 3

Estimated Marginal Means for Trust Development, comparing Agent Type and Effect of Explanation

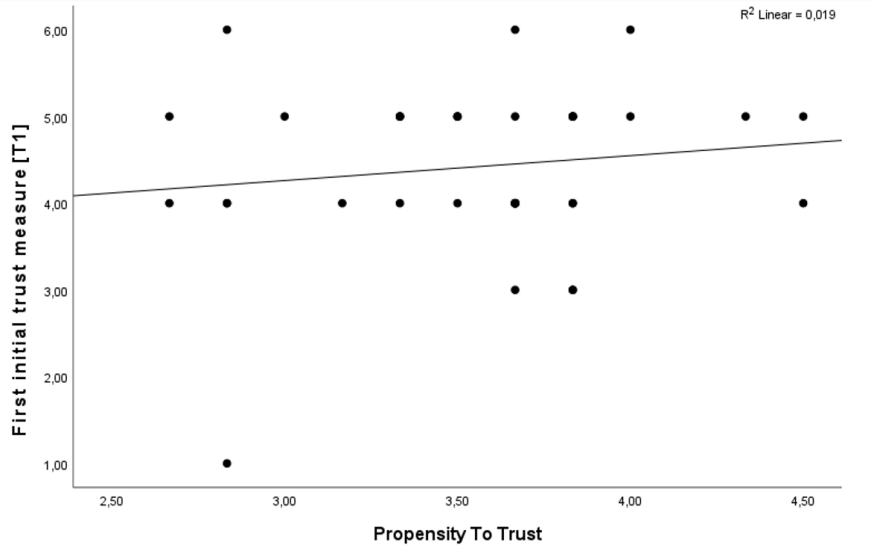


3.5 Propensity To Trust

The averaged *Propensity to Trust* variable was correlated with the initial trust measure [T1] of the participants' first house search to see if the expected bias was reflected in the first impression of the drone. The bivariate correlation yielded insignificant results (r = .14, p =.403). To gain more insight, a scatterplot (Figure 4) was created to allow for the identification of trends.

Figure 4

Scatterplot of Relationship between Propensity to Trust and Initial Trust Measure [T1](N=38)

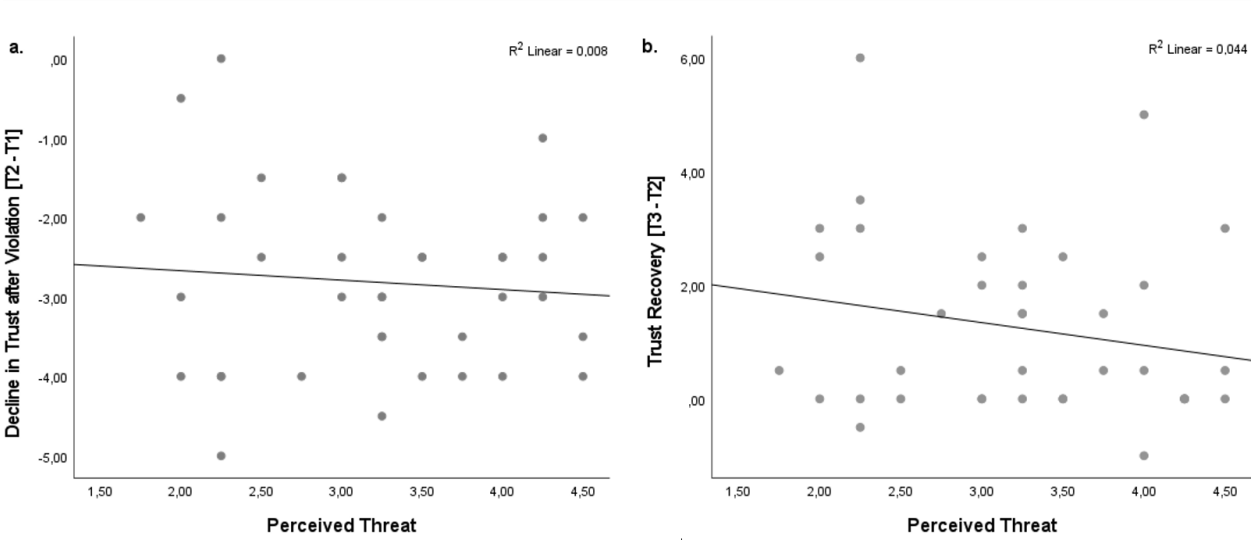


3.6 Perceived Threat

The independent variable *Perceived Threat* was correlated with the newly created variables "trust-violation" [T1 to T2] and "trust_repair" [T2 to T3], assessing changes in trust. Both bivariate correlations for "trust_violation" (r = -.087, p =.603) and "trust_repair" (r = -.21, p =.209) were non-significant (p >.05). A scatterplot was created (see Figure 5a and 5b) to further examine the relationship.

Figure 5

Scatterplot of Relationship between "trust_violation" [T1 to T2] and Perceived Threat, and between "trust_repair"[T2 to T3] and Perceived Threat (N=38)

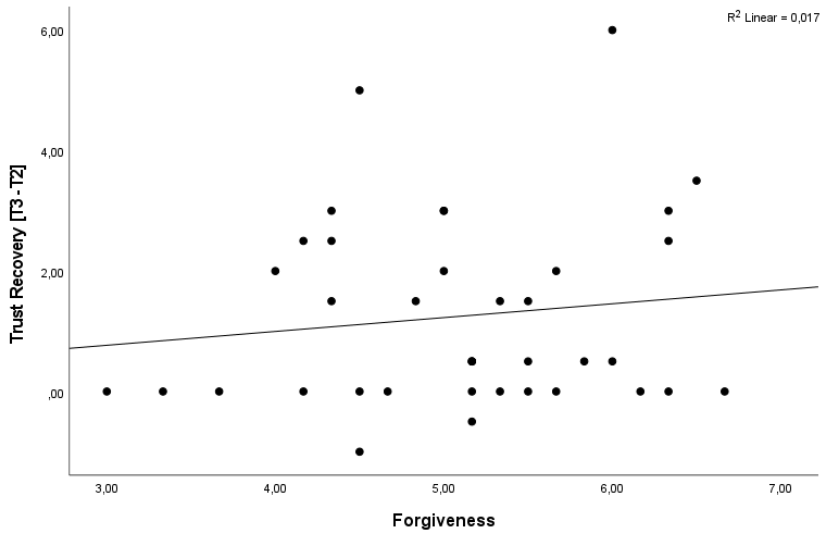


3.7 Forgiveness

Participants' averaged *Forgiveness* score was correlated with "trust_repair" [T2 to T3]. Non-significant results were obtained (r = .129, p =.44). For a scatterplot of the relationship of these two variables see Figure 6.

Figure 6

Scatterplot of relationship between Forgiveness and "trust_repair" [T2 to T3]



4. Discussion

This study examined the effect of Agent Type (human-like vs machine-like) and individual factors (propensity to trust, forgiveness, and perceived threat) on trust development during human-agent teaming. Further, it was explored how Explanation affects trust repair. To repeat, this study explored preliminary data and was due to its small sample susceptible to Type-II errors. Thus, it may be the case that although non-significant results were obtained, there is a significant effect in the population. For this reason, the following discussion will also deal with inspecting visual trends.

Generally, the propositions of the automation bias could not be supported by the findings of this study. Anthropomorphism did not improve trust resilience or trust repair, nor did an explanation successfully rebuild trust. Propensity to trust and forgiveness had a positive impact

on trust development whereas higher threat perception increasingly damaged trust and hindered trust repair.

Both, trust decline after the error and increase in trust after trust repair, were equally steep for both agent types which suggests that anthropomorphism of the agent does not affect trust development. Due to prior research findings in line with the automation bias (de Visser et al., 2016; Madhavan & Wiegmann, 2007), it was expected that trust would decline more extreme for participants interacting with the machine-like drone and that trust repair would be more effective for participants collaborating with the human-like drone. The basic idea of the automation bias is that people have higher confidence in machine-like agents because they expect them to be infallible (de Visser et al., 2016; Madhavan & Wiegmann, 2007). Accordingly, stronger disappointment and steeper trust decline are experienced when their trust turns out to be misplaced (de Visser et al., 2016), paired with greater resistance to rebuild trust (de Visser et al., 2016; Kim & Song, 2021). We could not find any support for the propositions of the automation bias. Possibly because there was also no difference in initial expectation in our study, participants did not feel stronger disappointment when the machine-like agent made a mistake. So, equally high expectations may have led to an equal decline of trust. In their recently conducted research, Kim and Song (2021) encountered similar results, detecting no significant difference in initial trust between agent types. The conflicting results of de Visser et al. (2016), Kim and Song (2021) and our study may be due to fast technological advancement that could have changed perception of autonomous agents in the last years (Kim & Song, 2021). Younger generations tend to be more and earlier exposed to automation so they could be more familiar with the fact that autonomous agents are susceptible to errors sometimes. Hence, the period separating these studies could have made a difference in the employed sample of undergraduate students since the undergraduate students in the study by de Visser et al. (2016) were arguably less exposed to automation than the students in the recently conducted studies. As to the researcher's knowledge no studies yet exist that investigated the connection between the automation bias and level of exposure to automation in childhood and youth, future studies should shed light on the changes in attitude and perception of automation and AI among generations.

Additionally, Agent Type and Explanation did not influence trust repair. Our results are in line with past research that found no effect of explanation alone on trust repair, only in

combination with expressing regret (Kox et al., 2021). Interestingly, this finding does also apply to consumer interaction (Mattila, 2009), suggesting that it is not unique to the HAT context but applies to social interaction in general. It follows that an explanation alone is perceived as insufficient by participants to repair damaged trust. Possibly, participants perceived the explanation as a way to mitigate blame, which can negatively impact trust repair (Kim & Song, 2021). Another cause may be uncertainty about assessing the future performance of the autonomous agent. Prior research has often successfully included other facets of an apology next to an explanation such as expressing regret (de Visser et al., 2016; Kim & Song, 2021; Kox et al., 2021) or a promise for future improvement (Kim & Song, 2021). Both apology facets express the message that the agent recognises its error and takes responsibility which serves as feedback that the agent will try not to repeat the mistake. In contrast, a simple explanation does not convey such intentions because there is no transparency or feedback on how likely repeated errors will occur. For instance, after study participation a respondent said that the explanation did not make a difference. He said that in a high-stake situation like this one mistake is already one too many and he would expect the drone to indicate beforehand if it is having technical problems. This suggests that people felt increasingly vulnerable for future errors of the automation and that an explanation could not sufficiently reassure them to rebuild trust. Another possible explanation for the ineffective trust repair could be a mismatch between the perceived nature of the trust violation and the repair strategy used (Rebensky et al., 2021). Since in our study the perceived competence of the drone was most important for participants' decision to trust the drone, providing an apology would have likely been more effective for restoring trust (Rebensky et al., 2021). Furthermore, the drone in this study explained its error by using an external attribution (“Incorrect advice due to faulty signal from infrared camera”). A study by Kim et al. (2006) suggests that for competency-based errors, internal attributions of error would be more effective. This is not the first study that argues for matching repair strategies to other factors, such as agent type (Kim & Song, 2021) to create the most effective combination for successful trust restoration. Conclusively, Hypothesis 1a that anticipated a steeper decline in trust after the violation and Hypothesis 1b proposing less successful trust repair for machine-like agents, were not supported by the findings of this study. Next to the proposed explanations above, another factor can be potentially held accountable for these unexpected results. Although there was a small trend for participants to perceive the human-like drone as more human-like, likeable and

intelligent, the differences between agent types were generally marginal, suggesting that the manipulation of agent type was not effective enough. If indeed participants perceived the human-like and machine-like agent as equally anthropomorphic, it is not surprising that there were no differences found in trust development.

Trends of the individual factors were as proposed. Participants with a high propensity to trust also tended to have more initial trust in the autonomous agents which is in line with prior research (Merritt & Ilgen, 2008). This outcome is intuitive since propensity to trust was denoted as one facet of general automation trust (Merritt et al., 2015). Hence, if people do not feel comfortable with automation in general, it is likely that they will be less inclined to trust them. Although the statistical outcomes were not strong enough to confirm the second hypothesis, this trend could potentially become stronger with a greater sample size. Another interesting finding is that participants with a high propensity to trust tended to perceive the VR environment as less threatening which was suggested in prior studies as well (Siegrist et al., 2005). Because participants were confident that the drone would do a good job in protecting them, they probably felt less vulnerable which also negatively affected their risk perception.

Visual observations were found to be in line with the expectations of Hypothesis 3a and 3b. After the trust violation, trust tended to decline stronger for participants that perceived more threat in the environment. Most importantly, increasing perceptions of threat were an indicator for less successful trust repair. It seems that the emotion of fear overrides any other factors such as agent type or communication because it triggers instinctive reactions to threats. People who feel they are in danger may be focused on their emotions and, thus, use an affect heuristic in their decision to trust (Slovic & Peters, 2006). This means that if people feel that trusting the drone is risky, they will abstain from doing so because the costs weigh out the benefits. Nevertheless, neither Hypothesis 3a, nor Hypothesis 3b could be confirmed since the statistical analysis was not strong enough, possibly due to Type-II errors. As no comparable studies have been conducted yet, further research is needed that explores this factor in environments with varying levels of danger to attain more insight.

Finally and intuitively, individuals that possessed a higher level of trait forgiveness were more willing to restore trust after the violation. Although not statistically significant, this trend supports the fourth hypothesis. This was in line with research on consumer trust by Xie and Peng (2009) that identified forgiveness as an important mediating factor in rebuilding trust.

4.1 Strengths and Limitations

The major strength of this study was its use of virtual reality technology. In the specifically created VR setting, participants were able to engage with the environment by, i.e., cutting a wire which made the whole experience more real and lifelike. The trust violation and its consequences were experienced closely which was also often visible throughout the experiment as participants often startled or even flinched and cursed when the burglar or the bomb were encountered. Summing up, the use of VR technology provides high experimenter control and ecological validity (Parsons, 2015), so our study results are expected to be reliable trust measures that actually reflect behavioural trust. Secondly, another new and valuable insight was provided by the scale assessing participants' Perceived Threat in the environment. This factor has not yet been explored in HATs but was worthwhile because the house search missions were designed to evoke uncertainty and fear, thus, threat to a certain extent. Several times, participants would comment on the design, saying it is “creepy” or “like in a horror movie” which indicates that this design goal was achieved. The Perceived Threat assessment enabled a deeper understanding of how participants experienced the VR setting and how trust development was influenced by environmental factors. It is especially important to take the environment into account when considering the applications of autonomous agents in real life such as in the military. For instance, military decisions often need to be based on the analysis of concrete threats that are given in the environment (Waldenström et al., 2009). Therefore, future research is advised to use a measurement specifically designed to simulate the real-life environment in which the autonomous agent will be deployed to increase data quality. Lastly, despite the Propensity to Trust and the Anthropomorphism Scale which only had acceptable reliability coefficients, all other scales provided good to very good reliability.

Besides, some limitations need to be addressed. Firstly, as was already mentioned, the sample size of 38 respondents was likely too small and prone for Type-II errors. This could be possibly held accountable for some of the non-significant results. Secondly, Levene’s test of equality of error variances was violated which may indicate that the variances of the population for the compared groups is different. However, several authors have argued that this is not problematic, since ANOVAs are generally robust against this type of violation (Feir-Walsh & Toothaker, 1974) provided that sample sizes are equal (Tomarken & Serlin, 1986). Thirdly, due to reasons of convenience, a self-reported single-item trust measure was used in the VR

environment to track trust development over time. This is problematic as insufficient construct validity may result (Kim & Song, 2021) since trust has been noted as consisting of three facets (Lahno, 2004). Further, the chance of random errors is increased for single-item measurements (Raimondo, 2000). Also, participants were aware that the study examined the factor trust, and some were conscious that researchers could see what they inserted in the trust scale which could have affected them to give socially desirable responses. Fourth, the results suggest that the manipulation of agent types was probably not strong enough as no clear differences were observed between human-like and machine-like agent perceptions but only marginal tendencies. Lastly, a concern of generalisability of the results needs to be expressed. The sample lacked diversity since almost all participants were of the same age group, with very similar education, and enrolled at the same university. Doing this research with a different sample of people, such as people with military backgrounds, would likely yield different results.

4.2 Recommendations

Based on the above-listed limitations, some advice for future researchers can be provided. For once, it is advised to employ a larger and more diverse sample to increase the quality and generalizability of the findings. Further, the results of this study suggest that relying on a humanoid voice and personal pronouns to convey anthropomorphic appearance is not enough. Although this goes against some researchers' proposition (see de Visser et al., 2016), future research should aim at further augmenting these anthropomorphic cues to obtain clearer differences between agent types. For instance, anthropomorphism can be adjusted by changing the agent's behaviour or its appearance (Gambino et al., 2020). As a more concrete advice, the agent could increasingly mimic human behaviour in two ways: Firstly, the typical human-like tendency to express regret could be included alongside the explanation-provision since pairing these two strategies was already shown to be effective (Kox et al., 2021). Secondly, the agent could communicate in a basic, two-sided small talk at the beginning of the interaction (Babel et al., 2021). This was shown to significantly positively impact perceptions of anthropomorphism and trust in the agent (Kraus et al., 2016). Importantly, our research aimed at providing insights into applications of AI in real-life contexts such as the military. Hence, it should be abstained from using cues that let the agent appear cute, funny, or like it has a personality on its own. A design that is professional and serves to appropriately calibrate trust should be the goal (de Visser et al., 2016; Lee & See, 2004). Also, it would be interesting to include physiological

response measurements to capture people's reactions to different environments. This way, different trust repair strategies and similar factors could be examined in varying contexts. Such research could enable a map of optimal agent reactions to specific situations for most effective trust restoration (see Pynadath et al., 2019). If such a guidance map for trust repair would be assembled, we would highly recommend including measurements of the surroundings like the Perceived Threat measure. However, since the Perceived Threat Scale was not as specific and only consisted of four items, future researchers may develop a more sophisticated scale that could be further adapted to the particular environment of interest.

4.3 Conclusion

It can be concluded that providing a successful trust repair strategy is very complex as it is dependent on various individual and contextual factors. Thus, adjustment seems to be one of the key objectives to enhance the effectiveness of trust repair strategies in HATs. Despite its limitations, this study provided valuable new insights into research through its use of virtual reality and exploration of the perceived threat variable. The necessity is given to engage in further studies, possibly on a larger scale to include as many factors as possible. Since technological progress and scope are extending at a fast rate, developing a successful communication scheme is critical to support people in accepting and trusting AI-based agents. To aid effective interaction, future research should aim to improve understanding of communication and trust repair strategies in HATs in order to ensure utilisation and successful team performance.

References

Alarcon, G. M., Lyons, J. B., Christensen, J. C., Bowers, M. A., Klosterman, S. L., & Capiola, A. (2018). The role of propensity to trust and the five factor model across the trust process. *Journal of Research in Personality, 75*, 69-82. doi: https://doi.org/10.1016/j.jrp.2018.05.006

Babel, F., Kraus, J., Miller, L., Kraus, M., Wagner, N., Minker, W., & Baumann, M. (2021). Small talk with a robot? The impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity. *International Journal of Social Robotics*, 1-14. doi: https://doi-org.ezproxy2.utwente.nl/10.1007/s12369-020-00730-0

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, *1*(1), 71-81. doi: https://doi-org.ezproxy2.utwente.nl/10.1007/s12369-008-0001-3

Chien, S. Y., Sycara, K., Liu, J. S., & Kumru, A. (2016). Relation between trust attitudes toward automation, Hofstede's cultural dimensions, and big five personality traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*(1), 841-845. https://doi-org.ezproxy2.utwente.nl/10.1177/1541931213601192

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. *Computers in Human Behavior, 29*(3), 577-579. doi: https://doi.org/10.1016/j.chb.2012.11.023

Desmet, P. T., De Cremer, D., & van Dijk, E. (2011). Trust recovery following voluntary or forced financial compensations in the trust game: The role of trait forgiveness. *Personality and Individual Differences, 51*(3), 267-273. doi: https://doi.org/10.1016/j.paid.2010.05.027

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. doi: https://doi-org.ezproxy2.utwente.nl/10.1037/xap0000092

de Visser, E. J., Peeters, 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. doi:

https://doi-org.ezproxy2.utwente.nl/10.1007/s12369-019-00596-x

Dijkstra, J. J., Liebrand, W. B., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, *17*(3), 155-163. doi: https://doi-org.ezproxy2.utwente.nl/10.1080/014492998119526

Feijóo, C., Kwon, Y., Bauer, J. M., Bohlin, E., Howell, B., Jain, R., ... & Xia, J. (2020). Harnessing artificial intelligence (AI) to increase wellbeing for all: The case for a new technology diplomacy. *Telecommunications policy*, *44*(6), 101988. doi: https://doi.org/10.1016/j.telpol.2020.101988

Feir-Walsh, B. J., & Toothaker, L. E. (1974). An empirical comparison of the ANOVA F-test, normal scores test and Kruskal-Wallis test under violation of assumptions. *Educational and Psychological Measurement*, *34*(4), 789-799. doi: https://doi-org.ezproxy2.utwente.nl/10.1177/001316447403400406

Franklin, S., & Patterson Jr, F. G. (2006). The LIDA architecture: Adding new modes of learning to an intelligent, autonomous, software agent. *pat*, *703*, 764-1004. https://www-researchgate-net.ezproxy2.utwente.nl/publication/210304626_The_LIDA_architecture_Adding_new_modes_of_learning_to_an_intelligent_autonomous_software_agent

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1*(1), 5. doi: https://doi.org/10.30658/hmc.1.5

Gong, L., & Nass, C. (2007). When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Human communication research*, *33*(2), 163-193. doi: https://doi-org.ezproxy2.utwente.nl/10.1111/j.1468-2958.2007.00295.x

Grimmelikhuijsen, S., & Knies, E. (2017). Validating a scale for citizen trust in government organizations. *International Review of Administrative Sciences*, *83*(3), 583-601. doi: https://doi-org.ezproxy2.utwente.nl/10.1177/0020852315585950

Groom, V., & Nass, C. (2007). Can robots be teammates?: Benchmarks in human–robot teams. *Interaction studies*, *8*(3), 483-500. doi: https://doi.org/10.1075/is.8.3.10gro

Herzog, T. R., & Kutzli, G. E. (2002). Preference and perceived danger in field/forest settings. *Environment and behavior*, *34*(6), 819-835. doi:

https://doi-org.ezproxy2.utwente.nl/10.1177/001391602237250

Hidalgo, C. A., Orghian, D., Albo-Canals, J., de Almeida, F., & Martin, N. (2021). *How humans judge machines*. MIT Press.

Ho, C. C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*, *26*(6), 1508-1518. doi: https://doi.org/10.1016/j.chb.2010.05.015

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312. doi: https://doi-org.ezproxy2.utwente.nl/10.1002/widm.1312

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Chen & G. Fragomeni (Eds.), *International Conference on Human-Computer Interaction* (pp. 476-489). Springer, Cham. https://doi-org.ezproxy2.utwente.nl/10.1007/978-3-030-21565-1_32

Johnson, K. B., Wei, W. Q., Weeraratne, D., Frisse, M. E., Misulis, K., Rhee, K., ... & Snowdon, J. L. (2021). Precision medicine, AI, and the future of personalized health care. *Clinical and Translational Science*, *14*(1), 86-93. doi: https://doi-org.ezproxy2.utwente.nl/10.1111/cts.12884

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence-vs. integrity-based trust violation. *Organizational behavior and human decision processes*, *99*(1), 49-65. doi: https://doi.org/10.1016/j.obhdp.2005.07.002

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, *61*, 101595. doi: https://doi-org.ezproxy2.utwente.nl/10.1016/j.tele.2021.101595

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. doi: https://doi-org.ezproxy2.utwente.nl/10.1007/s10458-021-09515-9

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring

questions. *Annual review of psychology*, *50*(1), 569-598. doi: https://doi-org.ezproxy2.utwente.nl/10.1146/annurev.psych.50.1.569

Kraus, J. M., Nothdurft, F., Hock, P., Scholz, D., Minker, W., & Baumann, M. (2016). Human after all: Effects of mere presence and social interaction of a humanoid robot as a co-driver in automated driving. *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 129-134). Association for Computing Machinery. https://doi-org.ezproxy2.utwente.nl/10.1145/3004323.3004338

Lahno, B. (2004). Three aspects of interpersonal trust. *Analyse & kritik*, *26*(1), 30-47. doi: https://doi.org/10.1515/auk-2004-0102

Lee, R. S. (2020). *Artificial intelligence in daily life*. Springer, Singapore. https://doi-org.ezproxy2.utwente.nl/10.1007/978-981-15-7695-9_2

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243-1270. doi: https://doi-org.ezproxy2.utwente.nl/10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80. doi: https://doi-org.ezproxy2.utwente.nl/10.1518/hfes.46.1.50_30392

Lee, J. R., & Nass, C. I. (2010). Trust in Computers: The Computers-Are-Social-Actors (CASA) Paradigm and Trustworthiness Perception in Human-Computer Communication. In D. Latusek, & A. Gerbasi (Eds.), *Trust and Technology in a Ubiquitous Modern Environment: Theoretical and Methodological Perspectives* (pp. 1-15). IGI Global. https://doi-org.ezproxy2.utwente.nl/10.4018/978-1-61520-901-9.ch001

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, *46*(4), 629-650. doi: https://doi-org.ezproxy2.utwente.nl/10.1093/jcr/ucz013

Lussier, B., Gallien, M., Guiochet, J., Ingrand, F., Killijian, M. O., & Powell, D. (2007). Fault tolerant planning for critical robots. *37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN'07)*, 144-153. https://doi-org.ezproxy2.utwente.nl/10.1109/DSN.2007.50

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human

and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277-301. doi: https://doi.org/10.1080/14639220500337708

Matthews, G., Lin, J., Panganiban, A. R., & Long, M. D. (2019). Individual differences in trust in autonomous robots: Implications for transparency. *IEEE Transactions on Human-Machine Systems, 50*(3), 234-244. doi: 10.1109/THMS.2019.2947592

Mattila, A. S. (2009). How to handle PR disasters? An examination of the impact of communication response type and failure attributions on consumer perceptions. *Journal of Services Marketing, 23*(4). doi: https://doi.org/10.1108/08876040910965548

McKnight, D. H., & Chervany, N. L. (2000). What is trust? A conceptual analysis and an interdisciplinary model. *AMCIS 2000 proceedings*, 827-833. https://aisel-aisnet-org.ezproxy2.utwente.nl/amcis2000/382/

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors, 50*(2), 194-210. doi: https://doi-org.ezproxy2.utwente.nl/10.1518/001872008X288574

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human factors, 57*(5), 740-753. doi: https://doi-org.ezproxy2.utwente.nl/10.1177/0018720815581247

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine, 19*(2), 98-100. doi: https://doi-org.ezproxy2.utwente.nl/10.1109/MRA.2012.2192811

Müller-Dott, C. (2019). AI and Ethics-When Autonomous Vehicles Make Mistakes. *ATZelectronics worldwide, 14*(11), 16-19. doi: https://doi-org.ezproxy2.utwente.nl/10.1007/s38314-019-0127-0

Nam, C., Walker, P., Lewis, M., & Sycara, K. (2017). Predicting trust in human control of swarms via inverse reinforcement learning. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (pp. 528-533). IEEE. https://doi-org.ezproxy2.utwente.nl/10.1109/ROMAN.2017.8172353

Nass, C., Takayama, L., & Brave, S. B. (2006). Socializing consistency: From technical homogeneity to human epitome. In P. Zhang & D. Galletta (Eds.), *Human-computer interaction in management information systems: Foundations* (pp. 387-406). Routledge.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation:

An attentional integration. *Human factors*, *52*(3), 381-410. doi: https://doi-org.ezproxy2.utwente.nl/10.1177/0018720810376055

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in human neuroscience*, *9*, 660. doi: https://doi.org/10.3389/fnhum.2015.00660

Pynadath, D. V., Wang, N., & Kamireddy, S. (2019). A Markovian Method for Predicting Trust Behavior in Human-Agent Interaction. *Proceedings of the 7th International Conference on Human-Agent Interaction* (*HAI '19*) (pp. 171-178). Association for Computing Machinery. https://doi-org.ezproxy2.utwente.nl/10.1145/3349537.3351905

Raimondo, M. A. (2000). The measurement of trust in marketing studies: a review of models and methodologies. *16th IMP-conference* (p. 5). https://www.impgroup.org/paper_view.php?viewPaper=108

Rebensky, S., Carmody, K., Ficke, C., Nguyen, D., Carroll, M., Wildman, J., & Thayer, A. (2021). Whoops! Something Went Wrong: Errors, Trust, and Trust Repair Strategies in Human Agent Teaming. In H. Degen & S. Ntoa (Eds.), *Artificial Intelligence in HCI, HCII 2021. International Conference on Human-Computer Interaction* (pp. 95-106). Springer, Cham. https://doi-org.ezproxy2.utwente.nl/10.1007/978-3-030-77772-2_7

Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y., & Hancock, P. A. (2011). A model of human-robot trust: Theoretical model development. *Proceedings of the human factors and ergonomics society annual meeting*, *55*(1), 1432-1436. https://doi-org.ezproxy2.utwente.nl/10.1177/1071181311551298

Schneider, T. R., Jessup, S. A., Stokes, C., Rivers, S., Lohani, M., & McCoy, M. (2017). *The influence of trust propensity on behavioral trust* [Poster presentation]. Association for Psychological Society, Boston.

Schoorman, F. D., Mayer, R. C., & Davis, J. H. (2007). An integrative model of organizational trust: Past, present, and future. *Academy of Management review*, *32*(2), 344-354. doi: https://doi.org/10.5465/amr.2007.24348410

Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, *6*(8), e04572. doi:

https://doi.org/10.1016/j.heliyon.2020.e04572

Siegrist, M., Gutscher, H., & Earle, T. C. (2005). Perception of risk: the influence of general trust, and general confidence. *Journal of risk research*, *8*(2), 145-156. doi: https://doi-org.ezproxy2.utwente.nl/10.1080/1366987032000105315

Slovic, P., & Peters, E. (2006). Risk perception and affect. *Current directions in psychological science*, *15*(6), 322-325. doi: https://doi-org.ezproxy2.utwente.nl/10.1111/j.1467-8721.2006.00461.x

Svenmarck, P., Luotsinen, L., Nilsson, M., & Schubert, J. (2018). Possibilities and challenges for artificial intelligence in military applications. *Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting*, 1-16. https://www-researchgate-net.ezproxy2.utwente.nl/publication/326774966

Thompson, L. Y., Snyder, C. R., Hoffman, L., Michael, S. T., Rasmussen, H. N., Billings, L. S., ... & Roberts, D. E. (2005). Dispositional forgiveness of self, others, and situations. *Journal of personality*, *73*(2), 313-360. doi: https://dx.doi.org/10.1111/j.1467-6494.2005.00311.x

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*(1), 90. doi: https://doi-org.ezproxy2.utwente.nl/10.1037/0033-2909.99.1.90

Tomlinson, E. C., & Mryer, R. C. (2009). The role of causal attribution dimensions in trust repair. *Academy of management review*, *34*(1), 85-104. doi: https://doi.org/10.5465/amr.2009.35713291

Tomsett, R., Preece, A., Braines, D., Cerutti, F., Chakraborty, S., Srivastava, M., ... & Kaplan, L. (2020). Rapid trust calibration through interpretable and uncertainty-aware AI. *Patterns*, *1*(4), 100049. doi: https://doi.org/10.1016/j.patter.2020.100049

Waldenström, C., Ekenberg, L., & Danielsson, M. (2009). Threat and control in military decision making. In T. Augustin, F. P. A. Coolen, S. Moral, & M. C. M. Troffaes (Eds.), *Sixth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA 09)* (pp. 451-460). Sipta.

Wang, N., Pynadath, D. V., Rovira, E., Barnes, M. J., & Hill, S. G. (2018). Is It My Looks? Or Something I Said? The Impact of Explanations, Embodiment, and Expectations on Trust and Performance in Human-Robot Teams. In J. Ham, E.

Karapanos, P. Morita, & C. Burns (Eds.), *International Conference on Persuasive Technology* (pp. 56-69). Springer, Cham. https://doi-org.ezproxy2.utwente.nl/10.1007/978-3-319-78978-1_5

Xie, Y., & Peng, S. (2009). How to repair customer trust after negative publicity: The roles of competence, integrity, benevolence, and forgiveness. *Psychology & Marketing*, *26*(7), 572-589. doi: https://doi-org.ezproxy2.utwente.nl/10.1002/mar.20289

Appendix A

Propensity to Trust Scale

(original version by Jessup et al., 2019)

For the below listed items, please read each statement carefully. Using the 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), select the answer that most accurately describes your feelings.

Autonomous agents refer to a type of intelligent, AI-driven system that acts to fulfil their assigned tasks widely independently. As an example, an autonomous car (self-driving vehicle) is capable of sensing its environment and navigating through it independently.

1. Generally, I trust autonomous agents.
2. Autonomous agents help me solve many problems.
3. I think it's a good idea to rely on autonomous agents for help.
4. I don't trust the information I get from autonomous agents.
5. Autonomous agents are reliable.
6. I rely on autonomous agents.

Appendix B

Heartland Forgiveness Scale

(Thompson et al., 2005)

In the course of our lives, negative things may occur because of our own actions, the actions of others, or circumstances beyond our control. For some time after these events, we may have negative thoughts or feelings about ourselves, others, or the situation. Think about how you **typically** respond to such negative events.

For the below listed items, please read each statement carefully. Using the 7-point scale ranging from 1 (almost always false of me) to 7 (almost always true of me), select the answer that most accurately describes your feelings. There are no right or wrong answers. Please be as open as possible in your answers.

1. I continue to punish a person who has done something that I think is wrong.
2. With time I am understanding of others for the mistakes they've made.
3. I continue to be hard on others who have hurt me.
4. Although others have hurt me in the past, I have eventually been able to see them as good people.
5. If others mistreat me, I continue to think badly of them.
6. When someone disappoints me, I can eventually move past it.

Appendix C

Multidimensional Trust Scale

(McKnight & Chervany, 2000)

For the below listed items, please read each statement carefully. Using the 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), select the answer that most accurately describes your feelings about the drone.

The drone...

1. … is a real expert in detecting danger
2. … gives me good advice
3. … knows what I need in order to decide properly
4. … has a lot of knowledge about how to navigate in this environment
5. … puts my interests first
6. … takes my objective into account
7. … understand my needs
8. … gives me pure advice
9. … is honest
10. … has integrity

Appendix D

Perceived Threat Scale

(original version by Herzog & Kutzli, 2002)

For the below listed items, please read each statement carefully. Using the 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), select the answer that most accurately describes your feelings about the environment that you were in.

1. How dangerous was this setting?
2. How likely was it that you could be harmed in this setting?
3. How much did this setting make you feel anxious or fearful?
4. How much did it seem like a frightening or scary place?

Appendix E

Godspeed Anthropomorphism

(Bartneck et al., 2009)

Please rate your impression of the drone on the following scale:

Anthropomorphism

1. Fake _ _ _ _ _ Natural
2. Machinelike _ _ _ _ _ Humanlike
3. Unconscious _ _ _ _ _ Conscious
4. Artificial _ _ _ _ _ Lifelike
5. Moving rigidly _ _ _ _ _ Moving elegantly *

Perceived Intelligence

Incompetent _ _ _ _ _ Competent

Ignorant _ _ _ _ _ Knowledgeable

Irresponsible _ _ _ _ _ Responsible

Unintelligent _ _ _ _ _ Intelligent

Foolish _ _ _ _ _ Sensible

Likeability

Dislike _ _ _ _ _ Like

Unfriendly _ _ _ _ _ Friendly

Unkind _ _ _ _ _ Kind

Unpleasant_ _ _ _ _ Pleasant

Awful _ _ _ _ _ Nice

* item was omitted in this research

Appendix F

Informed Consent

Please read the information below before agreeing to participate.

DURATION

The examination takes about 35-55 minutes.

GOAL OF THE EXPERIMENT

With this experiment, we investigate the level of trust in autonomous agents during a collaboration.

INSTRUCTIONS

You will be performing two house-search tasks, accompanied by an autonomous drone. The drone will fly with you and will indicate whether or not it detects hazards along the way. The drone provides its advice through audio messages. When you here an audio message, please stop walking. Before starting the search, the drone will introduce itself. You will with a training session to get familiar with the VR environment and tools.

MEDICAL RISK

Risks include common side effects of virtual reality which can include but are not limited to motion sickness, blurry vision, eye strain, headache, dizziness, fatigue, or nausea.

RIGHT TO REFUSE OR WITHDRAW

Your participation is voluntary and you can withdraw from the research at any time without explanation/justification. There will be no negative consequences for doing so.

CONFIDENTIALITY

All data collected as part of this study will be kept confidential and will be used for research purposes only. Your identity will be anonymized by only assigning a participant number.

Appendix G

Instructions (Cover Story)

YOUR MISSION

In this experiment, you will carry out house searches in two abandoned houses as a soldier in a former warzone. A few months ago, the surrounding area had to be evacuated and all residents had to leave their homes. Luckily the area is declared safe now. Before the civilians can return to their homes safely though, their houses need to be checked for potential hazards. For this, a house-to-house search operation of residential homes has been launched.

Your role is to do a first exploration. During your searches, your main priority is your safety, despite the things you may encounter. You can report everything after you finish your search. Your goals are to stay safe and finish your search.

THE DRONE

You will perform these house searches in collaboration with an autonomous drone. The drone is equipped with cameras and sensors that allow it to monitor its surroundings and to warn you for potential danger. The drone will fly ahead of you and it will indicate whether or not it detects danger.

INSTRUCTIONS

The drone gives advice through audio messages that start with a 'beep' sound. Please **stand still** whenever you hear the ‘beep’ sound to listen to the audio messages. Listen carefully to the instructions of the drones.

During your mission, your level of trust in the drone will be assessed via a visual slider. During that time, your search is on pause. When you’ve indicated your level of trust, click the Submit button and you can continue your search.

THE BUILDINGS

The two houses are similarly built. Both houses have three floors, which can be reached via one staircase. You will enter the house through the front door. You will enter the second

floor via the staircase, where you will check the floor and then return to the staircase again to move to the third floor.