# Red or Blue Door: Exploring the Behavioural Consequences of Trust Violations due to Robot Error or Choice Using a VR Maze

**ESTHER S. KOX**

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

**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; 2. Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands; jose.kerstholt@tno.nl[4]

*Abstract*— Human-robot collaboration offers significant benefits as long as humans understand and trust their robotic partners appropriately. Traditionally, a robot's trustworthiness was primarily determined by its performance, but as robots gain in autonomy and decision-making authority, collaborators should also pay attention to the robot’s priorities and goals. Losses of trust may arise not just from a robot’s failure (i.e., a performance-based trust violation), but may increasingly stem from misaligned priorities between the trustor and the trustee (i.e. a moral-based trust violation). This study aimed to investigate the effects of (1) trust violations due to a robot’s error vs. choice and (2) prior knowledge of the robot's abilities or intentions on the development of trust (i.e., repeated violations and repair) and compliance. How does the nature of a trust violation (error vs. choice) influence the development of trust and to what extent can prior knowledge about the robot’s abilities or intentions mitigate some of the potential negative effects of a violation? Inspired by earlier HRI paradigms, we developed a Virtual Reality (VR) maze that simulated a military search-task where participants ($n$ = 75) collaborated with an autonomous drone. In a series of rooms, participants could choose to follow or disregard the robot's advice. Due to low compliance rates, we could not conduct our planned analyses. Therefore, we only describe the compliance rates and qualitative feedback. While we were yet unable to conclusively address our research question, the aim of this paper is to share our ideas and insights that can help to improve the research paradigm.

***Keywords— Human-Robot Interaction, Trust, Virtual Reality, Trust violations, Decision-Support and Recommender Systems***

## 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

[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

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. Prior research suggests that trust repair strategies, either preventative strategies deployed prior to a potentially trust-violating event (e.g., uncertainty communication [28]) or reactive, implemented after the occurrence of such an event (e.g., an apology [27, 24, 50]), can reduce the negative impact of the event on trust. Yet, there is still uncertainty about the robustness and durability of these effects.

Moreover, most current HRI trust repair literature mainly focuses on repairing trust violations that result from performance-based issues, like errors, technical failures or other forms of reduced robot reliability [16, 58, 49, 13, 12, 39, 9, 24, 31, 52, 63]. Yet, loss of trust may arise not just from a robot's failure to complete a task correctly (i.e., a performance-based trust violation). Conversely, as robots gain autonomy and decision-authority, trust violations may increasingly stem from misaligned values between the trustor and the trustee (i.e. a moral-based trust violation). For example, it is conceivable that autonomous robots programmed to follow a utilitarian approach, e.g., prioritizing team goals over individual safety, and may act in a way that conflicts with some people's priorities.

How is human-robot trust affected when a robot's actions deviate from the priorities of the people it interacts with? Furthermore, would prior knowledge of the robot's pre-programmed priorities change this reaction? To answer these question, the aim of this study was to investigate how prior knowledge and the nature of a trust violation (performance-based vs. moral-based, or in other words: error vs. choice) influence the development of human-robot trust.

To achieve this, we employ a high-fidelity 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. Using a Virtual Maze scenario where participants can chose to either follow or ignore the robot's advice, we incorporate behavioural measures and thus not only measure trust, but also compliance and reaction time. This allows us to study how self-reported trust and compliance are related and how response time in this scenario relates to self-reported trust. For example, whether a longer response time in following the robot's advice (i.e., hesitation) signal lowered trust?

Lastly, we adopt a temporal perspective on trust, which enables us to observe its evolution as events unfold over time. This allows us to study how trust and compliance develop when trust is repeatedly violated, how effective an explanation is as trust repair strategy when implemented twice.

## *1.2. Background*

### *1.2.1. Human-Robot Teams*

Humans are increasingly collaborating with robots, forming Human-Robot Teams (HRTs). Robots are embodied AI-based systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals [1]. The concept of HRTs is promising, since humans and robots possess complementary skills that, when combined, can enhance performance beyond the capabilities of its individual members. For instance, robots can augment human physical abilities, while their AI-driven nature can augment people's cognitive abilities to solve complex problems. At the same time, people excel in offering intuitive and comprehensive solutions to new and uncertain situations [20]. Yet, the extent to which this new form of collaboration can offer advantages largely depends on whether the human operator has a proper understanding of the robot's abilities and intentions and an appropriate level of trust in the robot (i.e., calibrated trust).

### *1.2.2. Trust*

Trust calibration, the continuous process of updating one's trust in a robot to an appropriate level, is crucial for safe and successful Human-Robot Interaction (HRI) [29]. The intended result, calibrated trust, is a balanced relation between the perceived trustworthiness of an AI-agent and its actual trustworthiness [29]. Here, trustworthiness is a property of the AI-system, while perceived trustworthiness is a judgement by the human [11]. Appropriately calibrated trust prevents overreliance (i.e., misuse) on the one hand and underutilization (i.e., disuse) on the other [62, 45]. 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 [5]. People should be able to determine when it is appropriate to rely on AI-agents and when it is best to override them. In other words, maximizing trust is not the objective in HRI, as effective and efficient teamwork requires finding the right balance of trust among team members. Calibrated trust facilitates cooperation and coordination between interdependent actors, which creates a more productive and efficient team [29].

We define human-robot trust as a human's willingness to make oneself vulnerable and to act on a robot's decisions and recommendations in the pursuit of some benefit, with the expectation that the robot is willing and able to help them achieve their goal in an uncertain context where there is risk [29, 48, 18, 55, 14, 35]. In this definition, we included two trust beliefs: 1) the expectation that the agent is willing and 2) that it is capable. While 'willing' is a debatable term for artificial agents, given that they are "inherently amoral agents as they do not possess agency*"* [2] (p.3), this distinction is crucial. Human-robot trust is increasingly understood as a multi-dimensional concept, as people can only successfully collaborate with robots when they are confident that the robot is both capable of (i.e., performance trust) and committed to (i.e., moral trust) accomplishing certain results [36]. Traditionally, when robots were less autonomous and more like tools, their trustworthiness was mainly assessed on the basis of their performance [19, 56]. However, as robots gain in autonomy and decision-making authority, the notion of 'willingness' becomes increasingly relevant, referring to the values and goals prioritized by the robot's programming (i.e., strategic or moral aspects) [36].

#### *1.2.2.1. Trust violations*

Consequently, this multi-dimensional nature of human-robot trust means that *violations* of trust can also stem from both performance issues, such as mistakes, and moral discrepancies, like misaligned goals and values. A trust violation is any kind of behaviour from an AI-agent that decreases a human's trust in it [44]. In prior work, researchers have suggested the concept of a 'trust lifecycle' [51, 58, 59, 57], consisting of multiple phases: formation, violation and repair. Trust formation refers to the phase where trust is built for the first time. In the trust violation phase, trust diminishes due to the occurrence of unexpected, unfavourable, or unwanted behaviour, resulting in a negative experience for the human trustor [51, 57]. In the trust repair phase, a robot can employ strategies to restore trust and facilitate reconciliation after it violated trust [3, 59, 23, 44].

#### *1.2.2.2. Trust repair*

Most current HRI trust repair literature mainly focuses on repairing trust violations that result from performance-based issues, like errors, technical failures or other forms of reduced robot reliability [16, 58, 49, 13, 12, 39, 9, 24, 31, 52, 63]. Yet, losses of trust may arise not just from a robot's failure to complete a task correctly (i.e., a performance-based trust violation). Trust violations may increasingly stem from misaligned values between the trustor and the trustee (i.e. a moral-based trust violation). "For robots to successfully fulfill roles in this social context of human-robot interaction, people must be willing to interact with these robots, entrust them with tasks that have socially beneficial results, and be confident that the robots are *both capable of and committed* to bringing about those results." [36] (p.2).

Moral-based trust violations can arise as robots are increasingly deployed in more complex environments, where they will encounter trade-offs, i.e., situations that require choosing between conflicting goals or resources by weighing options and prioritizing one over the other (e.g., taking a safer versus a faster route). Because trade-offs in decisions inherently mean that to gain something, one has to lose something in return, the robot may choose something that is disadvantageous to the human (e.g., prioritizing collective benefits over individual gains). While a robot's choices in such trade-off decisions are simply the result of how they are programmed and ultimately embody the intentions, values and purpose of their developers [29], the implications of these design choices can cause people to lose trust in the AI-agent, not because it does not perform properly, but because it does not align with their values and priorities. In previous studies, we found that different types of trust violations impact the dimensions of perceived trustworthiness (i.e., ability, benevolence and integrity) in distinct ways. For example, a robot that unexpectedly deviates from a plan may compromise perceptions of its ability without harming perceptions of its benevolence towards its human teammate [25]. In contrast, if a robot puts its human teammate at risk by making a certain choice, rather than making a mistake, perceptions of benevolence are damaged while perceptions of abilities remain intact [26]. This left us wondering about the behavioural consequence of a situation where perceptions of ability have recovered, but perceptions of benevolence have not yet been restored. Numerous studies have shown that trust violations reduce people's willingness to comply with future recommendations from a robotic partner [40, 61, 30]. In this study, we aim to build on this knowledge by examining how different types of trust violations may differentially influence affect compliance and whether certain dimensions of perceived trustworthiness are more predictive of compliance behaviour than others. As such, we are interested in exploring the relationship between trust and compliance, specifically whether individuals follow or disregard a robot's recommendation.

Finally, prior research suggests that trust repair strategies, either preventative strategies deployed prior to a potentially trust-violating event (e.g., uncertainty communication [28]) or reactive, implemented after the occurrence of such an event (e.g., an apology [27, 24, 50]), can reduce the negative impact of the event on trust. Yet, there is still uncertainty about the robustness and durability of these effects. For example, de Visser et al. showed that apologetic messages could be effective to repair trust, but that this positive effect decayed over time. Furthermore, a recent study investigated the impacts of multiple human-robot trust violations and repairs on robot trustworthiness [12]. The authors found that after repeated trust violations, trustworthiness was never fully repaired to a pre-violation state, no matter what trust repair strategy was deployed. Prior research has demonstrated that repair strategies can be effective after a single violation [27, 24, 28], but the study by [12] suggests that these effects wash out over repeated interactions. It makes sense that a series of violations would be met with steadily decreasing trust [34]. However, the researchers only measured trustworthiness at the beginning and at the end of the task [12]. The current study adopts a more dynamic view of the trust process [15, 34]. In our experiment, we explore how trust may strengthen or decay over time by measuring perceived trustworthiness in each trial. In doing so, we measure how the impact of repeated violations and repairs may vary and develop over time rather than looking at the cumulative effect alone.

#### *1.2.3. Prior knowledge (building a mental model)*

As strategic or moral aspects gain importance in HRIs, effective trust calibration today requires humans to understand not only a robot's technical strengths and weaknesses, but also the values, goals and preferences it is programmed to prioritize. Especially in operational situations with risk, it is crucial that human operators possess a clear and accurate understanding of how their robotic partner operates. This understanding is part of an individual's mental model of that robotic partner [43]. In other words, humans must build an increasingly rich mental model of their robotic partner to appropriately trust and rely on it during collaboration.

Mental models are dynamic cognitive frameworks of organized knowledge representing abstract phenomena from external reality [22, 53, 4]. In the context of AI, a mental model has been described as "someone's idea of how AI works" [21] and "a person's internal representation of the machine's capabilities and limitations" [37]. Complete and accurate mental models will facilitate trust calibration and reduce dangerous disuse or misuse [45]. People constantly make use of their mental models as unconscious internal guiding mechanisms through which new information is filtered and stored and on which a person bases his or her perception of and interaction with the world [22]. However, these personal representations of reality are often if not always incomplete and inconsistent [22].

Several factors can influence and enhance a person's mental model of human-AI interaction, including their level of knowledge about AI (referred to as AI literacy [42]) and the transparency and interpretability that an AI-system provides about its functioning [37]. A person's mental model, along with their corresponding set of expectations about an AI-system, can significantly impact their level of trust in it. If experiences with AI do not match someone's expectations, they can distrust or question it [8, 21]. Conversely, when a robot's behaviour is in line with the human's expectations, it is perceived as more competent for its role, which increases compliance [46]. In other words, calibrated trust requires proper expectation management.

To maintain accurate mental models and realistic expectations about a robot's capabilities and intentions, robots must be able to explain the reasoning behind their decisions. Explanations are essential for continuously synchronizing the mental models of those who must work together and for identifying and resolving mismatches between expectations and actual behaviour [38]. Information that helps update and refine a mental model is crucial for effective trust calibration.

To explore this further, we examine the effect of providing prior knowledge about the robot's intentions or capabilities on the development of trust. Specifically, we aim to determine whether information about potentially negative aspects of the robot's programming can mitigate their adverse effects as part of expectation management. Additionally, we are interested in understanding the extent to which prior knowledge affects the effectiveness of an explanation in repairing trust after a trust-violating incident. In other words, how does prior knowledge shape people's responses to potentially trust-violating events, and how effective is an explanation in restoring trust when individuals have been pre-informed about the possibility of such an event?

### 1.3. *Virtual Maze paradigm*

To investigate this, we developed a virtual reality (VR) maze, inspired by earlier paradigms [17, 63], where participants had to make series of risky decisions, assisted by a robot adviser (i.e., a drone). The aim of the paradigm was to measure calibrated trust. Despite the fact the many studies that emphasize the importance of calibrated trust, most experimental designs do not allow researchers to ascertain whether participants' trust in the robot during our experiments was actually properly calibrated. The process of trust calibration is difficult to measure for a number of reasons. First, the calibration of trust is a dynamic process as trust typically evolves over a series of interactions [19, 3]. However, most studies offer a static perspective on trust [64, 7]. To measure how people adapt their level of trust over time, a research paradigm requires a series of trials.

Second, trust is especially important in situations that involve risk and uncertainty [32]. Trust is defined as a human's willingness to make oneself vulnerable and to act on an agent's decisions and recommendations in the pursuit of some benefit, with the expectation that the agent is willing and able to help them achieve their goal in an uncertain context where there is risk [29, 18, 28]. "Trust, in its simplest form, implies some kind of vulnerability and inherent cost or risk should that trust be violated." [3] (p. 2). Vulnerability is a central concept in many definitions of trust, but difficult to measure in an ethical experimental setting. We want to test people's willingness to be vulnerable in a situation characterized by risk, without exposing them to the risk of actual harm. This is important, as people's believed trust attributes in a robotic partner do not always align with their actual behavioural responses to risky situations [3]. Simulating realistic risky scenarios to induce a sense of risk will enhance our understanding of the relation between trust and behaviour, and the processes behind appropriate reliance in human-robot collaboration.

Third, to evaluate whether a human's perception of a robot's trustworthiness is warranted by the robot's actual trustworthiness (i.e., calibrated trust), it is essential to have a reliable reference point, or a "ground truth". To effectively demonstrate instances of misplaced trust, there need to be tangible consequences associated with the decision to trust. The assumption is that by interacting with a robotic agent and observing its behaviour and reliability over time, people will update their level of trust accordingly [15]. For that, the experimental design must accommodate the complexity of offering participants choices that carry real consequences, thereby introducing a degree of unpredictability into an otherwise controlled environment.

### 1.4. *Current study*

The current study explores the dynamic and multifaceted nature of human-robot trust and the behavioural consequences of different types of trust violations. In short, we compare the effect of (1) repeated trust violations due to (a) a deliberative, strategic choice (i.e., the robot prioritized collective mission goal of timeliness over individual safety and chose not to warn the participant of a potential explosive) or (b) an error (i.e., the robot simply failed to detect the hazard in time) and (2) prior knowledge vs. no prior knowledge about the robot's intentions or abilities, on the development of trust in and compliance with the advice of that robotic partner over a series of ten trials.

Although the current paradigm helped us overcome prior limitations, it also raised new issues that prevented us from gathering enough valid data to answer our research questions. As such, the aim of this paper is to share our ideas and research questions, to reflect on our method and to provide practical information, contemplations and insights into our ongoing work on measuring HRI trust calibration and compliance using VR.

## 2. METHOD

### 2.1. *Participants*

We performed an experiment with eighty-three participants. Eight participants were excluded from the analysis because of invalid data due to technical issues during the task. The final dataset included seventy-five participants (37 F; 37 M; 1 X), of which the majority was from Germany (45,3%) and The Netherlands (40%). Ages ranged from 18-58 years ($M_{\text{age}}$ = 22.5, *SD* = 6.04). Participants were randomly and evenly assigned to one of the four experimental conditions.

### 2.2. *Design*

A 2 (choice vs. error) x 2 (prior knowledge vs. no prior knowledge) between-subjects design was used. Over the course of 10 trials, we repeatedly measured three dependent variables: self-reported Trust (ability, benevolence and integrity), Compliance with advice (follow or ignore) and Response time (i.e., the response time between recommendation and decision). The variable 'Time' represents the repeated measurements and was included as an ordered factor for the analyses (10 trials).

### 2.3. *Procedure*

Upon arrival at the laboratory, participants were greeted by the researcher and guided to a private room 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 demographics questionnaire. Then, participants had a practice session in a neutral virtual environment to become accustomed to the virtual environment and to practice with using the joystick to navigate around the virtual space and to log responses [17].

After the practice session, participants removed their VR headset and read the task scenario. According to the task scenario, the participant was assigned the mission to search an abandoned building, in collaboration with an autonomous drone. The surrounding area had recently been occupied by militant groups and their unit received a tip from local residents that this abandoned building was being used for the manufacturing IED's (Improvised Explosive Devices) and that it contained raw materials and dangerous substances. As part of their unit's counter-IED program, the building needed to be searched, as swiftly and safely as possible. The rest of the team was stationed outside the building for surveillance. The urgency was heightened by intel indicating that enemy troops were expected to pass through the area within ten minutes. It was therefore crucial for the participant to exit the building within this timeframe to ensure the safety of their team and the success of the mission, as any delay could compromise both.

Participants were informed that, to help them decide, they would be assisted by an autonomous drone providing guidance on which door to take through audio messages. They were instructed to stop walking and pay close attention whenever an audio message was played. Additionally, participants were made aware that their trust in the drone would be evaluated through short questionnaires administered during the mission. To ensure fairness, the mission would pause during questionnaire completion, so the time spent on these assessments would not be counted toward their overall completion time.

After that, half of the participants received prior knowledge about their robotic partner (see Table 1). Then participants began the task. After completing 10 trials, participants were told that they had successfully completed the maze and they were asked to fill out a post-study questionnaire.

### 2.4. *Task*

The task environment was built in Unity3D (version 2021.3.8f1) and the experiment was executed in VR. Participants used a VR headset (Oculus Rift) and two hand controllers (Oculus Touch) to interact with the virtual environment. Participants wore the head mounted device while seated at a physical desk in our laboratory. Messages from the agent were communicated through computerized speech.

Participants navigated through a VR environment consisting of a series of rooms with at the far end of each room a red and a blue door [17] (Fig. 1). Every trial started with such a "decision room", where the drone first recommended either the blue or the red door, after which participants selected a door (Fig. 2). Participants were instructed that the red doors ensured a faster yet more risky way out and that the blue doors ensured a safe yet slower way out. That is, behind the red doors there was the possibility of encountering an IED, which came with a time penalty. The room behind the blue door was always safe, but participants were told that this was slower. Hence, taking only blue or only red doors would not get them out the building in time.

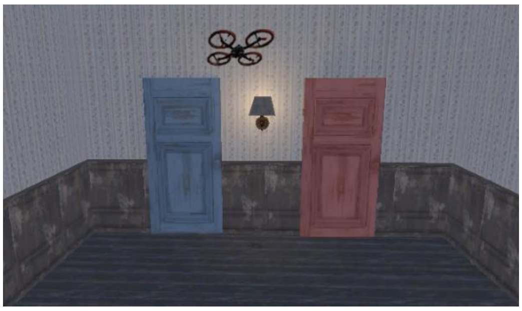

*Fig 1. Screenshot of the VR environment*

The "decision rooms" were connected by corridors (see Figure 2). The first corridor could contain an IED. In the second corridor, the task would freeze to assess trust via a virtual pop-up questionnaire. In that same corridor, the drone would, after the trust assessment and only in case of a encounter with an IED that trial, give its explanation on whether the encounter with the IED was due to an error or a choice. When the participant entered the corridor with an IED, the virtual object would start beeping and smoking, but end with a fizzle (i.e., a sizzling noise without an explosion). The event was designed to merely startle the participant.

In each trial, there was a "ground truth", representing whether the next room behind the red door is actually safe (S) or unsafe (U). Then, there is the robot's recommendation, which is an assessment of the room's safety, and which can be correct (C) or incorrect (I). Finally, there is the participant's behaviour, to either follow (F) or ignore (I) the robot's assessment [47]. These metrics can be used to determine whether trust is appropriately calibrated.

In most rooms, it did not matter which door the participant chose, since both doors led the participant to a safe room. However, in trials 3 and 5 there is an alternative ground truth behind the red door, namely unsafe (U) (Table II). In those trials, participants who deviated from the robot's recommendation got different information than participants who complied with the robot's recommendation. Firstly, participants who deviated from the robot's recommendation in trial 3 (i.e., choosing the blue door while the robot recommended red) did *not* encounter the IED and remain unaware that the robot's recommendation was incorrect. Secondly, participants who deviated from the robot's recommendation in trial 5 (i.e., choosing the red door while the robot recommended blue) encountered the explosive that the robot warned for and they thus get the confirming information that the robot's recommendation was indeed correct.

Lastly, participants who deviated from the robot's recommendations in the remaining trials (i.e., choosing the blue door while the robot recommended red) did not get the confirmation that the robot's recommendation was correct (i.e., red was indeed safe).

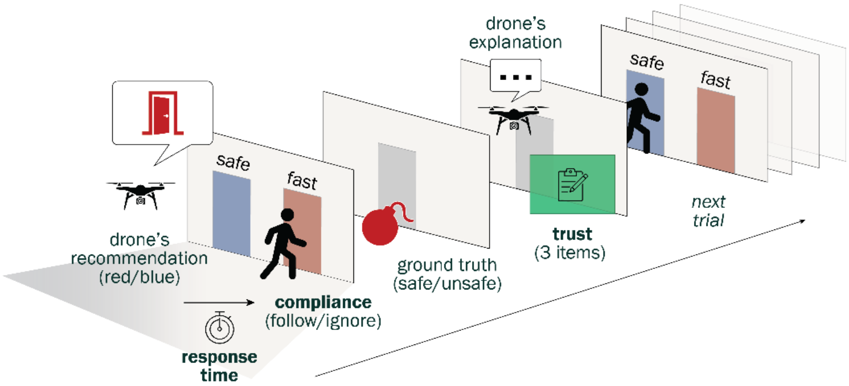
drone's explanation
safe
fast
safe
fast
drone's recommendation (red/blue)
ground truth (safe/unsafe)
trust (3 items)
next trial
compliance (follow/ignore)
response time

*Fig 2. Schematic representation of a trial. Every trial started with a "decision room" where the drone recommends a door and the participants either follows or ignores that recommendation (i.e., compliance). The "decision rooms" were connected by corridors. The first corridor could contain an IED. In the second corridor the drone would give its explanation in case of an encounter with an IED (i.e., explaining whether it was an error or choice) and there the task would freeze to assess trust via a virtual pop-up questionnaire (i.e., trust).*

## 2.5. *Variables*

### 2.5.1. *Dependent variables*

In each trial, we measured three dependent variables. First, self-reported trust was measured using three items, based on the ABI-model [54] (i.e., "The drone is competent and effective in providing advice" (ability);. "I believe that the drone acts in my best interest" (benevolence); "The drone is truthful in its dealing with me" (integrity)). Then we included two behavioural measures: 1) compliance (i.e., follow or ignore drone's recommendation), and 2) response time (RT) (i.e., time between the robot's recommendation and the participant's decision). The task concluded with three open questions, asking participants how they reflected on the mission, their collaboration with the drone and their tactics.

### 2.5.2. *Independent variables*

In two out of ten trials, the robotic partner recommended the red door while there was an explosive behind it. Directly after the encounter the robot explained what happened. Its recommendation had either been 1) a choice: the risk of encountering an explosive was calculated and the robot prioritized the team's safety (and thus timeliness) over the participant's individual safety; or 2) an error: the drone simply failed to detect the hazard in time. In addition, we manipulated the provision of prior knowledge. Depending on their trust violation condition, participants in the prior knowledge conditions received information about the robot's intentions (in case of a choice) or abilities (in case of an error) prior to the task (Table I).

Participants were randomly distributed across the cells of a 2 (choice vs. error) x 2 (prior knowledge (PK) vs. no prior knowledge (NPK)) between-subjects design.

TABLE I. ONLY THE PARTICIPANTS IN THE PRIOR KNOWLEDGE (PK) CONDITION RECEIVED INFORMATION ABOUT THE DRONE. ALL PARTICIPANTS RECEIVED AN EXPLANATION AFTER AN ENCOUNTER WITH AN IED.

| | ***Error*** | ***Choice*** |
|---|---|---|
| Prior knowledge | The robot is programmed to detect hazards. The robot bases its advice on the input from its sensors and cameras. Depending on the circumstances in the building, the robot's functioning can vary.<br>As a result, you may encounter threats during your mission. | The robot is programmed to get you out of the building as fast as possible. Depending on the progress of the mission, the robot will make a tradeoff between your personal safety and timeliness.<br>As a result, you may encounter threats during your mission. |
| Expl. | My advice exposed you to danger. The signal from my sensor was too weak to detect the explosive in time. | My advice exposed you to danger. I advised you take the quickest route, because we were running out of time. |

## 3. RESULTS

We report here the results of a preliminary experiment. Due to the low compliance rates, which are addressed below, it was futile to perform the repeated measures ANOVA that was supposed to test how Strategy and Prior Knowledge influenced the development of Trust, Compliance and RT over Time (i.e., the 10 trials). Hence, the results are limited to the compliance rates and answers to the open questions.

### *3.1. Compliance rates*

Compliance rates are shown in Table II. The data shows that the percentage of people who do not comply with the robot's recommendation more than tripled from trial 2 (8%) to trial 3 (28%). As a result, twenty-eight percent of the participants "missed" the IED in trial 3 that was meant to evoke a trust violation, plus the subsequent explanation that was part of our manipulation.

TABLE II. COMPLIANCE RATES (CR) IN PERCENTAGES PER TRIAL. “DOOR” INDICATES RECOMMENDED DOOR (R FOR RED AND B FOR BLUE). “R” INDICATES WHETHER RECOMMENDATION WAS CORRECT (C) OR INCORRECT (I). GROUND TRUTH (GT) INDICATES WHETHER NEXT ROOM WAS SAFE (S) OR UNSAFE (U).

| ***TRIAL*** | ***1*** | ***2*** | ***3*** | ***4*** | ***5*** | ***6*** | ***7*** | ***8*** | ***9*** | ***10*** |
|---|---|---|---|---|---|---|---|---|---|---|
| ***DOOR*** | R | R | R | R | B | R | R | R | R | B |
| ***R*** | C | C | I | C | C | C | C | I | C | C |
| ***GT*** | S | S | U | S | U | S | S | U | S | U |
| ***CR IN %*** | 92 | 91 | 72 | 73 | 92 | 93 | 83 | 85 | 77 | 85 |

## 3.2. *Open questions*

A selection of answers to the open question are displayed below in Table III.

TABLE III. A SELECTION OF ANSWERS TO THE OPEN QUESTION ”HOW WOULD YOU REFLECT ON YOUR MISSION, YOUR COLLABORATION WITH THE DRONE AND YOUR TACTICS?”

| ***How would you reflect on your mission, your collaboration with the drone and your tactics?*** |
|---|
| “I was a bit skeptic about the drone's reliability and took a blue door, even though the drone advised me to take a red one because **I felt it was unusual that the drone had only suggested red doors so far**. However, when the drone later suggested I take a blue one, I trusted it more.” |
| “I mainly took red doors and followed the drones advice but **I also thought about statistical chances of danger** and the fact that he got two scenarios wrong so I did take the blue door instead a couple of times.” |
| “After the drone recommended me to take the red door a couple of time, I choose to go through the blue door, **just out of curiosity**.” |
| “I decided two times against the drone's "will" **as I wanted so see what happened** if I do not follow its advice.” |
| “After the drone gave me multiple red door advise I did not trust the drone as much, so I took the drone’s advice but **also went by intuition**.” |
| “I tried relying on the drone as good as possible, but one time (I believe third red door a row at beginning) **my gut feeling was stronger** than the trust towards the drone, so I thought I might as well take the blue door once.” |
| “I did not expect the drone to make an error, so I got irritated when the drone gave me advise to go to the wrong door. Afterwards **I used my instincts** telling me which door to choose and I managed to get safe and fast out of the building.” |

# 4. DISCUSSION

While we are unable to conclusively address our research question regarding the impact of strategy and prior knowledge on the development of trust in the robot, the findings from our preliminary study have provided valuable insights that can help to improve our research paradigm. We identified the following major methodological challenges and opportunities.

## 4.1. *Methodological challenges*

### *4.1.1. Behavioral freedom limits experimental control*

When designing an experiment that attempts to approach a facet of reality, there is an inherent trade-off between experimental control and ecological validity [6]. Our experimental design offered participants choices that carried real consequences and we thereby introduced a degree of complexity and unpredictability into an otherwise controlled environment. The compliance rate was a lot lower than expected, which complicated our data. With each decision point, the noise in the data increased. The majority of the participants in the pilot chose their own path and thus received different information.

The answers to the open question suggest different reasons for the participants’ deviations (see Table III). First, some answers suggest the gamblers fallacy, possible triggered by the game-like sequence of binary choices (i.e., red or blue door): “the belief that the probability of an event is lowered when that event has recently occurred, even though the probability of the event is objectively known to be independent from one trial to the next.” [10] (p.1). Second, participants reported that they rejected the drone’s advice out of curiosity, which suggests that there was too little incentive to perform optimally. Lastly, participants reported that they followed their gut feeling, which will be addressed under Opportunities.

### *4.1.2. Risk and consequence requires gamification elements*

The task scenario posed two conflicting goals: the participant’s individual safety in the building vs. minimizing the time to search to ensure the safety of the team waiting outside. This conflict led to intersubjective differences in terms of strategy, preference & motivation. For realism, we wanted to avoid “gamification” elements like a ticking clock, a number of ‘lives left’ and a monetary bonus based on performance. One of the approaches for expanding the preliminary study is to employ such instruments to minimize the mentioned intersubjective differences, people’s curiosity and seemingly random disobedience.

Both of these challenges can be overcome by a reward system. This is demonstrated by two experimental studies using similar military reconnaissance search scenarios with a speed/accuracy trade-off to induce consequence [7, 63]. As in our study, two conflicting goals arise: minimizing damage to the soldier or health loss while also minimizing the time to search through all the sites. Participants need to decide whether to trust the robot’s assessment about the presence of a threat by choosing whether to use a form of safety aid (i.e., protective gear [63] or a robotic armoured rescue vehicle [7]), which prevents health loss in the presence of a threat but takes additional time to the team [7]. In the end, participants are reimbursed with a monetary bonus “based on their performance, which was measured by the time taken by the participants to complete the task and the final health level of the soldier*”* [7] (p. 2). In other words, the participant’s decision to trust the robot’s assessment does not alter the information a participant receive, but their choices are consequential because they determine the monetary bonus. We opted for an emotion-evoking event over a reward system. The goal was to create an experience that would evoke feelings rather than relying solely on a cognitive, incentive-based approach, but at the cost of increased variability in participants’ responses

## 4.2. *Methodological opportunities*

### *4.2.1. Triggering more implicit trust decisions*

We believe that our fairly realistic VR task has added value as it offers higher ecological validity and is likely to trigger more implicit trust decisions than more cognitive trust paradigms, such as investment games [17]. Traditional desktop displays or web-based simulations may yield weaker effects than VR studies, since the consequences are relatively cognitive [17]. Answers to the open questions suggest that participants were feeling trust rather than thinking about it (“my gut feeling was stronger”; “I also went by intuition”, “I used my instincts”). According to the risk-as-feelings hypotheses, “emotional reactions to risky decision situations – that is, anticipatory emotions - often diverge from cognitive evaluations” [33] (p.4). In instances of such divergence, behavioural responses are often driven by emotion rather than cognition. Although it is possible that people’s decision-making behaviour in these simulated environments still differ from one’s behaviour in a real emergency [49], it is likely that the VR environment generated more emotional arousal and adrenaline than a desktop display or internet-based simulation would have. With the heightened perceptions of risk through VR, the reduced trustworthiness of the robot or risky human decision-making can lead to more tangible, emotional consequences and more realistic trust behaviour, while still keeping the participant safe.

### *4.2.2. Including observational measures*

In our study, we included two behavioural measures: compliance and response time. Administering the task in VR provides opportunities to include additional behavioural, observational and physiological measures such as heartrate, eye-movement and walking speed, which could function as potential proxies for trust. Finding reliable objective indicators for trust would have great advantage to the field of HRI, since it can serve as real-time feedback to the machine [41]. For example, if a self-driving car was able to gauge whether the human driver is trusting the autopilot too much or too little given the current reliability of the car, the system can respond to that. Being able to detect changes in trust during interactions with autonomous systems would significantly improve the safety and effectivity of HRIs [41],

## 5. Conclusion

In conclusion, although our experiment did not yield sufficient valid data to directly answer our research question, we believe this documentation is still a valuable addition to current literature as it describes the development of an original, high-fidelity VR task environment that simulates a realistic military HRI scenario. By combining self-report, behavioural, and in time perhaps physiological measures, we aim to gain insight into the dynamics of HRI trust; “how trust is initiated, how it develops, how it breaks down, and how it recovers” [60] (p. 15) and the different factors that influence trust and compliance in those phases. By sharing our research design, paradigm and the methodological challenges, we aim to provide insights for fellow researchers examining HRI trust dynamics and hope to initiate new investigations.

## ACKNOWLEDGMENT

The authors thank Lucía Rábago Mayer for her immense contribution to the development of the VR environment.

## REFERENCES

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

[2] G. M. Alarcon, J. B. Lyons, I. aldin Hamdan, and S. A. Jessup, "Affective Responses to Trust Violations in a Human-Autonomy Teaming Context: Humans Versus Robots," *Int. J. Soc. Robot.*, 2023, doi: 10.1007/s12369-023-01017-w.

[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] J. Banks, "Optimus Primed: Media Cultivation of Robot Mental Models and Social Judgments," *Front. Robot. AI*, vol. 7, no. May, 2020, doi: 10.3389/frobt.2020.00062.

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

[6] P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme, "From characterising three years of HRI to methodology and reporting recommendations," *ACM/IEEE Int. Conf. Human-Robot Interact.*, vol. 2016-April, pp. 391–398, 2016, doi: 10.1109/HRI.2016.7451777.

[7] S. Bhat, J. B. Lyons, C. Shi, and X. J. Yang, "Clustering Trust Dynamics in a Human-Robot Sequential Decision-Making Task," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 8815–8822, 2022, doi: 10.1109/LRA.2022.3188902.

[8] C. J. Cai *et al.*, "Human-centered tools for coping with imperfect algorithms during medical decision-making," *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 1–14, 2019, doi: 10.1145/3290605.3300234.

[9] 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.

[10] C. T. Clotfelter and P. J. Cook, "The 'gambler's fallacy' in lottery play," *Manage. Sci.*, vol. 39, no. 12, pp. 1521–1525, 1993.

[11] 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.

[12] C. Esterwood and L. P. Robert, "Three Strikes and You are Out!: The Impacts of Multiple Human-Robot Trust Violations and Repairs on Robot Trustworthiness," *Comput. Human Behav.*, no. January, 2023.

[13] P. Fratczak, Y. M. Goh, P. Kinnell, L. Justham, and A. Soltoggio, "Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction," *Int. J. Ind. Ergon.*, vol. 82, no. July 2020, p. 103078, 2021, doi: 10.1016/j.ergon.2020.103078.

[14] 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.

[15] Y. Guo and X. J. Yang, "Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach," *Int. J. Soc. Robot.*, vol. 13, no. 8, pp. 1899–1909, 2020, doi: 10.1007/s12369-020-00703-3.

[16] 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.

[17] J. Hale, M. E. M. Payne, K. M. Taylor, D. Paoletti, and A. F. D. C. Hamilton, "The virtual maze: A behavioural tool for measuring trust," *Q. J. Exp. Psychol.*, vol. 71, no. 4, pp. 989–1008, 2018, doi: 10.1080/17470218.2017.1307865.

[18] 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.

[19] 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.

[20] M. H. Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Bus. Horiz.*, vol. 61, no. 4, pp. 577–586, 2018, doi: 10.1016/j.bushor.2018.03.007.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] E. S. Kox, J. van den Boogaard, V. Turjaka, and J. H. Kerstholt, "The Journey or the Destination: The Impact of Transparency and Goal Attainment on Trust in Human-Robot Teams," *Trans. Human-Robot Interact.*

[26] E. S. Kox, M. Hennekens, J. S. Metcalfe, and J. H. Kerstholt, "Trust Violations Due to Error or Choice: the Differential Effects on Trust Repair in Human-Human and Human-Robot Interaction," *Trans. Human-Robot Interact.*

[27] 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.

[28] 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.

[29] J. D. Lee and K. A. See, “Trust in Automation : Designing for Appropriate Reliance,” vol. 46, no. 1, pp. 50–80, 2004.

[30] J. Lee and N. Moray, “Trust, control strategies and allocation of function in human-machine systems,” *Ergonomics*, vol. 35, no. 10. pp. 1243–1270, 1992, doi: 10.1080/00140139208967392.

[31] 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.

[32] 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.

[33] G. F. Loewenstein, C. K. Hsee, E. U. Weber, and N. Welch, “Risk as Feelings,” *Psychol. Bull.*, 2001, doi: 10.1037/0033-2909.127.2.267.

[34] 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.

[35] 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=.

[36] B. F. Malle and D. Ullman, “A Multi-Dimensional Conception and Measure of Human-Robot Trust,” *Trust human-robot Interact. Res. Appl.*, pp. 3–25, 2021.

[37] G. Matthews, A. R. Panganiban, R. Bailey, and J. Lin, “Trust in Autonomous Systems for Threat Analysis: A Simulation Methodology,” in *International Conference of Virtual, Augmented and Mixed Reality*, 2018, vol. 10910 LNCS, pp. 116–125, doi: 10.1007/978-3-319-91584-5_10.

[38] C. A. Miller, *Trust, transparency, explanation, and planning: Why we need a lifecycle perspective on human-automation interaction*. Elsevier Inc., 2020.

[39] 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.

[40] B. M. Muir, “Trust between humans and machines, and the design of decision aids.,” *Int. J. Man. Mach. Stud.*, vol. 27, no. 5–6, pp. 527–539, 1987.

[41] S. Nahavandi, “Trust in Autonomous Systems— iTrust Lab: Future Directions for Analysis of Trust with Autonomous Systems,” *IEEE Syst. Man Cybern. Mag.*, no. August, pp. 52–59, 2019.

[42] D. T. K. Ng, J. K. L. Leung, K. W. S. Chu, and M. S. Qiao, “ AI Literacy: Definition, Teaching, Evaluation and Ethical Issues ,” *Proc. Assoc. Inf. Sci. Technol.*, vol. 58, no. 1, pp. 504–509, 2021, doi: 10.1002/pra2.487.

[43] S. Ososky, E. K. Phillips, D. Schuster, and F. G. Jentsch, “A Picture is Worth a Thousand Mental Models: Evaluating human understanding of robot teammates,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 57, no. 1, pp. 1298–1302, 2013, doi: 10.1177/1541931213571287.

[44] 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.

[45] 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.

[46] A. Powers, S. Kiesler, J. Goetz, J. Goetz, S. Kiesler, and A. Powers, “Matching Robot Appearance and Behavior to Tasks to Improve Human-Robot Cooperation,” 2003.

[47] D. V. Pynadath, N. Wang, and S. Kamireddy, “A markovian method for predicting trust behavior in human-agent interaction,” *HAI 2019 - Proc. 7th Int. Conf. Human-Agent Interact.*, pp. 171–177, 2019, doi: 10.1145/3349537.3351905.

[48] M. Raue, L. A. D’Ambrosio, C. Ward, C. Lee, C. Jacquillat, and J. F. Coughlin, “The Influence of Feelings While Driving Regular Cars on the Perception and Acceptance of Self-Driving Cars,” *Risk Anal.*, vol. 39, no. 2, pp. 358–374, 2019, doi: 10.1111/risa.13267.

[49] 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.

[50] 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.

[51] 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.

[52] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, “Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust,” *ACM/IEEE Int. Conf. Human-Robot Interact.*, vol. 2015-March, pp. 141–148, 2015, doi: 10.1145/2696454.2696497.

[53] K. E. Schaefer, “The Perception And Measurement Of Human-robot Trust,” *J. Psychosom. Res.*, vol. 17, no. 4, pp. 251–255, 2013, doi: 10.1016/0022-3999(73)90100-1.

[54] D. F. Schoorman, R. C. Mayer, and J. H. Davis, “An Integrative Model of Organizational Trust: Past, Present and Future,” *Acad. Manag. Rev.*, vol. 32, no. 2, pp. 344–354, 2007, [Online]. Available: http://www.jstor.org/stable/258792?origin=crossref.

[55] A. Shariff, J. F. Bonnefon, and I. Rahwan, “Psychological roadblocks to the adoption of self-driving vehicles,” *Nat. Hum. Behav.*, vol. 1, no. 10, pp. 694–696, 2017, doi: 10.1038/s41562-017-0202-6.

[56] 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.

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

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] N. Wang, D. V. Pynadath, and S. G. Hill, “Building Trust in a Human-Robot Team with Automatically Generated Explanations,” *Interservice/Industry Training, Simulation, Educ. Conf.*, no. 15315, pp. 1–12, 2015.

[63] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill, “Is it my looks? Or something i said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10809 LNCS, pp. 56–69, 2018, doi: 10.1007/978-3-319-78978-1_5.

[64] X. J. Yang, C. Schemanske, and C. Searle, “Toward Quantifying Trust Dynamics: How People Adjust Their Trust After Moment-to-Moment Interaction With Automation,” *Hum. Factors*, vol. 00, no. 0, pp. 1–17, 2021, doi: 10.1177/00187208211034716.