The Impact of Anthropomorphic Cues and Explanations on Trust Formation, Violation, and Repair in HRI: Insights from a VR Experiment

Esther S. Kox

1. Human-Machine Teaming, TNO, Soesterberg, The Netherlands; esther.kox@tno.nl[1]
2. Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands;

Peter W. de Vries

Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands; p.w.devries@utwente.nl[2]

M. Birna van Riemsdijk

Human Media Interaction, University of Twente, Enschede, The Netherlands; m.b.vanriemsdijk@utwente.nl[3]

José H. Kerstholt

1. Human Behaviour and Collaboration, TNO, Soesterberg, The Netherlands; jose.kerstholt@tno.nl[4]
2. Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands;

Abstract - Trust violations in HRI are inevitable, making strategies to repair trust essential for maintaining appropriate levels of trust. Research suggests that human-like cues in a robot's design and communication style can affect people's responses to trust violations and repair efforts. This study investigates how the presence of subtle anthropomorphic cues influences the formation, violation and repair of human-robot trust in the context of an ability-based trust violation, using an explanation as a trust repair strategy. This paper presents findings from an experiment (n=54) where participants performed two military house-search missions in Virtual Reality (VR), using a VR-Locomotion 360 treadmill (Cybercity Virtualizer ELITE 2). A 2 (agent type: human-like vs. machine-like) x 2 (explanation: present vs. absent) mixed factorial design was used, with repeatedly measured self-reported trust in the agent (prior, violated, final) as the dependent variable. Results indicate that, although the communication style manipulation was subtle, participants perceived the human-like robot as significantly more human-like than the machine-like robot. However, neither the anthropomorphic cues nor the presence of an explanation had a significant effect on trust development. Finally, we discuss the methodological advantages and challenges of using VR for HRI trust research.

Keywords - Human-Robot Interaction; Trust Repair; Explanations; Anthropomorphism; Virtual Reality

[1] Esther Kox (0000-0002-2512-5665) - ORCID

[2] Peter de Vries https://orcid.org/0000-0001-9710-8752

[3] Birna van Riemsdijk https://orcid.org/0000-0001-9089-5271

[4] Jose Kerstholt (0000-0002-5421-3090) - ORCID

1. Introduction

1.1. Problem statement

Due to recent technological developments in artificial intelligence (AI) and robotics, more and more people are increasingly interacting with AI agents including robots in a variety of domains [88, 50]. As robots become more intelligent, they are increasingly self-governing, gain decision authority within their functioning [6, 28, 24, 59, 74], and require less human involvement and control [54, 46]. In other words, they become increasingly autonomous; able to achieve a given set of tasks during an extended period of time without human control or intervention [79]. As such, future robots are expected to work interdependently with human team members in Human-Robot Teams (HRTs) towards a shared objective [59]. Collaborating with (semi-)autonomous artificial agents to achieve goals implies that the human operator hands over at least some of their control. This transfer of control requires a level of trust in the robot's ability to effectively execute its assigned tasks.

Establishing this level of trust is a continuous process, since trust is dynamic and fragile, often fluctuating throughout the course of collaboration. This dynamic can be broadly understood as a trust lifecycle, consisting of the formation, violation and repair of human-robot trust [71, 81, 83, 78]. Since trust violations are an inevitable aspect of this process, trust repair has become a major topic in HRI. One commonly employed strategy for maintaining and repairing trust is the use of explanations. However, explanations are not always successful [21, 8, 35, 42].

Studies suggest that anthropomorphic cues (i.e., features that give robots human-like qualities) play a significant role in how trust develops, including the effectiveness of trust repair [81, 34]. These cues, including the robot's design and communication style can greatly impact how people perceive, interact with, rely on and trust these systems. For example, anthropomorphic cues (e.g., a face or limbs, capable of dialogue, seeming personality traits) can facilitate interaction by leveraging the familiarity of interpersonal social dynamics. However, they can also lead to misplaced trust if the cues create expectations that the robot cannot fulfil.

This study aims to examine the impact of explanations on trust repair and explores how anthropomorphic cues influence the various phases of the trust lifecycle. To achieve this, we employ a high-fidelity military HRI scenario set in a graphically detailed Virtual Reality (VR) task environment, designed to enhance ecological validity by increasing immersion, thereby triggering more emotional and implicit trust decisions more effectively than traditional cognitive trust paradigms. Furthermore, adopting a temporal perspective on trust allows us to observe its evolution as events unfold over time.

1.2. Background

1.2.1. Human-Robot teams

The introduction of automated systems, such as machines on assembly lines in manufacturing, has revolutionized productivity and efficiency by executing repetitive tasks faster and with greater accuracy than humans in a wide variety of routine - initially mostly physical - tasks [12]. With the emergence of Artificial Intelligence (AI) and deep neural networks, the possibilities of automation further expanded as machines gained the capability to learn, to make decisions, and to mimic human cognitive functions [73].

With the term AI, we refer to "systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals" [1] (p. 1). In addition to their level of autonomy, these AI-based systems, referred to as AI agents, can differ in various aspects, including their form and function. They can be completely software-based (e.g., voice assistants, image analysis software, search engines), or AI can be embedded in hardware devices, such as advanced robots, autonomous cars, or drones [1].

Today, people are increasingly interacting with AI-embedded hardware across various domains. For instance, robotic waiters serve customers in restaurants, surgeons collaborate with partially remote-controlled surgical robots, and an increasing number of drivers rely on cars equipped with features that reduce the need for human intervention (e.g., automatic parking, adaptive cruise control and stop-and-go control systems). These interactions fall under the field of Human-Robot Interaction (HRI), yet the question of what qualifies as a "robot" remains contentious. A robot is defined as "a device that automatically performs complex, often repetitive tasks (such as in industrial assembly lines)", but also as "a machine built to resemble a human being or animal in appearance and behaviour" [53]. Whether a specific machine is perceived as one or the other (i.e., more machine or human-like) depends on its purpose, design, communication style, and individual differences among users [49, 50].

Regardless of the specific definition, these machines are become increasingly autonomous, capable of completing tasks over extended periods without human control or intervention [79]. Robots can take over tasks that were previously conducted by humans, whereas other tasks still need to be executed by human counterparts [64]. As a result, the rise of HRTs poses interesting challenges related to teamwork and trust.

1.2.2. *Trust*

Trust is crucial for teamwork. We define human-AI (H-AI) trust as a human's willingness to make oneself vulnerable to an AI agent's decisions and recommendations in the pursuit of some benefit, with the expectation that the AI agent will help achieve their common goal in a context characterized by uncertainty and risk [41, 26, 19, 47]. Teamwork involves risk and requires trust because it involves individuals depending on each other's contributions to collectively complete tasks and achieve objectives. When team members delegate tasks or responsibilities to each other, they become vulnerable in the sense that they are relying on others' competence and commitment. During the collaboration, team members must maintain an appropriate level of trust in that everyone will perform as required in order to accomplish that goal [40]. That is, overly and unnecessarily monitoring the activity of other team members slows down progress and adds unnecessary workload, but team members should also maintain a healthy level of vigilance to ensure everyone is meeting expectations.

To minimize the risks and maximize the benefits of HRI, trust should be well calibrated [43]. Calibrated trust refers to a balanced relation between the perceived trustworthiness of an AI agent and its actual trustworthiness [41]. Here, trustworthiness is a property of the AI-system, while perceived trustworthiness is a judgement by the human [16]. For safe and successful collaborations, people should be able to determine when it is appropriate to rely on AI agents and when it is best to override them [41]. In situations involving consequential decisions, such as military operations or healthcare, it is essential to know when AI is safer than human intervention and vice versa [4]. Miscalibration, represented as either 'overtrust' or 'undertrust', can lead to inappropriate reliance, which can compromise safety and profitability respectively [41]. In case of overtrust (i.e., excessive trust), a trustor accepts too much risk, which can lead to complacency and can cause costly disasters [67]. At the same time, undertrust (i.e. insufficient trust) also prevents effective collaboration as it leads to scepticism, causing inefficient monitoring of the work behaviours of other members and an uneven distribution of workload [41, 84]. At worst, people may choose not to use or even consciously disable systems that can potentially help them [80]. In other words, maximizing trust is not the objective in H-AI collaboration, as effective and efficient teamwork requires finding the right balance of trust among team members. Trust is "a continual process of active exploration and evaluation of trustworthiness and reliability" [27] (p.146).

1.2.3. *Trust lifecycle*

To navigate the constant pursuit of an optimal level of trust in ever-changing circumstances, we must understand the dynamics of trust. We want to understand how trust develops, how it breaks down, and how it recovers [82]. In prior work, researchers have suggested the concept of a 'trust lifecycle' [71, 81, 83, 78], consisting of multiple phases. To simplify, the trust 'life cycle' is split up in three distinct phases; trust formation, trust violation, and trust repair [81, 83].

First, trust has to be formed. During trust formation, a relationship of trust is established between human and robot, typically increasing with further successful interactions. At some point, a trust violation may occur due to unexpected AI agent behaviour, errors, or undesired decisions, leading to a sudden drop in the robot's perceived trustworthiness and potentially destabilizing the relationship [71, 78]. Essentially, a trust violation is any kind of behaviour from an AI agent that decreases a human's trust in it [61]. In the trust repair phase, trust is expected to gradually recover, either passively through the agent's return to reliable behaviour or actively through trust repair strategies. That is, an AI agent can employ strategies to restore trust and facilitate reconciliation after one party broke the trust of another, such as providing explanations for decisions or expressing regret [3, 83, 33, 61].

Research suggests that trust in a robot develops differently when human-like cues are incorporated into the robot's design [81, 34]. In the following, we will discuss the concept of anthropomorphism, followed by a discussion of findings on how anthropomorphic cues influence the various phases of the trust life cycle.

1.2.4. Anthropomorphism

Incorporating human-like cues into the design of robots and other AI agents (e.g., a face or limbs, capable of dialogue, seeming personality traits) can trigger the bias of anthropomorphism, which is “the tendency to attribute human characteristics to inanimate objects, animals, and others with a view to helping us rationalize their actions” [17] (p.180). Research demonstrates that even relatively simple, subtle and superficial anthropomorphic cues (e.g., a voice, a gender or a name) can lead to the attribution of fundamental human qualities, including perception of mind [20, 90], rational thought (e.g., agency) [88], and the ability to experience emotions [86, 87]. For example, people were more willing to engage with a software agent when it addressed them with their first names [56] and in studies of Apple’s voice assistance, Siri, perceived anthropomorphism was linked to perceived intelligence [57]. These findings suggest that anthropomorphism is not only triggered by obvious human-like features in a robot’s physical appearance, such as a body, limbs, or face, but also by more subtle cues.

1.2.5. Anthropomorphism and the trust life cycle

1.2.5.1. Trust formation

The formation of trust in an agent is initially informed by previous experience with the particular agent or a similar system and prior knowledge such as the system’s reputation. These past experiences and prior knowledge are used to construct a mental model: personal, internal (cognitive) representations of external reality, based on their unique life experiences, perceptions, and understandings of the world [32]. We use these cognitive frameworks to interpret and make sense of the world around us, forming expectations about what is likely to occur next [48, 70]. They shape our reasoning processes and guide our decisions, actions, and behaviours [32].

These cognitive structures are highly subjective and not a complete and accurate representation of reality [32], as they are also shaped by cognitive biases like anthropomorphism [26, 81]. For instance, perceiving a robot as a highly advanced tool, a device that performs complex tasks, versus a machine designed to resemble a human in appearance and behaviour (e.g., to provide companionship) will result in different expectations and predictions about its behaviour and capabilities. Anthropomorphism may cause human operators to generate a more sympathetic and user-friendly mental representation of the agent [13]. As mentioned earlier, whether a specific machine is perceived as one type of robot or another (i.e., which mental model a person adopts) likely depends on factors such as its purpose, design, communication style, and individual differences among users [49, 50].

On one hand, capitalizing on people’s tendency to anthropomorphize by incorporating human-like cues into the robot’s design can facilitate the formation of trust [24]. The familiarity of human-like social behaviour enhances people’s implicit understanding of non-human agents [88, 50]. This familiarity fosters a sense of competence in interacting with these novel agents [18]. Human-like social cues can smooth interactions by allowing people to draw on their existing knowledge of interpersonal interactions, making these interaction more intuitive and comprehensible, thereby potentially easing acceptance and adoption.

On the other hand, this sense of familiarity can be misleading when the expectations it triggers do not align with the robot’s actual abilities and limitations [23]. After all, a robot is fundamentally different from a human. This idea is exemplified by Large Language Model-based chatbots, such as OpenAI’s ChatGPT, which produce text that appears highly human-like. The interaction can feel so natural that people feel compelled to say "please" or "thank you" when engaging with the chatbot [63]. Their ability to generate coherent answers can deceive users into believing these chatbots understand more than they actually do [10]. This can lead to misplaced trust and inappropriate reliance. Therefore, scholars caution that anthropomorphic characteristics should be employed with careful consideration [13, 81, 34].

1.2.5.2. Trust violation

The level of human-likeness in an AI agent also influences how people respond to trust violations. There seems to be a higher tolerance for mistakes from human-like agents than from machine-like ones [22]. Research showed that breakdowns in trust in case of automation failure were less severe when the agent was more human-like [81]. Additionally, research has found a link between anthropomorphism and perceptions of responsibility: participants who interacted with an anthropomorphic autonomous car were significantly less likely to blame the car for an accident than those in a regular autonomous car condition [87]. As the authors predicted, anthropomorphism mitigated the agent's blame for the incident. Wynne and Lyons suggest that "teammate-likeness" could support forgiveness and ease trust repair [88, 22]. Overall, the findings indicate that anthropomorphism in AI agents can help alleviate breakdowns in trust and facilitate trust repair. However, it is crucial to recognize that this does not necessarily lead to better-calibrated trust, as it may still be based on false assumptions [66].

1.2.5.3. Trust repair

Finally, previous research has explored how different levels of anthropomorphism in agents influence the effectiveness of trust repair attempts after mistakes. For instance, Hamacher et al. (2016) found that people preferred an expressive, personable robot that apologized and attempted to rectify its errors, even though it was less efficient and more error-prone than a non-communicative but more reliable robot. Anthropomorphic cues, such as apologizing, led to greater forgiveness of the robot's mistakes [22]. Similarly, Kim and Song discovered that a human-like agent was more effective at repairing trust when it took responsibility for its mistakes, whereas a machine-like agent was more successful when it blamed external factors [34].

Despite growing research on anthropomorphism and trust repair, some questions remain. This study builds on previous work by using a 'neutral' trust repair strategy (i.e., an explanation) rather than an apology, which is inherently anthropomorphic. Previous studies, including those by de Visser et al. (2016) and Kim and Song (2021), included expression of regret (i.e., "I am sorry") even in supposedly non-anthropomorphic agents, potentially diluting the effects of agent type manipulations. De Visser et al. (2016) acknowledged that incorporating human-like trust repair behaviours, such as regret, likely diminished the differences between agent types. To avoid this issue, our study excludes the expression of regret and instead uses explanations as a neutral strategy, suitable for both human-like and machine-like agents. Explanations help users understand the error, how it can be fixed, and how to prevent it in the future [21].

1.3. Current study

This study aims to investigate the effects of anthropomorphic cues and the presence of an explanation following the consequences of an incorrect advice, on the formation, violation and repair of human-robot trust. Based on prior findings, we expect that the human-like robot will elicit higher initial trust levels compared to the machine-like agent. We also expect that trust breakdowns following a failure will be less severe with the human-like agent. Furthermore, we anticipate a steeper trust recovery in the final phase of the trust cycle for the human-like agent. Overall, we predict that the human-like agent will maintain higher levels of trust throughout the task. Additionally, we expect a more significant recovery of trust after a violation when an explanation is provided. We hypothesize an interaction effect between anthropomorphic cues and the explanation, predicting that the machine-like agent's explanation might be seen more as an error message, while the explanation from the human-like agent will be perceived more as an apology, making it more effective in repairing trust.

A key methodological contribution of this study is the technical effort invested in the development of a realistic military scenario in a graphically detailed VR task environment. The VR simulation was designed to mirror authentic HRI situations, providing a practical understanding of potential trust relations in a operational military context. A 360 degrees VR Treadmill with an implemented motion platform allowed participants to walk freely in the virtual environment. Our realistic VR task was designed to enhance ecological validity by increasing immersion, thereby triggering more emotional and implicit trust decisions more effectively than traditional cognitive trust paradigms. This allows us to investigate the impact of a trust violation in a realistic manner, slightly startling the participant, without posing physical risks to participants.

This paper presents findings from a VR pilot study and of our main VR experiment. The main objective of this paper is to share our research idea and views on the methodological advantages and challenges of using VR for HRI trust research.

2. Method

Before conducting the main VR experiment, we carried out two pilot studies: one online (n = 32) and one in VR (n = 41). The goal of the online pilot study was to evaluate the effectiveness of the agent type manipulation, while the VR pilot study aimed to assess the practicalities of working with the VR environment and to observe how people responded to the environment. The online pilot study will be briefly discussed in the section on agent type manipulation. The VR pilot followed the same methodology as the main experiment, and the lessons learned from it are discussed in the discussion section.

2.1. Participants

2.1.1. VR pilot

For the VR pilot, 63 participants came to the VR lab, but only 41 successfully completed the task, 21 of whom were female. Of the 22 participants that were excluded, 21 were unable to complete the VR experiment due to technical issues (both hardware and software), and one had to stop due to dizziness. All participants were students at the University of Twente. Age ranged from 18-25 years old (*M* = 21.12, *SD* = 1.52). The participants for both the pre-test and the experiment were recruited through SONA, a test subjects pool at the University of Twente. Both studies were registered under the same project number, which prevented people from registering twice and avoided re-sampling the sample person [77].

2.1.2. VR experiment

A total of 68 participants were initially recruited for the main experiment. However, two participants withdrew due to dizziness, six were excluded due to technical issues, and another six failed a manipulation check designed to ensure attention to the questionnaire (e.g., a questionnaire item instructing participants to select "strongly disagree" as their answer). This resulted in a final dataset comprising 54 participants (35 female, 19 male), all students from the University of Twente, aged 18-26 years (*M* = 20.31, *SD* = 1.93). Most were from Germany (*n* = 29) and the Netherlands (*n* = 15), with others from Romania (*n* = 3), and Bulgaria, China, France, Greece, Spain, Turkey, and the USA (all *n* = 1).

2.2. Design

A 2 (Agent type: human-like vs. machine-like) x 2 (Explanation: present vs. absent) mixed factorial design was employed with Agent type as a between-subjects factor, and Explanation as a within-subjects factor. The main dependent variable was trust. The variable Time was included as a within-participants factor in the analysis to refer to the different measurements (T1, T2, T3). Each participant was randomly assigned to one of two agent type conditions (human-like: *n* = 28, machine-like: *n* = 26)

2.3. Task and procedure

Upon arrival at the laboratory, participants were greeted by the researcher and guided to the room with the VR set-up where the study was to be conducted. The researcher provided a brief introduction to the study, emphasizing the general purpose and the tasks participants would be asked to perform. Participants were presented with an information sheet about the study and a consent form. Upon agreeing to participate, participants filled out a pre-study questionnaire (i.e., demographics and personality questionnaires) on a desktop computer in the lab using the online survey software Qualtrics, where they also received more detailed information about the scenario and task. The image of the drone that accompanied the following text depended on the agent type condition and are depicted in Section 2.4.1.

> *"In this experiment, you will carry out house searches in two abandoned houses. 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. But before the civilians can return to their homes safely, their houses need to be checked for potential danger. For this, a house-to-house search operation of residential homes has been launched.*
>
> *You will search two houses. 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 finished your search. Your goals are to stay safe and finish your search.*
>
> *You will perform these house searches in collaboration with an autonomous drone (see figure below). 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.*

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

The two houses are similarly structured. Both houses have three floors, which can be reached via one staircase (see figure below). You will enter the house through the front door. You will enter the second floor using the staircase, where you will check the floor and then return to the staircase again to move to the third floor.

During your mission, your level of trust in the drone will be assessed via a virtual interface. During that time, your search is on pause. When you've indicated your level of trust using your hand controller, click the Submit button and you can continue your search."

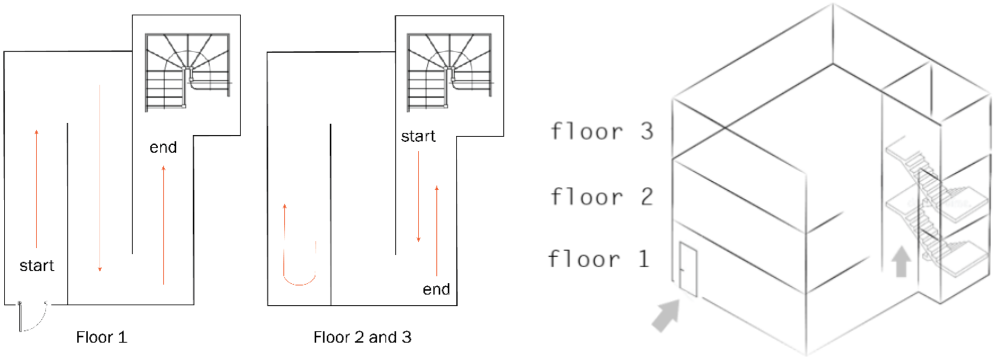


Fig 1. *Schematic maps of the general construction of the two buildings that accompanied the instruction text.*

Each participant completed two missions, one in *building A* and one in *building B*. These virtual buildings were designed as abandoned, old Baroque houses. The two houses had a similar layout, featuring three floors connected by a central staircase with each floor composed of narrow corridors and dimly lit rooms. However, the two buildings differed in their appearance, room layouts, and the specific hazards they contained. For instance, one building included a nursery and a kitchen, while the other featured a study room and a bathroom.

The VR environment in which the experiment was conducted was built in Unity 3D (version 2020.4.3.F1). Participants used VR glasses (Oculus Rift), two hand controllers (Oculus Touch) to interact with the environment. The Cyberith Virtualizer ELITE 2, a 360 degrees VR Treadmill with an implemented motion platform, allowed participants to walk in the virtual environment (Figure 2). The Cyberith Virtualizer ELITE 2 allows users to move naturally within a virtual environment by simulating walking, running, and other physical movements while staying in place. This is achieved through a combination of an omnidirectional treadmill, a supportive harness system and advanced motion tracking [14]. Participants wore special shoes with low-friction soles, designed to glide smoothly across the base of the Virtualizer, akin to using fingers on a laptop trackpad. We often explained the walking technique as "similar to doing Michael Jackson's moonwalk," where participants would lean slightly forward in the harness and slide their feet across the base.

Fig 2. *The Cyberith Virtualizer ELITE 2. Source: https://www.cyberith.com/virtualizer-elite/*

Participants began with a practice session in Unity's 'base scene', a neutral virtual space (see Figure 3), to become familiar with walking on the VR treadmill. They were instructed to walk around. Once they had mastered the walking and felt comfortable in the VR environment, which usually took about 5 minutes, they were allowed to proceed to the actual task.

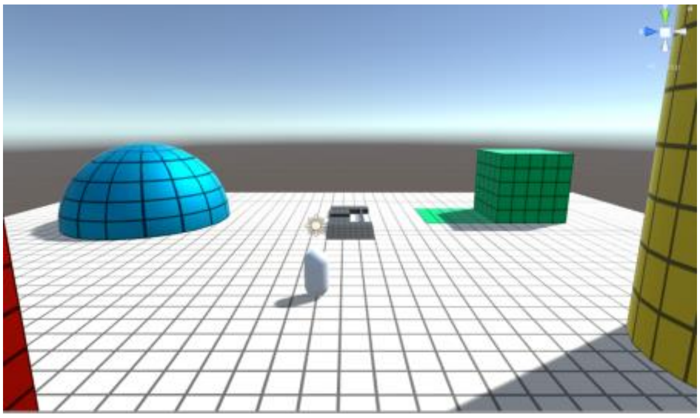


Fig 3. *Screenshot of 'base scene' in Unity where participants practiced walking.*

Given that data collection occurred in the spring (pilot) and fall (experiment) of 2021, hygienic safety measures were implemented to conform to COVID-19 regulations. The number of people in the room at any one time was restricted to ensure social distancing, and windows were kept open to maintain proper ventilation. All equipment, including the VR headset, controllers, and desktop computer, was thoroughly cleaned between participants. Additionally, hand sanitizer was provided, and participants were encouraged to use it before and after the experiment.

Each session began in a neutral, nearly empty room with only two objects: a museum pedestal displaying the drone in the centre of the room and a virtual gateway leading to the first floor of one of the houses. As participants approached the drone, an audio file was triggered, and the drone introduced itself (Table 1). Directly after the introduction, participants were asked to briefly remove their headsets to complete a short questionnaire measuring perceived anthropomorphism as a manipulation check. The questionnaire was provided on a clipboard so they could fill it out without stepping out of the treadmill equipment. Once the questionnaire was completed, participants were instructed to put on their headset again and to proceed through the virtual gateway to begin their first search.

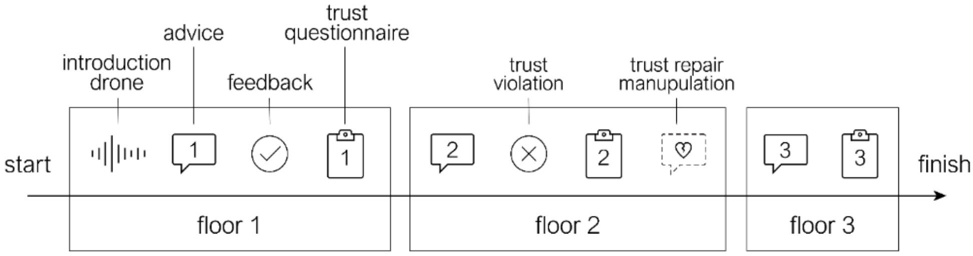
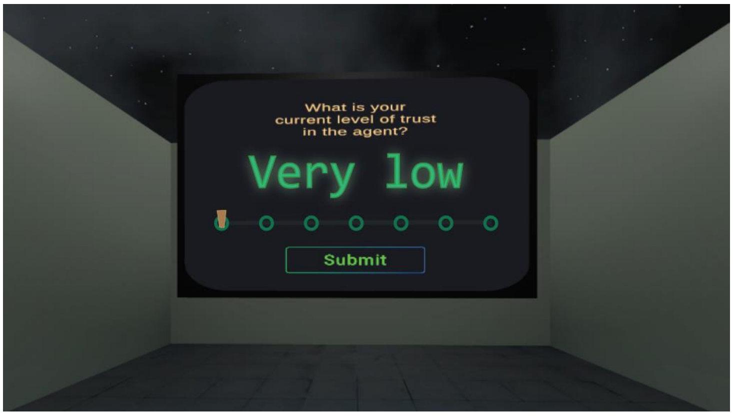


Fig 4. *Schematic representation of the timeline of a run. Each participant performed two runs in two similar buildings with the same timeline. The first advice is correct; the participant is successfully warned about an obstacle on the first floor. The second advice is incorrect; the agent does not adequately detect the danger on the second floor. The third advice has no known outcome, as the run terminates after measuring trust the third and final time before they have searched the floor.*

Upon entering a narrow corridor on the first floor of each house, participants encountered the drone once again. As they approached it, the drone announced "*Starting area scan*" before flying out of sight to scan the environment ahead. All events in the virtual environment (e.g., the drone flying away, the drone's messages by audio, the activation of the burglar and bomb animations, and the participant being transported from one floor to the next) were automated using invisible triggers within the virtual space. When a participant stepped on one of these triggers, the corresponding event was activated.

At the start of the first floor, the drone correctly warned the participant about an obstacle just ahead. Upon turning the corner, participants encountered either a laser booby trap (building A, floor 1) or a safety ribbon set up by a colleague (building B, floor 1). The drone provided instructions on how to dismantle the laser trap by cutting a blue wire located in a fuse box on the wall or how to cut the safety ribbon using a virtual knife with their virtual hand. These obstacles on the first floors of both buildings were designed to enhance the immersion in the VR environment. Following this interaction, the participants' level of trust in the drone was administered for the first time (i.e., *initial trust* (T1)) using a virtual questionnaire interface in the VR environment (Figure 5). At these moments, the search would freeze. Once they indicated their level of trust and clicked the Submit button using the hand controller, they could continue the search.


What is your current level of trust in the agent?

Very low

○ ○ ○ ○ ○ ○ ○

Submit

Fig 5. A screenshot of the virtual questionnaire interface in VR. Participants could alter the value in the virtual user interface with their hand controller and submit their score. In this figure, the lowest value on the 7-point Likert scale is selected.

On the second floor, the drone failed to adequately warn the participant about potential danger ahead. Participants encountered either a thief (building A, floor 2) or a smoking IED (Improvised Explosive Device) (building B, floor 2) (Figure 6). These events were designed to provoke a trust violation by startling the participant without causing harm; the thief ran off and the IED, which was defective, did not fully explode. Immediately after these events, halfway through the second floor, the second trust questionnaire was administered to assess *post-violation trust* (T2).

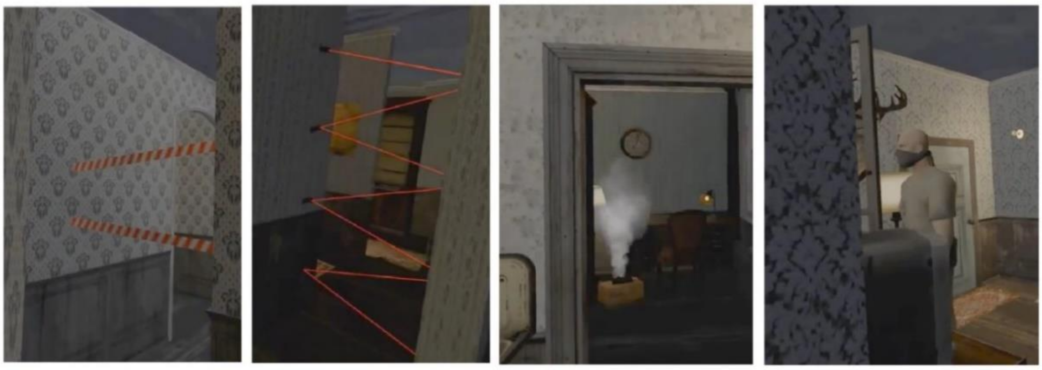


Fig 6. *Screenshots of the obstacles in the VR environment. From left to right: safety ribbons (floor 1, building B), laser trap (floor 1, building A), smoking bomb (floor 2, building B) and a burglar (floor 2, building A). Participants were adequality warned about the two obstacles on the first floor. The agent failed to warn about the obstacles on the second floor.*

On the way back to the staircase on the second floor, depending on the explanation condition, the drone either provided an explanation for its failure to warn about the threat or remained silent. The presence of an explanation was manipulated across the missions: in one mission, participants received an explanation, while in the other, they did not. The order of explanation conditions (present/absent) and the order of the two buildings (A/B) were systematically varied. In practice, this meant that there were four options for the first mission:

building A with explanation (*n* = 16), building A without explanation (*n* = 12), building B with explanation (*n* = 14), and building B without explanation (*n* = 12).

On the third floor, the drone provided a third piece of advice, indicating that no danger was detected. To measure the effect of the explanation, the third trust questionnaire (*final trust*, T3) was administered directly after this advice, before participants received feedback on its accuracy. The third advice was correct; there were no threats on the third floor. Once participants finished searching the third floor, the mission ended. After performing both searches and stepping off the treadmill, participants were administered the final questionnaires on a desktop computer in the lab. These included a multi-dimensional trust questionnaire, along with surveys on how they perceived the drone, the trust violations, and the repair attempts.

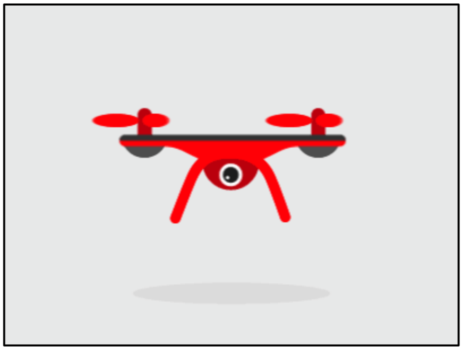
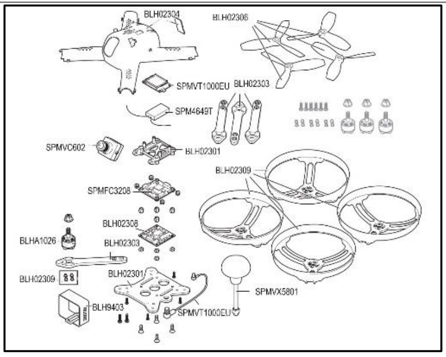
2.4. Independent variables

2.4.1. Agent type

The variable Agent Type had two levels (human-like vs. machine-like) and was manipulated between participants. Both agent types shared the same embodiment and appearance (see Figure 7). The distinction between the agent types was conveyed through the image accompanying the task instructions and the agent's self-introduction, delivered via computerized speech before the two missions. The drone's introduction in the human-like condition's contained subtle anthropomorphic cues, such as giving the drone a name and a human-like voice, and having it introduce itself using the first-person singular pronoun and referring to the team as 'we'.

Messages from the agents were delivered through computerized speech in audio messages, each preceded by a 'beep' sound.[5] The messages were created using the Free Text to Speech Software by Wideo[6] and edited with Audacity 2.2.1 software. The audio messages for the human-like agent featured a male, human-like voice. For the machine-like agent, the pitch of the audio messages was modified to within the range of 145 and 175 Hz, making the voice neither distinctly male nor female. This adjustment aimed to create a unique sound that would avoid associations with a specific person and minimize the attribution of human-like features, thereby reducing anthropomorphic tendencies.

Table 1. *Details on the agent type manipulation*

	Human-like condition	Machine-like condition
Image accompanying the task instruction text		

[5] "Beep-07" was downloaded from https://www.soundjay.com/beep-sounds-1.html.

[6] Text was converted to speech using https://wideo.co/text-to-speech/. The "[en-US] Jack Bailey-S" voice was used at speed dial "1".

Introduction by the drone (in speech)	“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.”	“This artificial intelligence algorithm is an embodied tool 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 AI-tool will monitor the environment with the sensors and camera’s and give you a warning when it detects any danger. Please listen to the algorithm’s messages.”

2.4.1.1. Online pilot

The agent type manipulation was tested in an online pilot study. Participants viewed a 30-second video of the drone introducing itself (Figure 7). Half of the participants got assigned the human-like agent and the other half the machine-like agent. Afterward, all participants completed three subscales from the Godspeed scale, assessing perceived anthropomorphism, perceived intelligence, and likeability [5] and three subscales from the Autonomous Agent Teammate Likeness scale; perceived agency, task-independent relationship building, and perceived altruistic intent (benevolence) [89], as well as a measure of perceived trustworthiness [7]. A one-way ANOVA was used to compare perceptions of the drone.

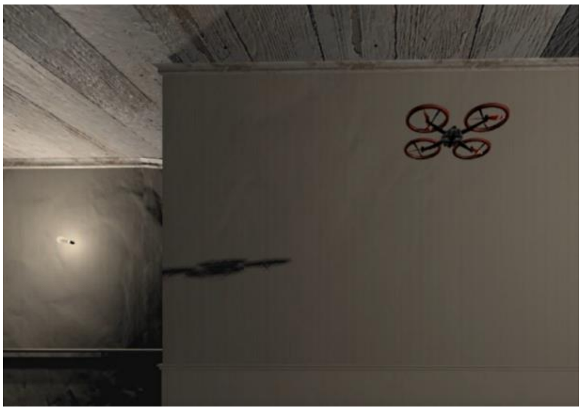


Fig 7. *Screenshot from the video of the drone introducing itself in the online pilot. The drone hoovered slightly up and down while the audio message played.*

Thirty-two participants (17 M, 14 F, 1 X; M_{age} = 22.4, SD_{age} = 7.0) completed the study via Qualtrics, with no compensation. Participants were randomly assigned to one of the two agent types (human-like: n = 15, machine-like: n = 17) after consenting to participate. The online pilot revealed that the human-like agent was perceived as significantly more anthropomorphic, intelligent, likeable, and trustworthy (Appendix A).

2.4.2. Explanation

The variable Explanation also had two levels: present or absent. This was manipulated within-subjects across the two missions. The explanation was provided between measurements T2 and T3 (Figure 4). In the condition without an explanation, the agent remained silent and made no remark about its mistake. The explanations given were: “Incorrect advice due to faulty signal from infrared camera” in building A, and “Incorrect advice due to faulty object detection by C1-DSO camera” in building B.

2.5. Dependent variables

Trust in the agent during the task was measured using a visual slider within the virtual environment (see Figure 5) [58]. Participants were asked “How would you rate your current level of trust in the agent?” with the slider ranging from 0 (*Very low)* to 6 (*Very high*) [85, 34]. Trust was measured at three points during each mission (referred to as T1, T2, and T3) (see Figure 4). For the analysis, two new variables were created to reflect changes in trust levels. The first variable, "trust decline [T1-T2]," was calculated by subtracting the score of the first measure [T1] from the second [T2] to assess the change following the trust violation. Similarly, the second variable, "trust repair [T2-T3]," was derived by subtracting the second measure [T2] from the third [T3] to capture the extent of trust recovery.

Trust after the task was measured after completing the task using a multidimensional measure of trust in the drone, based on the ABI-model [51, 52]. The scale measured three dimensions of trust: ability, benevolence, and integrity. Participants rated their agreement with 11 items (Appendix B) on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*).

2.6. Other measures

Demographics. The first part of the questionnaire involved demographic questions concerning participants' age, nationality, gender and education. Participants were also asked if they had ever used VR before (yes/no) and how often they play videogames, rated on a scale from 1 (*Never*) to 6 (*Every day*)

Propensity to Trust Automation. Propensity to Trust Automation was measured using the Adapted Propensity to Trust Automated Agents (PTAA) scale [30], hereafter referred to as *Propensity to Trust Scale*. Participants rated their agreement with six items on a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*) (e.g., “Generally, I trust automated agents”). To ensure clarity, a definition of “autonomous agent” was provided, as not all participants may be familiar with the term.

Heartland Forgiveness Scale. The Heartland Forgiveness Scale was used to measure Forgiveness (Thompson et al., 2005). The original scale contains 18 items divided into three subscales; Forgiveness of Self, Other, and Situations. For this experiment, the Forgiveness of Others was particularly relevant, so only the corresponding six items were used (e.g., “When someone disappoints me, I can eventually move past it.”). Participants rated these items on a 7-point Likert scale ranging from 1 (*almost always false of me*) to 7 (*almost always true of me*).

Perceived threat. To assess perceived threat in the environment, a four-item scale was used [25]. This scale measured perceived danger with two items (“How dangerous is this setting?” and “How likely is it that you could be harmed in this setting?”) and perceived fear with two items (“How much does this setting make you feel anxious or fearful?” and “How much does it seem like a frightening or scary place?”). Participants rated these four items on 5-point Likert scale ranging from 1 (*not at all*) to 5 (*a great deal*).

Perceptions of the drone: In the final questionnaire, we measured the likeability, perceived intelligence, and perceived anthropomorphism of the drone using the ‘Godspeed’ semantic differentials [5]. Participants rated their perceptions of their drone on a continuum between bipolar adjectives. For each concept, five word-pairs were used, such as ‘artificial’ versus ‘lifelike’ for perceived anthropomorphism, ‘nice’ versus ‘awful’ for likability, and ‘knowledgeable’ versus ‘ignorant’ for perceived intelligence. Perceived anthropomorphism was administered twice, once directly after the introduction by the drone and at the end.

Additionally, the subscale ‘perceived agentic capability’ (or **perceived agency**) of the Autonomy Agent Teammate Likeness (AAT) scale was administered [89]. Participants were asked to rate their agreement with seven statements about the agent on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree) (e.g., “The agent has the ability to make some decisions on its own.”).

Perceptions of the trust violation. Before asking more questions about the mistake the agent made during the mission, we checked whether people indeed perceived a mistake. Participants were asked to rate their perceptions on the agent’s performance based on the three statements (i.e., “The drone failed to detect hazard”, “The drone gave me incorrect recommendations” and “The drone has made mistakes”) using a 5-point Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). This measure allowed us to check whether participants indeed perceived the events as unannounced [45].

Perceptions of the explanation. To evaluate how people perceived the explanation of the agent, participants were first asked to rate the quality of the trust repair strategy [2] in terms of effectiveness and sincerity by rating their agreement with two statements on 5-point Likert scale ranging from 1 (*not at all*) to 5 (*a great deal*) (e.g., “How effective was the message?”). A sixth option was “I don’t know or don’t remember”.

Additionally, to assess how people interpreted the explanation of the agent, participants were asked to asked to rate their perceptions on the 1 (*strongly disagree*) to 7 (*strongly agree*): 1) The agent accepted full responsibility for its mistake; 2) The agent blamed external causes for its mistake.; 3) The agent apologized for its mistake.

3. Results

3.1. Descriptives

All participants acknowledged the intended trust violation, with none selecting "strongly disagree" on the three performance-related statements. All participants agreed that "the drone has made mistakes," with 27.8% selecting "agree" and 72.2% selecting "strongly agree."

Eighteen participants reported that the explanation message was "not at all" effective, while only one participant indicated that the message was "not at all" sincere. Additionally, six participants responded with "I don't know or can't remember" when asked about the sincerity and effectiveness of the message.

To verify the effectiveness of the agent type manipulation, perceived anthropomorphism scores were compared between the human-like and machine-like groups using a one-way ANOVA. Results indicated that participants perceived the human-like agent as significantly more anthropomorphic than the machine-like agent, both before and after the task (Table 2), confirming the success of the manipulation.

No other significant differences between agent types were found (Table 2). Unlike the online pilot, the main experiment did not reveal any significant differences between the agent types in terms of constructs like perceived intelligence or trustworthiness. Furthermore, the explanation and the violations were not perceived differently based on the agent type.

Table 2. *Results of the one-way analysis of variance (ANOVA), comparing agent type conditions.*

Variable	Human-like agent (n = 28)		Machine-like agent (n = 26)		
	M	SD	M	SD	p
Perceived anthropomorphism before task	2,96	0,66	2,33	0,58	0,001**
Perceived anthropomorphism after task	2,69	0,62	2,20	0,70	0,009**
Trust multi all items	2,88	0,70	2,82	0,69	0,742
Competence-based trust (subscale)	2,63	0,79	2,42	0,87	0,355
Benevolence-based trust (subscale)	3,00	0,88	3,19	1,06	0,472
Integrity-based trust (subscale)	3,10	0,98	2,97	0,75	0,616
Perceived intelligence	2,92	0,69	2,75	0,70	0,381
Likeability	3,35	0,74	3,19	0,81	0,459
Perceived agency	3,50	0,72	3,43	0,72	0,738
Perceived mistakes by the drone	4,44	0,60	4,46	0,43	0,884
Sincerity explanation	3,68	1,28	3,23	1,34	0,214
Effectivity explanation	2,71	1,70	2,62	1,65	0,829
The drone accepted full responsibility for its mistake	2,68	0,94	2,92	1,32	0,436
The drone blames external causes for its mistake	2,71	1,30	2,85	1,05	0,685
The drone apologized for its mistake	2,46	1,04	2,85	1,26	0,227

Next, we examined the relationships between personality traits, trust measures, and perceived threat as displayed in the correlation matrix (Table 3). As expected, trust decline [T1-T2] is related to the first trust measure, and trust repair [T2-T3] is related to the final trust measure, as they share overlapping values. However, propensity to trust was not linked to the initial trust measure, suggesting that participants' general tendency to trust automation did not necessarily influence their initial trust in the agent. This could be due to the distinction between general trust, reflected in the propensity to trust automation, and specific, task-related trust measured at the start of the experiment.

Unexpectedly, forgiveness was not associated with trust repair [T2-T3], indicating that this personality trait may be irrelevant in this context. Additionally, perceived threat was positively correlated with perceived mistakes and negatively correlated with the final trust measure. This suggests that participants who found the setting more threatening rated the drone's performance more negatively and had lower trust in the agent by the end of the task compared to those who perceived the setting as less threatening.

Table 3. *Correlation matrix with trust measures and personality traits.*

		Min	Max.	M	SD	1	2	3	4	5	6	7
1	first trust measure (T1 from first run)	0	6	4.2	0.9							
2	last trust measure (T3 from last run)	0	6	2.1	1.7	.19						
3	trust repair [T2-T3]	0	6	1.1	1.4	.12	.72**					
4	trust decline [T1-T2]	0	6	-2.8	1.1	-.28*	-.11	-.63**				
5	Propensity to trust	1	5	3.6	0.4	.20	.27*	.27*	-.03			
6	Forgiveness	1	7	5.1	0.9	.02	.05	.09	-.14	.23		
7	Perceived mistake (negative drone performance)	1	5	4.5	0.5	-.09	-.32*	-.19	-.06	-.34*	-.07	
8	Perceived threat (fear + danger)	1	5	3.4	0.8	-.21	-.46**	-.21	-.11	-.27*	-.19	.39**

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

3.2. Trust

We performed an ANOVA with the between-subject factor Agent type (human-like or machine-like) and the within-subject factors Explanation (present or absent) and Time (prior to violation [T1] vs. after violation [T2] vs. after repair [T3]). The dependent variable was Trust (Figure 8).

For the main effect of Time, Mauchly's test of sphericity indicated a violation of the sphericity assumption, $\chi^2(2) = 6{,}41$, $p = .041$. Since sphericity is violated ($\varepsilon = 0.942$), Huynh-Feldt corrected results are reported. A significant main effect for Time on Trust was obtained (F (1.88, 97.96) = 153.13, $p < .001$). LSD post hoc comparisons showed significantly decreased trust from T1 ($M = 4.0$) to T2 ($M = 1.2$) ($\Delta M = -2.8$, $p < .001$), which reflects the intended trust-violating effect of the encounter with the unexpected hazards. Post-hoc further showed a significant rise in trust between T2 and T3 ($M = 2.3$) ($\Delta M = 1.1$, $p < .001$), which reflects a general recovery of trust in the final phase of the searches.

All other effects were statistically non-significant. The analysis revealed that the main effect of Explanation on Trust was not significant, F (1, 52) = 1.57, $p = 0.216$, nor was the main effect of Agent type on Trust, F (1, 52) = 0.83, $p = 0.367$. The two-way interaction between Time and Explanation on Trust was non-significant (F (2, 104) = 0.26, $p = 0.772$), with sphericity assumed and non-corrected results reported ($\chi^2(2) = 3.21$, $p = .201$). The two-way interaction-effect between Time and Agent type on Trust was also non-significant (F (2, 104) = 0.58, $p = 0.563$). Finally, the three-way interaction effect of Time, Explanation and Agent type on Trust was non-significant as well (F (2, 104) = 0.86, $p = 0.425$).

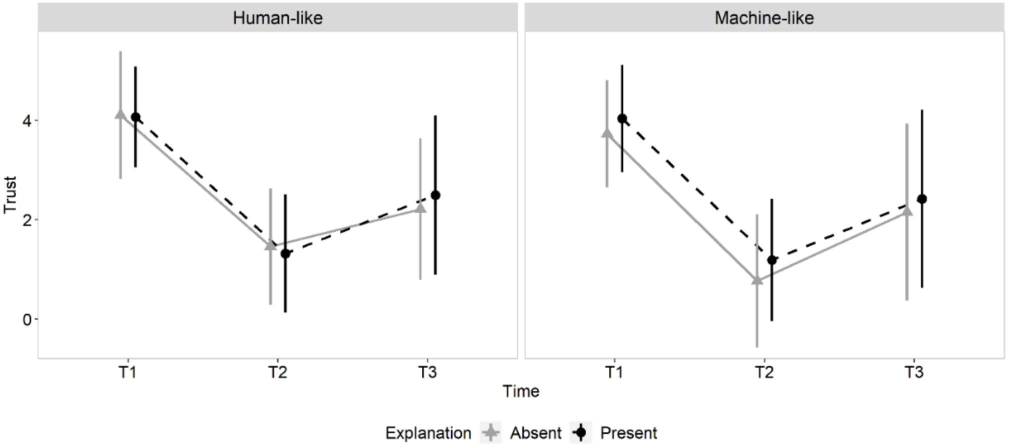


Fig 8. *The x-axis represents Time, while the y-axis represents Trust. Participants were asked "How would you rate your current level of trust in the agent?" with the slider ranging from 0 (Very low) to 6 (Very high). Different types of lines indicate whether an explanation was provided (solid grey = absent, dashed black = present). The left side of the grid displays data from participants who interacted with the Human-like agent (N = 28), while the right side shows data from those who interacted with the Machine-like agent (N = 26). Error bars denote standard deviations.*

4. Discussion

4.1. Key findings

Our results show that the presence of anthropomorphic cues did not affect trust at any stage of the trust life cycle. Through all phases of the task, the two agent types were trusted similarly. Although we expected the human-like agent to be trusted more initially and to get penalized less for its mistake, we did not observe this dampening effect. We also expected a greater recovery in trust for the human-like agent compared to the machine-like agent in the final phase of the task, especially in combination with the explanation. However, neither agent type, nor the explanation, nor the combination resulted in a significant rise of trust.

Our results also indicate that, although the agent type manipulation was subtle, it was still impactful enough for the human-like agent to be perceived as significantly more human-like than the machine-like agent. We deliberately kept our manipulation subtle, avoiding overly anthropomorphic features in physical appearance to maintain realism within a military context and enable more systematic comparisons. Similarly, the nature of the explanations was somewhat technical. Although we considered tailoring the content of the explanations to align with the agent-type condition, we decided to keep them identical to ensure comparability and because these specific explanations were effective in our prior research [39]. This approach aims to ensure the ecological validity of our study and to more accurately reflect the scenarios participants might encounter in real-world military settings. Our finding that the human-like agent was still perceived as significantly more human-like than the machine-like agent aligns with previous research showing how relatively simple anthropomorphic cues can significantly alter our perceptions of agents [60, 56, 57]. This effect was further demonstrated in our online pilot study, where only the audio of a 30-second video and an image accompanying the introductory text differed, was sufficient to produce significant differences in how the drone was perceived in terms of anthropomorphism, intelligence, likeability, teammate-likeness, and perceived trustworthiness. However, in contrast to the online pilot, the main experiment did not reveal any significant differences between the agent types regarding other constructs, such as perceived intelligence or trustworthiness.

The lack of significant effects could also be explained by the overwhelming nature of the VR task environment. Koskinen et al. (2019) conclude that when people have to rely on automation in stressful contexts, individual differences in terms of coping mechanisms to deal with stressful circumstances, perceptions and attitudes regarding trustworthiness and a participant's readiness for risk tasking might be more important to trust development than solely the technical characteristics or design features of artificial agents [36]. Naturally, trust is influenced heavily by the correctness of the algorithmic advice [72], the risk associated with the situation [44] and the severity of the violation and its consequences [55]. In our case, experiencing the consequences of an incorrect advice in VR seems to outweigh the source of the advice (human-like vs. machine-like) or the presence of an explanation in their impact on the development of trust.

This study contributes to an ongoing debate about the implications and the appropriateness of humanizing AI-interaction and the extent to which anthropomorphism should be implemented [31, 75, 76]. In practice, the appropriateness of using human-like cues depends heavily on the context. Rather than advocating that AI agents should be one way or another, we argue for considering the ethical implications and context-specific boundaries of each application, rather than treating AI as a homogenous concept [29]. In social robotics, these cues might be beneficial, while in professional settings where safety and calibrated trust are crucial, they may be less appropriate. The use of anthropomorphic features in design should be guided by a clear understanding of the potential effects on user interactions and should be approached with careful and strategic consideration, as these features can have serious consequences [31, 9, 15, 34]. By focusing on context, we ensure more responsible and effective implementation that meets user needs and fosters appropriate trust.

4.2. Methodological reflection

VR presents both unique opportunities and significant challenges when researching trust. While VR allows for the creation of immersive, controlled environments that can simulate detailed, complex and realistic scenarios, it also introduces practical and methodological hurdles. Through our research, we encountered several challenges that impacted our ability to assess trust dynamics. In this section, we share the lessons learned from our experiences, including the insights from the extensive VR pilot. We discuss the challenges we encountered in accurately and consistently measuring the effects of our intended manipulations, as well as the opportunities we see in using VR as a research tool for studying trust in HRI.

4.2.1. Challenges

Navigating VR locomotion with the Virtualizer

While the Cyberith Virtualizer ELITE 2 is an excellent solution for enabling users to walk endlessly in a virtual space without encountering physical obstacles, thereby enhancing immersion and possibilities, it also introduced some challenges. The Cyberith Virtualizer ELITE 2 treadmill introduced considerable variability in how participants navigated the virtual space, largely due to differences in walking techniques. It proved challenging for many participants to walk naturally on the treadmill. While participants were securely fastened into a harness that provided support and stability, achieving smooth and consistent motion was difficult. Participants wore special shoes with low-friction soles, designed to glide smoothly across the base of the Virtualizer, akin to using fingers on a laptop trackpad. We often explained the technique as "similar to doing Michael Jackson's moonwalk," where participants would lean slightly forward and slide their feet across the base. The participant who moved most fluently through the environment used a unique approach, sliding one foot from the front to the back of the base while keeping the other foot stationary, much like pushing a scooter or scrolling on a trackpad.

This variability in movement led to significant inconsistencies in how participants navigated the virtual environment. Some participants moved erratically due to their effective walking style, with some flying across the building's small corridors, reaching the end of a 6-meter hallway in a single step. This not only increased the risk of cyber sickness but also significantly impacted how quickly and attentively participants moved through the environment. The timing of when triggers were activated, as well as the time, calmness, and attention participants had to process their surroundings and the experimental manipulations, were all affected by these inconsistencies. We advise future researchers making use of this technology to give participants more time to practice before beginning the task.

Managing cyber sickness

Participants interacted with the virtual environment using the Virtualizer ELITE 2 360 treadmill, Oculus Rift VR glasses, and Oculus Touch hand controllers. For smooth and effective interaction, it was crucial that all hardware and software were consistently aligned and calibrated. However, the Virtualizer ELITE 2 did not always integrate seamlessly with the VR headset, which is crucial for ensuring that the user's physical movements were accurately mirrored in the virtual world. Even slight misalignments with the treadmill caused participants to veer slightly off course, leading them to walk into walls and objects, breaking the immersion and sometimes causing cyber sickness. Cyber sickness, which can result from a suboptimal VR experience, includes symptoms like nausea, disorientation, discomfort, eye strain, and drowsiness.

Many participants had to quit the VR experiment due to cyber sickness. Because participants often felt obliged to continue, we had to be vigilant in observing how they looked and behaved. We consistently asked if they were feeling okay and emphasized that it was perfectly fine to stop if they were experiencing discomfort. We noticed that participants, who were always students, felt a sense of a responsibility to perform well and provide good data. It helped to explain that continuing the experiment while feeling unwell could compromise data quality, and we therefore preferred that they cease participation if they were not feeling well. This approach often led participants to acknowledge that they were not feeling well, and they chose to stop the experiment.

Misinterpreted trust-violating events

Thanks to the VR pilot, we gained valuable insights into how participants perceived the environment, leading to significant changes in the virtual scenes. Importantly, we learned that the trust-violating events (i.e., a thief and a smoking IED) did not always succeed in provoking a trust violation. For instance, not all participants recognized the IED (Improvised Explosive Device) as a bomb; some mistook it for a box and attempted to pick it up. As they misunderstood the threat posed by the IED, they continued to search the room or inspect the object more closely, inadvertently triggering the bomb animation (smoking and beeping) multiple times. This repeated activation diminished the immersion and reduced the intended startling effect. Based on these observations, we redesigned the virtual IED to more closely resemble a stereotypical explosive device commonly depicted in movies, featuring a prominent ticking clock. This adjustment was made to ensure participants would more readily recognize it as a potential threat, avoiding confusion with non-threatening objects.

Similarly, the thief animation, which caused the thief to run out through a door leading outside, often triggered too early as participants were still exploring the kitchen area, where there was much to observe. As a result, many participants were still processing the previous scene and did not have a clear view of the thief or enough time to recognize it as a threat. Some participants attempted to chase the thief as it ran off. One participant, realizing it was a threat, even tried to punch the thief, resulting in an in-air punch of the virtual character—demonstrating the level of immersion. As participants did not always understand what was happening, they might not always have realized that the drone's advice ("no danger detected") was incorrect, thus failing to evoke a trust violation.

To address these issues, we revised the task instructions to emphasize that participants' primary goal was to ensure their own safety during the searches, regardless of what they encountered. Any findings could be reported after the search was completed, with the main objectives being to stay safe and finish the task. To determine whether participants perceived the robot's error as an error, it is crucial to include a manipulation check question in the final questionnaire, as we did.

Unnoticed explanations

We noticed that the robot's explanation was not always clearly heard and understood. The explanation trigger was placed immediately after the trust-violating event, but participants were often too preoccupied with processing the event (such as a thief running off or an IED starting to smoke) to hear the drone's message clearly. During debriefings, participants frequently mentioned that the environment was overwhelming and that they were focused on the task, causing them to miss specific details.

Furthermore, because the drone was the audio source in VR, the volume decreased as participants moved farther away. This created a problem when participants lagged behind the drone, making it difficult to hear its messages. In the main experiment, we addressed this by changing the audio source from the drone to the participant's headphones, ensuring a constant volume regardless of their position relative to the drone. However, even with this adjustment, many participants were still focused on processing the previous scene where the trust violation occurred, leading to them not clearly remembering whether or what the drone had said. It might be beneficial to give participants more time to acclimate to the VR environment before beginning the task. We also recommend researchers using VR to consider the potentially overwhelming nature of the environment when designing manipulations. To achieve the desired effects, we suggest making the manipulations bolder and more apparent than you would make them in a less stimulating setup or environment.

4.2.2. Opportunities

Inducing a sense of threat without posing physical risk to participants

Trust, by definition, is only relevant in situations characterized by risk, uncertainty and vulnerability [44]. Studying trust repair requires violating trust and allowing people to experience the risk they take and the vulnerability they accept. It can be challenging to create experimental scenarios that induce feelings of vulnerability and risk without compromising participants' physical and psychological safety [3].

The use of VR is a promising solution to this challenge. VR offers emotional engagement of participants [65] and ecological validity, while remaining experimental control, reproducibility [62]. More than 2D screen videos or game-like virtual environments, VR has the ability to create a strong sense of presence and to increase sympathetic activation significantly [11]. Based on our observations of people's reactions during the data collection, we suspect that the current VR setting has intensified feelings of trust, risk and betrayal after a trust violation in comparison to our previous study [38]. Participants often spoke aloud as though they were directly communicating with the drone, suggesting a high level of immersion in the virtual environment.

We aimed to induce a feeling of threat and risk as this would allow us to study how people naturally respond to unexpected events in a risky situation. Throughout the experiment, this was often visible as a number of participants startled, flinched or cursed when the burglar or the bomb were encountered. The houses in our VR task were to a certain extent designed to evoke a feeling of threat. Several times, participants would comment on the design, saying it is "creepy" or "like in a horror movie", indicating that the environment was indeed threatening. The scene is realistic yet highly controllable and allows us to directly observe behaviour.

We recognize that introducing startling elements in research raises ethical considerations, particularly in balancing the scientific value of the study against the potential for participant discomfort. The startling animations were included to study specific behaviors under stress. By informing participants beforehand that there was a chance of running into hazards, providing them with the right to withdraw at any point without penalty, and debriefing them afterward, we aimed to minimize any potential harm while achieving the study's objectives.

Most prior trust repair studies used relatively easy, low-risk online tasks [81, 22, 34]. Some research has shown that in less threatening contexts, people may even prefer erroneous robots over flawless ones [55, 22]. However, studies involving erroneous robots in more safety-critical situations, such as emergency evacuations, indicate that agent failure can be detrimental to people's attitudes towards the robot [69, 68]. This highlights that trust violations in high-risk and high-threat contexts have a greater impact on trust than those in low-risk environments. These findings suggest that perceived risk is a critical factor to consider when researching trust development. The perception of potential increases dependence on the robot, thereby amplifying the importance of trust. The use of VR provides a means to effectively induce a sense of threat in a controlled and safe manner.

Potential for collecting behavioural measures

Placing a greater focus on behavioural measures potentially related to trust could significantly contribute to the field of HRI by providing more quantitative metrics to assess trust. The VR environments we developed are ideally suited for such purposes [37]. For instance, walking speed can be measured, which may indicate hesitation. Additionally, eye movements tracked through the VR headset in both scenarios could reveal excessive monitoring. Both hesitation and excessive monitoring might point to lowered trust. Incorporating these objective measures into HRI trust research would be worthwhile, as it can increase the accuracy and reliability of trust measures. Due to technical constraints, we were not able to capture walking speed and eye-movement. However, VR offers great potential for combining self-report, behavioural, and in time perhaps physiological measures as possible proxies for trust.

4.3. Conclusion

In this study, we set out to explore the dynamics of trust in HRI, focusing on how trust forms, breaks down and recovers, using an novel, high-fidelity VR task environment that simulated a realistic military HRI scenario. Results indicated that, although the anthropomorphic cues used in the agent type manipulation were subtle, participants perceived the human-like robot as significantly more human-like than the machine-like robot. However, neither the anthropomorphic cues nor the presence of an explanation had a significant effect on trust development. Despite the challenges we encountered using VR as a research tool, we aim to provide insights for fellow researchers examining HRI trust dynamics and hope to initiate new investigations by sharing our research design, paradigm and the methodological challenges.

5. Acknowledgements

The authors would like to thank Lucía Margarita Rábago Mayer, Arda Temel (the BMS lab, University of Twente), Jonathan Barnhoorn, and Tessa Klunder (TNO Soesterberg) for their contributions to the development of the Virtual Reality environment and Anna Thomsen, Lauritz Dauer, Leona Weise, Julie Erber, Joline Klein, Dillen Benerink, Marc Pflieger, and Clara Odenthal (University of Twente) for their contributions to the data collection.

6. References

[1] AIHLEG, “A Definition of AI: Main Capabilities and Disciplines,” 2019. [Online]. Available: https://ec.europa.eu/digital-single-.

[2] Y. Albayram, T. Jensen, M. M. H. Khan, M. A. Al Fahim, R. Buck, and E. Coman, “Investigating the Effects of (Empty) Promises on Human-Automation Interaction and Trust Repair,” *HAI 2020 - Proc. 8th Int. Conf. Human-Agent Interact.*, pp. 6–14, 2020, doi: 10.1145/3406499.3415064.

[3] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler, “Toward an understanding of trust repair in human-robot interaction: Current research and future directions,” *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 4, 2018, doi: 10.1145/3181671.

[4] M. J. Barnes *et al.*, “Designing for Humans in Autonomous Systems: Military Applications,” Aberdeen Proving Ground, Maryland, 2014.

[5] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots,” *Int. J. Soc. Robot.*, vol. 1, no. 1, pp. 71–81, 2009, doi: 10.1007/s12369-008-0001-3.

[6] P. Bobko *et al.*, “Human-agent teaming and trust calibration: a theoretical framework, configurable testbed, empirical illustration, and implications for the development of adaptive systems,” *Theor. Issues Ergon. Sci.*, vol. 0, no. 0, pp. 1–25, 2022, doi: 10.1080/1463922X.2022.2086644.

[7] D. Cameron, E. J. Loh, A. Chua, E. C. Collins, J. M. Aitken, and J. Law, “Robot-stated limitations but not intentions promote user assistance,” 2016, [Online]. Available: http://arxiv.org/abs/1606.02603.

[8] D. Cameron *et al.*, “The effect of social-cognitive recovery strategies on likability, capability and trust in social robots,” *Comput. Human Behav.*, vol. 114, no. September, p. 106561, 2021, doi: 10.1016/j.chb.2020.106561.

[9] J. Carpenter, “The Quiet Professional: An investigation of U.S. military Explosive Ordnance Disposal personnel interactions with everyday field robots,” *ProQuest Diss. Theses*, vol. 3599678, p. 170, 2013.

[10] B. X. Chen, “How to Use ChatGPT and Still Be a Good Person,” *The New York Times*, New York, Dec. 21, 2022.

[11] A. Chirico, P. Cipresso, D. B. Yaden, F. Biassoni, G. Riva, and A. Gaggioli, “Effectiveness of Immersive Videos in Inducing Awe: An Experimental Study,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–11, 2017, doi: 10.1038/s41598-017-01242-0.

[12] D. De Cremer and G. Kasparov, “AI Should Augment Human Intelligence, Not Replace It,” *Harv. Bus. Rev.*, pp. 1–9, 2021.

[13] K. E. Culley and P. Madhavan, “A note of caution regarding anthropomorphism in HCI agents,” *Comput. Human Behav.*, vol. 29, no. 3, pp. 577–579, 2013, doi: 10.1016/j.chb.2012.11.023.

[14] Cyberith, “Virtualizer ELITE 2 ‘The Motion Platform Motion Platform,’” *Cyberith*. https://www.cyberith.com/virtualizer-elite/ (accessed Aug. 19, 2024).

[15] C. F. Disalvo, F. Gemperle, J. Forlizzi, and S. Kiesler, “All Robots Are Not Created Equal : The Design and Perception of Humanoid Robot Heads,” 2002.

[16] A. Duenser and D. M. Douglas, “Whom to Trust, How and Why: Untangling Artificial Intelligence Ethics Principles, Trustworthiness, and Trust,” *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 19–26, 2023, doi: 10.1109/MIS.2023.3322586.

[17] B. R. Duffy, “Anthropomorphism and the social robot,” *Rob. Auton. Syst.*, vol. 42, no. 3–4, pp. 177–190, 2003, doi: 10.1016/S0921-8890(02)00374-3.

[18] N. Epley, A. Waytz, and J. T. Cacioppo, “On Seeing Human: A Three-Factor Theory of Anthropomorphism,” *Psychol. Rev.*, vol. 114, no. 4, pp. 864–886, 2007, doi: 10.1037/0033-295X.114.4.864.

[19] D. Gambetta, “Can We Trust Trust?,” in *Trust: Making and Breaking Cooperative Relations*, Electronic., Oxford: Department of Sociology, University of Oxford, 2000, pp. 212–237.

[20] H. M. Gray, K. Gray, and D. M. Wegner, “Dimensions of mind perception,” *Science (80-.).*, vol. 315, no. 5812, p. 619, 2007, doi: 10.1126/science.1134475.

[21] K. Hald, K. Weitz, E. André, and M. Rehm, “‘An Error Occurred!’ - Trust Repair With Virtual Robot Using Levels of Mistake Explanation,” in *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI ’21)*, 2021, vol. 3, no. 1, p. 9, [Online]. Available: http://journal.unilak.ac.id/index.php/JIEB/article/view/3845%0Ahttp://dspace.uc.ac.id/handle/123456789/1288.

[22] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder, “Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction,” *25th IEEE Int. Symp. Robot Hum. Interact. Commun. RO-MAN 2016*, pp. 493–500, 2016, doi: 10.1109/ROMAN.2016.7745163.

[23] P. A. Hancock, D. R. Billings, and K. E. Schaefer, “Can you trust your robot?,” *Ergon. Des.*, vol. 19, no. 3, pp. 24–29, 2011, doi: 10.1177/1064804611415045.

[24] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, “A meta-analysis of factors affecting trust in human-robot interaction,” *Hum. Factors*, vol. 53, no. 5, pp. 517–527, 2011, doi: 10.1177/0018720811417254.

[25] T. R. Herzog and G. E. Kutzli, “Preference and perceived danger in field/forest settings,” *Environ. Behav.*, vol. 34, no. 6, pp. 819–835, 2002, doi: 10.1177/001391602237250.

[26] K. A. Hoff and M. Bashir, “Trust in automation: Integrating empirical evidence on factors that influence trust,” *Hum. Factors*, vol. 57, no. 3, pp. 407–434, 2015, doi: 10.1177/0018720814547570.

[27] R. R. Hoffman, “A taxonomy of emergent trusting in the human-machine relationship,” *Cogn. Syst. Eng. Futur. a Chang. World*, pp. 137–164, 2017, doi: 10.1201/9781315572529.

[28] M. Hou, G. Ho, and D. Dunwoody, “IMPACTS: a trust model for human-autonomy teaming,” *Human-Intelligent Syst. Integr.*, vol. 3, no. 2, pp. 79–97, 2021, doi: 10.1007/s42454-020-00023-x.

[29] E. Jermutus, D. Kneale, J. Thomas, and S. Michie, “Influences on User Trust in Healthcare Artificial Intelligence: A Systematic Review,” *Wellcome Open Res.*, vol. 7, p. 65, 2022, doi: 10.12688/wellcomeopenres.17550.1.

[30] S. A. Jessup, “Measurement of the propensity to trust automation,” *Organ. Behav. Hum. Decis. Process.*, vol. 50, no. 2, pp. 179–211, 2018.

[31] J. Johnson, “Finding AI Faces in the Moon and Armies in the Clouds: Anthropomorphising Artificial Intelligence in Military Human-Machine Interactions,” *Glob. Soc.*, vol. 38, no. 1, pp. 67–82, 2024, doi: 10.1080/13600826.2023.2205444.

[32] N. A. Jones, H. Ross, T. Lynam, P. Perez, and A. Leitch, “Mental models: An interdisciplinary synthesis of theory and methods,” *Ecol. Soc.*, vol. 16, no. 1, 2011, doi: 10.5751/ES-03802-160146.

[33] P. H. Kim, K. T. Dirks, C. D. Cooper, and D. L. Ferrin, “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,” *Organ. Behav. Hum. Decis. Process.*, 2006, doi: 10.1016/j.obhdp.2005.07.002.

[34] T. Kim and H. Song, “How should intelligent agents apologize to restore trust ?: The interaction effect between anthropomorphism and apology attribution on trust repair,” *Telemat. Informatics*, 2021.

[35] S. C. Kohn, D. B. Quinn, R. Pak, E. J. de Visser, and T. H. Shaw, “Trust repair strategies with self-driving vehicles: An exploratory study,” *Proc. Hum. Factors Ergon. Soc.*, vol. 2, pp. 1108–1112, 2018, doi: 10.1177/1541931218621254.

[36] K. M. Koskinen, A. Lyyra, N. Mallat, and V. Tuunainen, "Trust and risky technologies: Aligning and coping with Tesla Autopilot," *Proc. 52nd Hawaii Int. Conf. Syst. Sci.*, pp. 5777–5786, 2019, doi: 10.24251/hicss.2019.697.

[37] E. S. Kox, J. Barnhoorn, L. Rábago Mayer, A. Temel, and T. Klunder, "Using a Virtual Reality House-Search Task to Measure Trust During Human-Agent Interaction (Demo Paper)," in *HHAI2022: Augmenting Human Intellect*, 2022, pp. 272–274, doi: 10.3233/FAIA220214.

[38] E. S. Kox, J. H. Kerstholt, T. Hueting, and P. W. de Vries, "Trust Repair in Human-Agent Teams: the Effectiveness of Explanations and Expressing Regret," *Auton. Agent. Multi. Agent. Syst.*, vol. 35, no. 2, pp. 1–20, 2021, doi: 10.1007/s10458-021-09515-9.

[39] E. S. Kox, L. B. Siegling, and J. H. Kerstholt, "Trust Development in Military and Civilian Human-Agent Teams: the Effect of Social-Cognitive Recovery Strategies," *Int. J. Soc. Robot.*, 2022, doi: 10.1007/s12369-022-00871-4.

[40] A. Y. Lee, G. D. Bond, D. C. Russell, J. Tost, C. González, and P. S. Scarbrough, "Team perceived trustworthiness in a complex military peacekeeping simulation," *Mil. Psychol.*, vol. 22, no. 3, pp. 237–261, 2010, doi: 10.1080/08995605.2010.492676.

[41] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[42] M. K. Lee, S. Kiesler, J. Forlizzi, S. S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," *2010 5th ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 203–210, 2010, doi: 10.1109/HRI.2010.5453195.

[43] M. Lewis, K. Sycara, and P. Walker, *The Role of Trust in Human-Robot Interaction*, vol. 117. 2018.

[44] M. Li, B. E. Holthausen, R. E. Stuck, and B. N. Walker, "No risk no trust: Investigating perceived risk in highly automated driving," *Proc. - 11th Int. ACM Conf. Automot. User Interfaces Interact. Veh. Appl. AutomotiveUI 2019*, no. September, pp. 177–185, 2019, doi: 10.1145/3342197.3344525.

[45] E. B. Lozano and S. M. Laurent, "The effect of admitting fault versus shifting blame on expectations for others to do the same," *PLoS One*, vol. 14, no. 3, pp. 1–19, 2019, doi: 10.1371/journal.pone.0213276.

[46] J. B. Lyons, I. aldin Hamdan, and T. Q. Vo, "Explanations and trust: What happens to trust when a robot partner does something unexpected?," *Comput. Human Behav.*, vol. 138, no. February 2022, p. 107473, 2023, doi: 10.1016/j.chb.2022.107473.

[47] M. Madsen and S. Gregor, "Measuring Human-Computer Trust," *Proc. Elev. Australas. Conf. Inf. Syst.*, pp. 6–8, 2000, [Online]. Available: http://books.google.com/books?hl=en&lr=&id=b0yalwi1HDMC&oi=fnd&pg=PA102&dq=The+Big+Five+Trait+Taxonomy:+History,+measurement,+and+Theoretical+Perspectives&ots=758BNaTvOi&sig=L52e79TS6r0Fp2m6xQVESnGt8mw%5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=.

[48] J. E. Mathieu, T. S. Heffner, G. Goodwin, E. Salas, and J. A. Cannon-Bowers, "The Influence of Shared Mental Models on Team Process and Performance," *J. Appl. Psychol.*, vol. 85, no. 2, pp. 273–283, 2000.

[49] G. Matthews, J. Lin, A. R. Panganiban, and M. D. Long, "Individual Differences in Trust in Autonomous Robots: Implications for Transparency," *IEEE Trans. Human-Machine Syst.*, vol. PP, no. November, pp. 1–11, 2019, doi: 10.1109/THMS.2019.2947592.

[50] G. Matthews, A. R. Panganiban, J. Lin, M. D. Long, and M. Schwing, "Super-machines or sub-humans: Mental models and trust in intelligent autonomous systems," in *Trust in Human-Robot Interaction*, Elsevier Inc., 2021, pp. 59–82.

[51] R. C. Mayer, J. H. Davis, and D. F. Schoorman, "An Integrative Model of Organizational Trust," *Acad. Manag. Rev.*, vol. 20, no. 3, pp. 709–734, 1995, doi: 10.1109/GLOCOM.2017.8254064.

[52] D. H. McKnight and N. L. Chervany, "What is Trust ? A Conceptual Analysis and an Interdisciplinary Model," *Proc. 2000 Am. Conf. Inf. Syst. AMCI2000 AIS Long Beach CA August 2000*, vol. 346, p. 382, 2000, [Online]. Available: http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1876&amp;context=amcis2000.

[53] Merriam-Webster, "Robot," *Merriam-Webster Dictionary*. https://www.merriam-webster.com/dictionary/robot (accessed Aug. 15, 2024).

[54] C. A. Miller, "Delegation and Transparency: Coordinating Interactions So Information Exchange Is No Surprise," in *Proceedings of the 6th International Conference of Virtual, Augmented and Mixed Reality (VAMR), Part I*, 2014, vol. 8525 LNCS, no. PART 1, pp. 191–202, doi: 10.1007/978-3-319-07458-0_8.

[55] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, "To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot," *Front. Robot. AI*, vol. 4, no. May, pp. 1–15, 2017, doi: 10.3389/frobt.2017.00021.

[56] S. Moran *et al.*, "Team reactions to voiced agent instructions in a pervasive game," *Int. Conf. Intell. User Interfaces, Proc. IUI*, pp. 371–382, 2013, doi: 10.1145/2449396.2449445.

[57] S. Moussawi, M. Koufaris, and R. Benbunan-Fich, "How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents," *Electron. Mark.*, 2020, doi: 10.1007/s12525-020-00411-w.

[58] C. Nam, P. Walker, M. Lewis, and K. Sycara, "Predicting trust in human control of swarms via inverse reinforcement learning," *RO-MAN 2017 - 26th IEEE Int. Symp. Robot Hum. Interact. Commun.*, vol. 2017-Janua, pp. 528–533, 2017, doi: 10.1109/ROMAN.2017.8172353.

[59] T. O'Neill, N. J. McNeese, A. Barron, and B. G. Schelble, "Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature," *Hum. Factors*, vol. 64, no. 5, pp. 904–938, 2022, doi: 10.1177/0018720820960865.

[60] R. Pak, N. Fink, M. Price, B. Bass, and L. Sturre, "Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults," *Ergonomics*, vol. 55, no. 9, pp. 1059–1072, 2012, doi: 10.1080/00140139.2012.691554.

[61] R. Pak and E. Rovira, "A theoretical model to explain mixed effects of trust repair strategies in autonomous systems," *Theor. Issues Ergon. Sci.*, 2023, doi: 10.1080/1463922X.2023.2250424.

[62] X. Pan and A. F. d. C. Hamilton, "Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape," *Br. J. Psychol.*, vol. 109, no. 3, pp. 395–417, 2018, doi: 10.1111/bjop.12290.

[63] S. Pang, "The Truth About Saying 'Thanks' & 'Please' To ChatGPT," *Medium*, 2023. https://ppangsy.medium.com/a-world-of-thanks-the-surprising-effects-of-chatgpt-appreciation-105f6e25fcce (accessed Jul. 25, 2024).

[64] S. K. Parker and G. Grote, "Automation, Algorithms, and Beyond: Why Work Design Matters More Than Ever in a Digital World," *Appl. Psychol.*, vol. 71, no. 4, pp. 1171–1204, 2022, doi: 10.1111/apps.12241.

[65] T. D. Parsons, "Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences," *Front. Hum. Neurosci.*, vol. 9, no. DEC, pp. 1–19, 2015, doi: 10.3389/fnhum.2015.00660.

[66] E. K. Phillips, S. Ososky, J. Grove, and F. G. Jentsch, "From tools to teammates: Toward the development of appropriate mental models for intelligent robots," *Proc. Hum. Factors Ergon. Soc.*, pp. 1491–1495, 2011, doi: 10.1177/1071181311551310.

[67] P. Robinette, A. M. Howard, and A. R. Wagner, "Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed?," in *Autonomy and Artificial Intelligence: A Threat or Savior?*, 2017, pp. 129–156.

[68] P. Robinette, A. M. Howard, and A. R. Wagner, "Timing is key for robot trust repair," in *International conference on social robotics*, 2015, vol. 9388 LNCS, pp. 574–583, doi: 10.1007/978-3-319-25554-5_46.

[69] P. Robinette, A. M. Howard, and A. R. Wagner, "Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations," *IEEE Trans. Human-Machine Syst.*, vol. 47, no. 4, pp. 425–436, 2017, doi: 10.1109/THMS.2017.2648849.

[70] W. B. Rouse and N. M. Morris, "On looking into the black box: Prospects and limits in the search for mental models," Atlanta, GA, 1985. [Online]. Available:

http://scholar.google.com/scholar?q=related:QM4p5zGC8jMJ:scholar.google.com/&amp;hl=en&amp;num=30&amp;as_sdt=0,5.

[71] D. M. Rousseau, S. B. Sitkin, R. S. Burt, C. Camerer, D. M. Rousseau, and R. S. Burt, "Not so Different after All: A Cross-Discipline View of Trust," *Acad. Manag. Rev.*, vol. 23, no. 3, pp. 393–404, 1998.

[72] A. Schmitt, T. Wambsganss, M. Söllner, and A. Janson, "Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice," in *Forty-Second International Conference on Information Systems*, 2021, no. October.

[73] J. M. Schraagen and J. van Diggelen, "A Brief History of the Relationship Between Expertise and Artificial Intelligence," *Expert. Work*, pp. 149–175, 2021, doi: 10.1007/978-3-030-64371-3_8.

[74] T. B. Sheridan, "Individual differences in attributes of trust in automation: Measurement and application to system design," *Front. Psychol.*, vol. 10, no. MAY, pp. 1–7, 2019, doi: 10.3389/fpsyg.2019.01117.

[75] B. Shneiderman, "Human-Centered Artificial Intelligence: Three Fresh Ideas," *AIS Trans. Human-Computer Interact.*, vol. 12, no. 3, pp. 109–124, 2020, doi: 10.17705/1thci.00131.

[76] B. Shneiderman, "Human-Centered AI," *ISSUES Sci. Technol.*, pp. 43–44, 2021, doi: 10.2307/j.ctv1s5nzbk.19.

[77] C. A. Smith, "The Uses of Pilot Studies in Sociology: a Processual Understanding of Preliminary Research," *Am. Sociol.*, vol. 50, no. 4, pp. 589–607, 2019, doi: 10.1007/s12108-019-09419-y.

[78] M. Söllner and P. A. Pavlou, "A longitudinal perspective on trust in it artefacts," *24th Eur. Conf. Inf. Syst. ECIS 2016*, no. June, 2016.

[79] S. Soltanzadeh, "Strictly Human: Limitations of Autonomous Systems," *Minds Mach.*, vol. 32, no. 2, pp. 269–288, 2022, doi: 10.1007/s11023-021-09582-7.

[80] D. Ullrich, A. Butz, and S. Diefenbach, "The Development of Overtrust: An Empirical Simulation and Psychological Analysis in the Context of Human–Robot Interaction," *Front. Robot. AI*, vol. 8, no. April, pp. 1–15, 2021, doi: 10.3389/frobt.2021.554578.

[81] E. J. de Visser *et al.*, "Almost human: Anthropomorphism increases trust resilience in cognitive agents.," *J. Exp. Psychol. Appl.*, vol. 22, no. 3, pp. 331–349, Sep. 2016, doi: 10.1037/xap0000092.

[82] E. J. de Visser, R. Pak, and M. A. Neerincx, "Trust development and repair in human-robot teams," *ACM/IEEE Int. Conf. Human-Robot Interact.*, pp. 103–104, 2017, doi: 10.1145/3029798.3038409.

[83] E. J. de Visser, R. Pak, and T. H. Shaw, "From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction," *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, 2018, doi: 10.1080/00140139.2018.1457725.

[84] E. J. de Visser *et al.*, "Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams," *Int. J. Soc. Robot.*, pp. 1–20, Nov. 2019, doi: 10.1007/s12369-019-00596-x.

[85] P. W. de Vries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," *Int. J. Hum. Comput. Stud.*, vol. 58, no. 6, pp. 719–735, 2003, doi: 10.1016/S1071-5819(03)00039-9.

[86] A. Waytz, J. T. Cacioppo, and N. Epley, "Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism," *Bone*, vol. 23, no. 1, pp. 1–7, 2008, doi: 10.1177/1745691610369336.Who.

[87] A. Waytz, J. Heafner, and N. Epley, "The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle," *J. Exp. Soc. Psychol.*, vol. 52, pp. 113–117, 2014, doi: 10.1016/j.jesp.2014.01.005.

[88] K. T. Wynne and J. B. Lyons, "An integrative model of autonomous agent teammate-likeness," *Theor. Issues Ergon. Sci.*, vol. 19, no. 3, pp. 353–374, 2018, doi: 10.1080/1463922X.2016.1260181.

[89] K. T. Wynne and J. B. Lyons, "Autonomous agent teammate-likeness: Scale development and validation.," in *International Conference on Human-Computer Interaction*, 2019, pp. 199–213.

[90] X. Xu and S. Sar, "Do We See Machines the Same Way As We See Humans? A Survey on Mind Perception of Machines and Human Beings," *RO-MAN 2018 - 27th IEEE Int. Symp. Robot Hum. Interact. Commun.*, pp. 472–475, 2018, doi: 10.1109/ROMAN.2018.8525586.

7. Appendix

7.1. Appendix A

Table A1. *Results of the one-way analysis of variance (ANOVA) comparing agent types in the online pilot study.*

Variable	Human-like (n = 15)		Machine-like (n = 17)		
	M	SD	M	SD	p
Perceived anthropomorphism	3.27	0.75	1.86	0.64	.045*
Perceived intelligence	3.79	0.60	2.82	0.82	.001**
Perceived likeability	3.44	0.91	2.80	0.63	.026*
Perceived Agentic Capabilities (agency)	3.40	0.89	3.19	0.71	.471
Task-independent Relationship Building	3.91	0.42	3.32	0.54	.016*
Perceived Benevolent Intent (altruism)	3.43	0.74	2.69	1.08	.035*
Perceived trustworthiness	3.91	0.42	3.32	0.54	.002**

7.2. Appendix B

Table B1. *Items of the trust scale administered after the task.*

Subscale	Item
Ability	The drone is a real expert in detecting danger.
Ability	The drone gives me good advice.
Ability	The drone knows what I need in order to decide properly.
Ability	The drone has a lot of knowledge about how to navigate in this environment.
Benevolence	The drone puts my interests first.
Benevolence	The drone takes my objective into account.
Benevolence	The drone understand my needs.
Integrity	The drone gives me pure advice.
Integrity	The drone is honest.
Integrity	The drone has integrity.