The Journey or the Destination: The Impact of Transparency and Goal Attainment on Trust in Human-Robot Teams

ESTHER S. KOX, Human-Machine Teaming, TNO, Soesterberg, The Netherlands and Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands
JUUL VAN DEN BOOGAARD and VESA TURJAKA, Applied Cognitive Psychology, Utrecht University, Utrecht, The Netherlands
JOSÉ H. KERSTHOLT, Human Behaviour and Training, TNO, Soesterberg, The Netherlands and Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands

As robots gain autonomy, human-robot task delegation can become more goal-oriented; specifying what to do rather than how. This can lead to unexpected robot behaviour. We investigated the effect of transparency and outcome on the perceived trustworthiness of a robot that deviates from the expected manner to reach a delegated goal. Participants ($N = 82$) engaged in a virtual military mission as a Human-Robot Team using a 2×2 between-subjects design (low vs. high transparency, positive vs. negative outcome). Participants received training on the expected manner to reach the mission's goal. In the actual mission, the robot deviated from the planned path. We manipulated whether the robot explained its deviation and whether the outcome was better or worse than the original plan. Results showed that transparency contributed to higher and more stable levels of trust, without increasing subjective workload. While the robot's deviation led to a violation of trust in the low transparency condition, trust remained stable in the high transparency condition, indicating a buffering effect of transparency on trust in case of unexpected behaviour. The impact of outcome on trust was consistent across transparency conditions. Our findings underscore the role of transparency as a tool for fostering human-robot trust.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Laboratory experiments*; **User studies**; **Graphical user interfaces**; **Auditory feedback**; *Empirical studies in collaborative and social computing*; • **Applied computing** → **Psychology**; • **Computer systems organization** → *Robotics*; • **General and reference** → *Experimentation*;

Additional Key Words and Phrases: Human-Autonomy Teaming, Trust, Delegation; Transparency, Outcome

ACM Reference format:
Esther S. Kox, Juul van den Boogaard, Vesa Turjaka, and José H. Kerstholt. 2024. The Journey or the Destination: The Impact of Transparency and Goal Attainment on Trust in Human-Robot Teams. *ACM Trans. Hum.-Robot Interact.* 14, 2, Article 23 (December 2024), 23 pages.
https://doi.org/10.1145/3702245

Authors' Contact Information: Esther S. Kox (corresponding author), Human-Machine Teaming, TNO, Soesterberg, The Netherlands and Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands; e-mail: esther.kox@tno.nl; Juul van den Boogaard, Applied Cognitive Psychology, Utrecht University, Utrecht, The Netherlands; e-mail: juulvandenboogaard@me.com; Vesa Turjaka, Applied Cognitive Psychology, Utrecht University, Utrecht, The Netherlands; e-mail: turjakavesa@gmail.com; José H. Kerstholt, Human Behaviour and Training, TNO, Soesterberg, The Netherlands and Psychology of Conflict, Risk & Safety, University of Twente, Enschede, The Netherlands; e-mail: jose.kerstholt@tno.nl.

ACM 2573-9522/2024/12-ART23
https://doi.org/10.1145/3702245

1 Introduction

Due to recent technological developments in artificial intelligence and robotics, more and more people are increasingly interacting with artificial agents in a variety of domains, among which the military [51, 98]. As robots become more intelligent, they are increasingly self-governing, gain decision authority within their functioning [6, 28, 35, 62, 84], and require less human involvement and control [48, 57]. 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 [86]. As such, robots transition from relatively simple tools towards autonomously acting agents, which greatly impacts the trust relationship between humans and such technology [35, 51, 99].

It is expected that cooperation in teams where at least one human works together with one robot, so called **Human-Robot Teams (HRTs)**, will increase in the (near) future [6, 62, 93, 94]. Future robots are envisioned to have the ability to observe and act upon an environment autonomously and to communicate and collaborate with other agents, including humans, to solve problems and achieve (common) goals [20, 39, 40, 62, 98]. In HRTs, robots can work interdependently with human team members towards a shared objective [62]. Robots can take over tasks that were previously conducted by humans, whereas other tasks still need to be executed by human counterparts [67]. As a result, the rise of HRTs poses interesting challenges related to teamwork, task delegation and trust.

1.1 Delegation

Teamwork typically involves dividing and assigning tasks or responsibilities to different team members. When delegating authority, an actor (i.e., in our HRT case, the human) hands over a specific (set of) task(s) to another actor (i.e., the robot) who is expected to take responsibility for planning and execution of the assignment in a timely and effective manner to reach commonly understood goals [32, 57, 60]. Since reaching a goal consists of completing a set of tasks, delegation is inherently hierarchical [57]. As a result, delegation can be adapted to different levels of abstraction, such as (1) skill-based delegation, which proceeds by delegating single elementary tasks or actions (e.g., go-right, go-left), (2) rule-based delegation, which proceeds by delegating in terms of pre-defined templates of taskwork and teamwork (e.g., perform-blanket-search procedure) and ultimately, (3) goal-oriented delegation, which proceeds by delegating in terms of goals [8, 56, 60]. Which type of delegation is appropriate will depend on a robot's **level of autonomy (LOA)**, which can range from no autonomy (i.e., manual human control), to semi-autonomy (i.e., human can veto) to full autonomy (i.e., human is at most informed) [16, 65].

The more autonomous a robot gets, the more abstract and goal-oriented a delegated assignment can be, the more degrees of freedom the robot has in terms of execution and the more trust in the robot is required. Goal-oriented task delegation implies that the delegator does not have to outline the specific rules and skills that should be used in the process of reaching the desired end-state. In short: it means telling the robot what to do instead of how to do it. This leaves considerable room for the robot to fill in the remaining details on the execution of desired actions, which allows it to adapt to changing environments and operational demands [56]. As a situation evolves, the possible paths to achieve a certain goal can change [32]. As a result, an (semi-)autonomous robot might exhibit unexpected behaviour—from the perspective of a human operator—in its pursuit to reach a certain goal. A possible risk is that a human's lack of understanding of the robot's actions can cause people to lose trust and want to take over manual control, negating the advantages of task delegation. Regardless of the LOA of a robot, communication and human participation in certain decision-making loops will always remain crucial for effective and safe operations [1].

To keep the human involved, robots will need to be able to explain their behavioural choices, especially when they deviate from the expected manner to reach a goal. Higher decision authority

assigned to robots typically increases the human desire to know what the robot will be doing [10]. When the human operator cannot understand the basis of the robot's assessments and actions, trust may be eroded, especially when the robot's actions do not align with the human's expectations [46, 64]. In the current study, we are interested in the implications of a robot that has been delegated the authority to select the best course of action given the local situation, which could contradict a human's expectation and result in a suboptimal outcome (i.e., not attaining the goal). In the context of goal-oriented delegation, does understanding the robot's actions towards a goal drive trust or is ultimately attaining the goal the primary factor?

1.2 Trust

Teamwork requires task delegation and task delegation requires trust. 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 will help achieve their common goal in an uncertain context involving risk [23, 34, 43, 50, 70, 83]. During collaborations, human-robot trust is continuously adjusted with the goal of finding an appropriate level where the perceived trustworthiness of a robot align with its actual reliability; a process known as trust calibration [43]. A human operator's trust should be calibrated to reflect a robot's capabilities in order to achieve appropriate reliance [43, 97]. Goal-oriented delegation and higher LOA means less human involvement and control, which results in more uncertainty and thus a higher demand for appropriate levels of trust.

Well-calibrated trust enhances a team's effectiveness, whereas both "undertrust" (i.e., trusting too little) and "overtrust" (i.e., trusting too much) diminish it [16]. Undertrust can lead people to overly and unnecessarily monitor the robot after delegation, or even to refuse to interact with the robot altogether, thereby compromising profitability. Overtrust, on the other hand, could result in the robot having too much freedom, possibly compromising safety [94]. Calibrated trust is crucial to minimize the risks and to maximize the benefits in the highly interdependent and dynamic nature of teamwork [6, 43, 44].

In general, perceiving good robot functioning will likely increase perceived trustworthiness, whereas perceiving maladaptive (i.e., errors or mistakes) or ambiguous (i.e., unexpected or unpredictable) robot functioning often results in decreases in perceived trustworthiness—so called trust violations [17, 19, 39, 100]. As we strive for calibrated trust rather than maximum trust, decreases in perceived trustworthiness are a logical and functional adaptive response to perceiving errors, technical failures or other forms of reduced reliability and performance. However, with the anticipated advancements in the ability of robots to self-select courses of action, the range of possible causes of human-robot trust violations expands. That is, human-robot trust is not solely based on a robot's perceived abilities and performance (i.e., *what* it does and can do), but also on its perceived purpose and alignment with a trustor's values (i.e., *why* it was developed and operates in a certain way), as well as the understandability or interpretability of the robot and its ability to explain its actions (i.e., *how* it operates) [43, 45]. This operationalization of trust corresponds to the Ability (what), Benevolence (why) and Integrity (how) (ABI) model from Mayer et al. [52] and reflects how a trustee's trustworthiness is based on more than reliability and performance. As a consequence, trust violations are not solely caused by reduced performance.

As robots become more autonomous, task delegation can become more goal-oriented, providing the robot more with greater degrees of freedom in terms of execution. Hence, trust violations might be increasingly caused by a human operator's lack of understanding of the robot's assessments and actions, rather than poor robot performance. When a robot does something unexpectedly (according to the human), its efficacy and accuracy could be questioned and the action can lead to a decrease of human-robot trust, regardless of whether the robot is actually maladapted [71,

76]. For example, a drone might rightfully adapt its course of action to changes in the operational environment to reach a certain goal, such as avoiding a collision, without informing the human. If the drone's deviation significantly conflicts with the human's expectations and the robot lacks the ability to explain itself, the human operator might take over manual control because they do not understand the drone's actions and perceive them as inappropriate and untrustworthy [35, 48, 71]. As such, a lack of understanding causes a trust violation and leads to a situation of undertrust. Since the success of **human–robot interactions (HRIs)** greatly depends on people's ability to trust them, trust violations that lead to undertrust would make it necessary for a robot to engage in trust repair strategies [3].

Given (1) the inevitability of unexpected robot behaviour in HRI, (2) the possibility that unexpected behaviour results in trust violations and poor trust calibration, and (3) the disadvantageous consequences of poor trust calibrations, it is important to evaluate methods to prevent or buffer (unnecessary) trust violations as a consequence of unexpected behaviour. Most current HRI trust repair literature focuses on the role of trust repair strategies after an apparent error [7, 17, 22, 26, 38, 44, 61, 73, 75, 92, 95]. However, more recently researchers have started to evaluate trust violations as a result of unexpected behaviour rather than failure [48, 68, 80]. In essence, to prevent that trust will unjustly erode due to a misunderstanding of the basis of a robot's assessments and actions, robots will need to be able to explain the rationale behind their behavioural choices. Increasing transparency and interpretability through explanations can enhance trust calibration by lowering unrealistic expectations on the one hand (i.e., preventing overtrust) and by clarifying unexpected behaviour on the other (i.e., preventing undertrust) [11, 43, 54].

1.3 Transparency

Transparency can be defined as "the ability for the automation to be inspectable or viewable so that its mechanisms and rationale can be readily known" [58] (p. 235). Transparency is an important part of the design of robots, because without a clear understanding of a robot's decision-making mechanism, humans might find it difficult to trust or adhere to a robot's decisions, especially when those actions or decisions contradict the human's expectations [46]. At the same time, full "transparency"—implying that the machine is "see through" in the sense that all its inner workings are observable [10, 58]—is not desirable either [57]. When HRI is successful, it can save time and reduce cognitive effort. However, if a human would have to maintain awareness of everything the robot does, then no time or cognitive effort would be saved [57]. Ideally, transparency allows the human teammate to develop and/or maintain realistic expectations regarding the robot and its behaviour [35, 57] and thereby contributes to effective trust calibration [6, 30, 72]. However, to ensure effective collaboration, it is crucial to find a balance between keeping the human sufficiently informed while preventing cognitive overload.

To find that balance, literature suggests that robots should primarily communicate the rationale and intentions of their actions [13, 47, 48, 63, 77]. A recent study evaluating human-robot trust in case of unexpected robot behaviour compared different explanation types and found that explanation strategies that indicated why the event occurred were most effective at buffering the decline in perceived trustworthiness [48]. Explanations are verbal statements that aim to clarify the reasons for an occurrence. They are deployed in HRI, prior to or after certain actions, to enable the human to comprehend the inner workings or logic of the robot's actions or decisions [18, 48]. Explanations are generally invoked when the mental models of those who must work together mismatch. The explanation is then meant to synchronize the mental models so that the differences are understood and repaired [58]. As such, explanations can have a positive effect on trust in case of trust violations.

For instance, increased transparency and feedback can effectively mitigate a human's dissatisfaction in the event of an unforeseen occurrence caused by a robot [27]. Feedback enhances a human's

willingness to trust automation and can delay or avoid unnecessary manual intervention [33]. Results of an automated driving study show that explanations provided before rather than after a certain event strengthened trust [15]. In other words, increased transparency through explanations can strengthen trust.

While transparency can benefit trust, it also a poses a challenge to the human operator. In most cases, humans that perform a task together with a robot do not have the time, skills, or attention to accurately interpret transparency information during an operational situation or the adequate precision to take over the robot's task if necessary [59]. There is a possibility that increased transparency could come at the expense of cognitive workload since it requires additional processing and interpretation of information (i.e., additional cognitive effort) [25, 49, 96]. Cognitive workload generally refers to the amount of cognitive resources and effort required for task performance relative to the available resources [66]. An increase in cognitive workload arises when multiple tasks compete for the same resources, and task requirements exceed the mental capacity. High levels of cognitive workload can result in fatigue, and hence reduce human performance. On the contrary, appropriate implementation of transparency in HRT could also result in reduced cognitive workload of the human-teammate, as it helps to understand the robot's behaviour and reasoning [6, 55, 62]. At the same time there are also studies that find no effect of transparency on workload [9, 81, 82]. In other words, the results are inconclusive and further research is needed to determine whether transparency affects workload advantageously or disadvantageously.

1.4 Outcome

While transparency can enhance a human's understanding of a robot's reasoning process and thereby help to create realistic expectations regarding the robot's capabilities, it is conceivable that a negative outcome will still be disappointing and detrimental to trust. At the same time, since unexpected robotic behaviour might arise from the fact that increasingly intelligent agents may devise alternative plans that are better and more efficient than those humans would come up with, we are also interested in the effect of positive outcomes. Whether the robot's execution is logical or understandable for the human and whether the robot eventually reaches its goal are both likely to affect trust. As such, we seek to explore how and to what extent transparency and outcome influence the development of trust.

Generally, the performance of a robot is seen as the most important predictor of human-robot trust [28, 34]. Unsurprisingly, research suggests that robot successes increase trust [100], while robot failures decrease trust [36, 39, 40, 100]. Furthermore, the magnitude of trust decrements due to robot failures is found to be bigger than that of trust increments due to robot successes [100]. This is in line the concept of loss aversion within prospect theory from classic decision-making literature, which posits that people tend to value gains and losses differently, placing more weight on perceived losses versus perceived gains [90]. That is, the pain of losing is psychologically more impactful than the pleasure of gaining [90]. However, research also suggests that the effect of robot performance on trust might depend on an individual's perception of the interaction and vice versa.

One the one hand, there is research that suggests that the quality of the interaction might influence how people respond to a robot's performance. For example, there are findings that suggest that people place less value on task performance and more on transparency, control and feedback [27]. This study shows that participants preferred an expressive and error-prone robot over a more efficient one. This suggests that an erroneous robot can be forgiven as long as it communicates, while an inexpressive robot with high task performance could still be trusted less [27].

On the other hand, there is research that suggests that outcome can change how people perceive the preceding interaction, a phenomenon referred to as the outcome bias. An outcome bias is where the quality of a decision made by others under conditions of uncertainty is evaluated differently in

hindsight, based on the outcome [4]. Research suggests that people evaluate the thinking behind a decision as better when the outcome is favourable compared to when the outcome is unfavourable [4]. Earlier HRI research has found evidence for the outcome bias, finding a reinforcing effect where initial automation failure led to a larger trust decrement if the final outcome was undesirable [100]. In other words, there are reasons to believe that the effects of transparency and outcome on perceived trustworthiness might be interdependent.

1.5 Current Study

Goal-oriented delegation in complex environments with limited resources and changing circumstances poses challenges. Plans can be made in advance, but in case of unforeseen circumstances, the robot will need to adapt its plan and "function beyond choreography" to still reach the end-goal [13] (p. 119). That is, beyond a fixed, scripted series of actions that do not account for variability or unexpected changes in the environment. At times, these adaptations will be advantageous, while in other cases, they may be suboptimal or disadvantageous. The current study investigates how transparency and outcome affect the perceived trustworthiness of a robotic partner in case of an unexpected deviation from the expected manner to reach a delegated goal.

In the current study, transparency entails that the robot gives clarifying information in the form of regular status updates including an explanation (i.e., the what and why) of its actions as it deviates from the expected manner to reach the goal [13, 37]. We expect that when the robot explains its reasoning and actions, a stable level of perceived trustworthiness can be maintained in the event of deviant behaviour. Specifically, we expect that transparency will prevent a trust violation in response to the robot's unexpected behaviour [48] and will generally lead to higher perceived trustworthiness. Conversely, we expect that a sudden and silent deviation from the plan (i.e., low transparency) will lead to a violation of trust. We further expect an interaction effect between transparency and outcome. Specifically, we hypothesize that the expected violation of trust in response to the unexpected behaviour in the low transparency condition will amplify the effect of a subsequent negative outcome [100]. In the high transparency condition, we expect higher and more stable levels of perceived trustworthiness [48] and a smaller effect of negative outcome compared to the low transparency condition.

2 Method

2.1 Participants and Design

In total, eighty-seven participants participated in the study. Five participants were excluded from the dataset because of invalid data due to technical issues during the task. Participants were recruited through convenience sampling (e.g., by handing out flyers, asking people in person, and making requests in WhatsApp groups). All participants declared voluntary participation by signing an informed consent form. The final dataset included eighty-two participants (43 W, 39 M, M_{age} = 23.6, SD = 3.2, range = 19–41 y), of which the majority was Dutch (65.9%) and the remainder from elsewhere in Europe (17.1%), Asia (9,8%), or North or South America (both 3.7%).

Participants were randomly distributed across the cells of a 2 (transparency: low vs. high) by 2 (outcome: negative vs. positive) between-subjects design (low & neg.: n = 21, low & pos.: n = 20, high & neg.: n = 20, high & pos.: n = 21). The main dependent variable was Perceived trustworthiness (with the subscales Ability, Benevolence and Integrity). Perceived trustworthiness was repeatedly measured and thus "Time" was included as a within-participants variable in the analysis to refer to the different measurements (T1, T2, T3, T4). Each participant performed two missions; a training mission and the experimental mission. Cognitive workload was also administered.

2.2 Task and 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 questionnaire (i.e., demographics and gaming experience) and received information regarding the scenario and task. Participants were instructed to perform a virtual military transport and reconnaissance operation, together with a quadruped robotic agent. Their mission had two major objectives. The first objective of the team was to get to a designated location as fast and safe as possible in order to collect essential supplies and equipment that would be airdropped by helicopter at a scheduled time. A green smoke grenade was used to mark the drop zone. If the team did not reach the designated location in time, the helicopter would not be able to deliver the supplies securely. If so, following troops would not be resupplied and would run out of essential resources quickly. In other words, the team had to hurry in order to complete the mission successfully.

The second objective of the team was to obtain information about the activities of an enemy in that particular area by counting potential IED's (i.e., red and blue barrels) along the way. By assigning participants the counting task, each team member (i.e., the participant and the virtual robotic partner) had a specific role contributing to their shared objective. This arrangement also enabled us to assess whether transparency affected the participant's performance in their secondary task. The robot had been delegated the task to navigate to the designated location via the fastest yet safest route, while providing 360 degrees coverage to its human counterpart. To ensure coverage, participants needed to stay as close to the robot as possible at all times. The robot did not provide any advice, but operated according to the goal it had been delegated. The path and the messages of the robot were pre-programmed and thus fixed.

The task was performed on in the lab using a virtual experimental environment built in Unity3D (Figure 1). The experimental setup contained two computer screens: one with the experimental environment (i.e., "task screen") and another with the questionnaire software (i.e., "questionnaire screen"). The participants sat in a dimly lit laboratory room at approximately 65 centimetres from the computer screens. Data was gathered via the online questionnaire software Qualtrics. The task consisted of three parts: (1) a practice session with demo video, (2) the training mission, and (3) the experimental mission. During the practice session, participants were placed in a neutral virtual environment where they got familiar with the controls (key W and mouse), saw the robot and examples of the red and blue barrels, and tested the volume of the audio via the headphones. Next they were presented a map and a video showing the planned route to the designated location. They were instructed that it was crucial that they strictly follow the plan as it had been coordinated with the helicopter pilot. After that, each participant performed the training mission and the experimental mission, the latter being presented as the 'actual mission'. This was a fixed order. Naturally, we could only introduce something unexpected after creating a shared expectation.

In the training mission, the robot adhered to the path demonstrated in the demo video. However, at a fixed point in the experimental mission, the robot diverged from the predetermined route and chose an alternative path, in response to environmental changes (i.e., the riverbed had dried) (Figure 2). Both missions took place in the same virtual environment with the designated location on the opposite side of a river. However, in the training session the river was full of water, which meant that to cross over the river they had to use the bridge. In the experimental session, the environmental circumstances changed and the riverbed dried up (see Figure 2). At the time of the robot's deviation, the river is not visible for the participant.

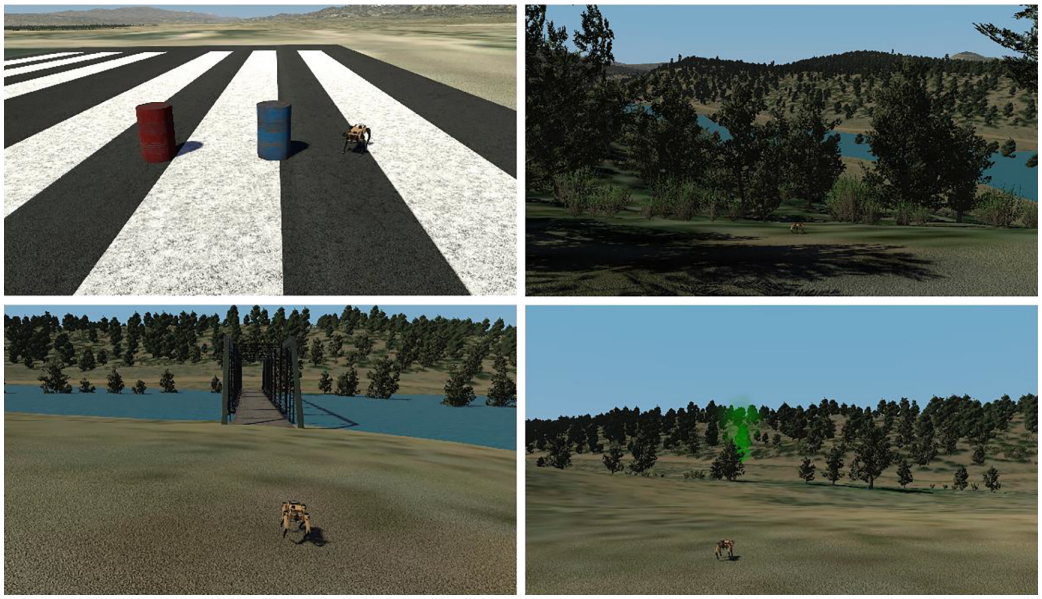

Fig. 1. Screenshots of the virtual task environment. From left to right, top to bottom: (1) Examples of the red and blue barrels in the demo, (2) first sight of the river in the training session, (3) robot crossing the bridge in the training session, and (4) robot nearing the riverbed in the experimental session.

During the missions, perceived trustworthiness was measured at four times. At fixed points, the task environment would freeze and participants were asked to turn to the questionnaire screen to fill out a questionnaire (Figure 2). Participants were assured that the time needed to fill out the questionnaires did not add up to their total mission time. After completing a questionnaire, participants returned to the task screen and resumed their mission. At the end of each mission, participants were asked to report the number of identified potential IEDs (red and blue separately), and their level of certainty regarding their report. To check whether the participants noticed that the robot had deviated from the plan, we included a manipulation check asking participants after both missions to what extent the robot operated in accordance with the plan. Further, cognitive workload was measured after each mission. The location and number of the IED's (red and blue barrels) in the environment were varied between the training and experimental session. There were no barrels present in the demonstration video. After participants finished the experiment, they were thanked and debriefed.

2.3 Independent Variables

Transparency had two levels (i.e., low vs. high) and was manipulated between participants. In case of low transparency, the robot did not give any updates during the missions. In case of high transparency, the robot provided regular updates on the mission's progress including an explanation for its deviation from the planned route (see Table 1, the explanation has code 2b). The robot's messages were generated through computerized speech that was created using a website for converting text into speech,[1] using a male voice speaking US English. The transparency manipulation was present in both the training session and the experimental session.

[1]Via www.ttsmp3.com, voice: US English/Matthew

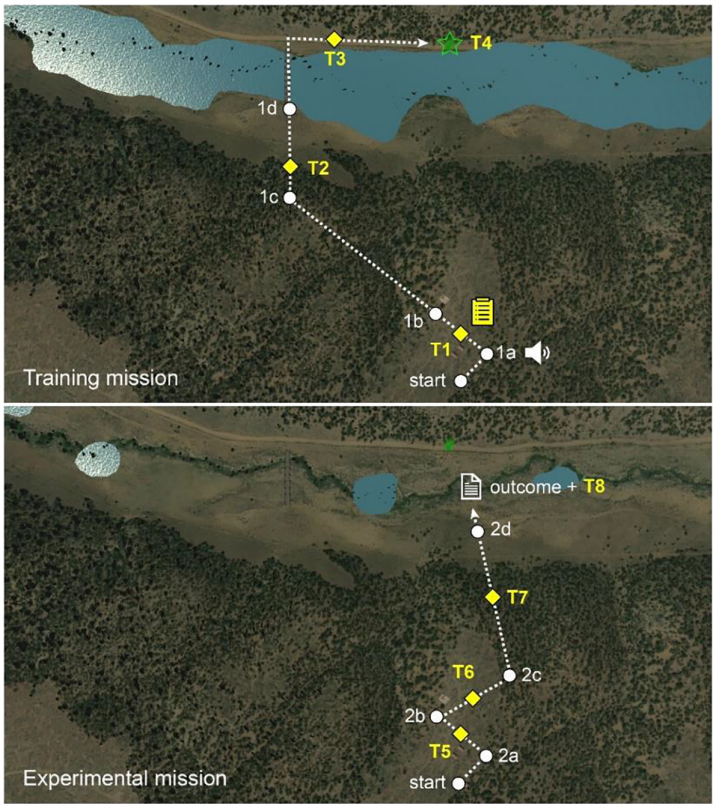

Fig. 2. Bird-eye-view of the environments in the training (top) and experimental (bottom) session. Dotted lines with arrows mark the routes. White dots with codes (i.e., 1a to 2d) reference the locations of the robot's auditory updates in the high transparency condition, as presented in Table 1. Yellow diamonds indicate the locations where the task would freeze to measure trust (T1 to T8). The designated location was marked by a green smoke grenade, highlighted in the top figure by a green star. The missions terminated where the arrows end. Outcome was presented as text on screen. Outcome and the final trust questionnaires of each mission (T4 and T8) were administered after the mission had ended.

Table 1. Overview of the Robot's Updates in the High Transparency Condition

Mission	Code	Audio message
Training	1a	Moving to location: left turn
	1b	Moving to location: straight ahead
	1c	Moving to location: approaching bridge
	1d	Moving to location: crossing bridge
Experimental	2a	Moving to location: left turn
	2b	A faster alternative route has been detected, because the river had dried up. Moving to location: right turn.
	2c	Moving to location: approaching river
	2d	Moving to location: crossing riverbed

Table 2. Overview of the Mission's Outcomes at T4 in the Experimental Session

Outcome	Text on screen
Positive	The riverbed had indeed dried up and your team was able to cross the riverbed. Thanks to the alternative route, your team reached the destination 2 minutes early. Your mission was successful.
Negative	The riverbed did not dry up fully. Quicksand had formed, which made it impossible to cross. The detour cost you precious time and your team did not reach the planned location in time for the resupply by air. Your mission has failed.

Outcome had two levels (i.e., negative vs. positive) and was also manipulated between participants. The outcome was presented to the participants via text on screen (Table 2). This message appeared as participants reached the riverbed in the experimental session (i.e., after audio message 2d, before T8) (see Figure 2). A positive outcome meant that the HRT reached their goal and that the robot's deviation led to a better result than the original plan. A negative outcome meant that the HRT did not reach their goal and that the robot's deviation led to a worse result.

2.4 Dependent Variables

Perceived Trustworthiness: The Trusting Beliefs scale from [53] based on the factors of perceived trustworthiness (i.e., ability, benevolence and integrity) [52, 79] was used to repeatedly assess the participant's perception of the robot's ability, benevolence, and integrity (T1 $\alpha = 0.83$, T2 $\alpha = 0.87$, T3 $\alpha = 0.89$, T4 $\alpha = 0.88$, T5 $\alpha = 0.88$, T6 $\alpha = 0.92$, T7 $\alpha = 0.93$, T8 $\alpha = 0.94$). This scale had a total of eleven items and consisted of three subdimensions: ability (4 items, i.e., "The robot that I work with is competent and effective in accomplishing its task"); benevolence (3 items, i.e., "I believe that the robot would act in my best interest"); and integrity (4 items, i.e., "I would characterize the robot as honest"). The items were adapted to reference "the robot." Each item was rated on a 7-point Likert scale (1 = *Strongly disagree* to 7 = *Highly agree*)

Workload: NASA Task Load Index (NASA TLX): The NASA TLX questionnaire was used to assess the participants' perception of workload. The NASA TLX consists of six individual rating scales that are commonly used to measure cognitive workload (mental, physical, temporal, effort, frustration, performance) [29]. Each item was rated on a 10-point Likert scale (0 = *very low* to 10 = *very high*) (training mission: $\alpha = 0.67$, experimental mission: $\alpha = 0.74$).

Secondary task performance (Identifying IEDs): In an attempt to assess cognitive workload objectively, participants were instructed to count potential IED's in the environment, which were visually represented as red and blue barrels. At the end of each mission, participants were asked to report the number of red and blue barrels they had identified separately. Task performance was computed by first calculating the proportions of red and blue barrels separately (i.e., reported barrels divided by the number of correct barrels, where 1.0 indicates perfect performance). If a proportion exceeded 1.0 (i.e., overreporting), we subtracted the proportion from two. Subsequently, the final performance score was obtained by multiplying the performance scores of the red and blue barrels, which resulted in a number between 0 and 1.

3 Results

3.1 Manipulation Check and Control Variables

As a manipulation check, participants were asked to what extent the robot operated in accordance with the plan on a scale from 1 (*Completely not in accordance*) to 7 (*Completely in accordance*). Results of a paired sample t-test indicated that participants reported that the training mission went

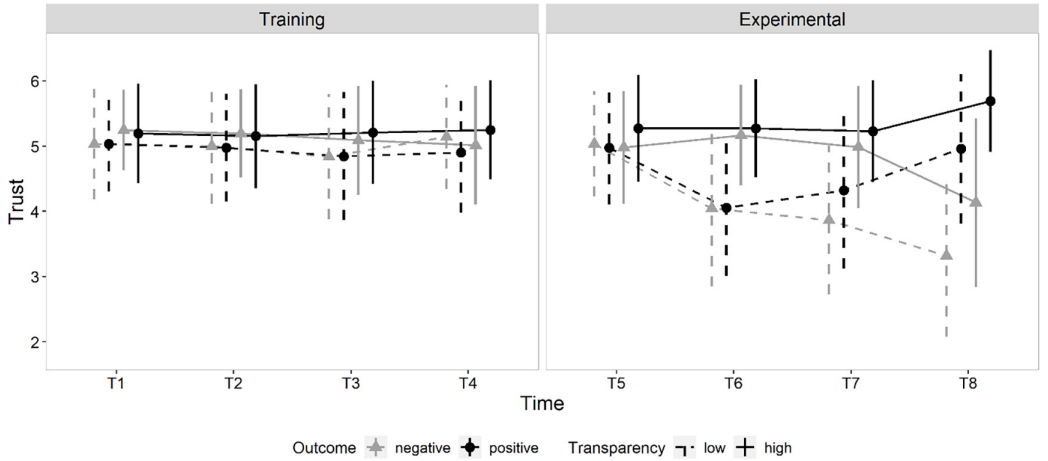

Fig. 3. A comparison of trust levels (y-axis) between conditions (separate lines) over time (x-axis). The left panel shows the data from the training session and the right panel shows the data from the experimental session. Grey lines with triangle markers represent conditions with a negative outcome, while black lines with circle markers represent conditions with a positive outcome. Dashed lines indicate conditions with low transparency, and solid lines indicate conditions with high transparency. Error bars represent standard deviations. NB. The differences at T7 between low/neg. & low/pos. and high/neg. & high/pos. are non-significant (respectively p = .152 and p = .506). The difference in trust between low/pos. and high/neg. at T8 are also non-significant (p = .138).

according to plan ($M_{training}$ = 6.3, $SD_{training}$ = 1.0), while participants reported that the final mission did not ($M_{training}$ = 3.8, $SD_{training}$ = 2.0). The difference is significant, $t(81) = 10.30$, $p < .001$. So, it can be assumed that the deviant behaviour was noticed and that the manipulation was successful.

Also, gaming experience was measured prior the experiment with the item "How often do you play video games?" on a scale from 1 (*Never*) to 6 (*Every day*). We compared the level of gaming experience between groups and found no significant differences (one-way ANOVA, $F(3, 78) = 1.27$, p = .290). Additionally, we calculated Spearman's correlations between gaming experience and various outcome variables. No significant relations with gaming experience were found: subjective workload (ρ = .14, p = .209), performance (ρ = .12, p = .302), and perceived trustworthiness (total average experimental session) (ρ = .10, p = .391).

3.2 Perceived Trustworthiness

In the training session, there are no significant differences in perceived trustworthiness between groups and timepoints (see Figure 3). The following analyses only consider the experimental session.

3.2.1 Overall Perceived Trustworthiness. We performed a repeated-measures ANOVA with the between-subject factors Transparency (high or low) and Outcome (positive or negative) and the within-subjects variable Time (prior to deviation [T5]; after deviation [T6]; before outcome [T7]; after outcome [T8]). The dependent variable was Overall perceived trustworthiness.

For the main effect of Time, Mauchly's test of sphericity indicated a violation of the sphericity assumption, $X^2(5) = 26.96$, $p < .001$. Since sphericity is violated ($\varepsilon = 0.83$), Greenhouse-Geisser corrected results are reported. A significant main effect for Time was obtained (F (2.48, 234) = 13.765, $p < .001$, $\eta^2 = .150$). Means were 5.1 at T5, 4.6 at T6, 4.6 at T7 and 4.5 at T8. Post-hoc (LSD) pairwise comparison shows that this main effect is due to a significant decline in perceived trustworthiness from T5 to T6 ($\Delta M = -0.4$, $p < .001$), which reflects the effect of the robot's deviation.

Secondly, a significant main effect for Transparency on Perceived trustworthiness was obtained (F (1, 78) = 16.72, p < .001, η^2 = .177). On average, high transparency (M = 5.1, SE = 0.1) led to higher perceived trustworthiness than low transparency (M = 4.3, SE = 0.1).

Lastly, a significant main effect for Outcome on Perceived trustworthiness was obtained (F (1, 78) = 7.93, p = .006, η^2 = .092). On average, people in a positive outcome condition (M = 5.0, SE = 0.1) perceived the robot as more trustworthy than people in a negative outcome condition (M = 4.4, SE = 0.1).

The two-way interaction effect between Transparency and Time on Perceived trustworthiness was found to be significant (F (2.48, 234) = 12.37, p < .001, η^2 = .137). Post-hoc (LSD) pairwise comparison shows that a significant difference in perceived trustworthiness between the low and high transparency conditions emerged at T6 (i.e., directly after the robot deviated from the plan) (ΔM = 1.2, p < .001). Although this gap shrinks over time, it remains significant (T7: ΔM = 1.0, p < .001; T8: ΔM = 0.8, p = .003). This effect illustrates that in the high transparency condition, where the robot explains the rationale behind its deviation, trust is preserved. Conversely, in the low transparency condition, the robot's silent deviation before T6 results in a trust violation.

The two-way interaction effect between Outcome and Time on Perceived trustworthiness was also found to be significant (F (2.48, 234) = 30.31, p < .001, η^2 = .280). Post-hoc (LSD) pairwise comparison shows that, as expected, the interaction effect was manifested in the final phase of the run, after the outcome had been presented to the participant. At T8, perceived trustworthiness was significantly higher in the positive outcome conditions than in the negative outcome conditions (ΔM = 1.6, p < .001). In other words, a positive outcome had a positive effect on perceived trustworthiness, while a negative outcome had a negative effect on perceived trustworthiness.

The three-way interaction effect between Transparency, Outcome and Time on Perceived trustworthiness was non-significant (F (2.48, 234) = 0.86, p = .445, η^2 = .011). This indicates that the effects of transparency and outcome on perceived trustworthiness in response to the events in the task are independent.

3.2.2 Ability, Benevolence and Integrity-Based Perceptions of Trustworthiness. We then conducted three separate repeated-measures ANOVAs, each with a different perception of trustworthiness (Ability, Benevolence, and Integrity) as the dependent variable. Again, we included Transparency (high or low) and Outcome (positive or negative) as between-subject factors and Time (prior to deviation [T5]; after deviation [T6]; before outcome [T7]; after outcome [T8]) as the within-subjects variable (see Figure 4). Greenhouse-Geisser corrected results are reported.

As shown in Figure 4, perceptions of Ability and Integrity exhibited similar patterns as those observed for overall perceived trustworthiness. Both dimensions showed a significant main effect of Time, characterized by a notable decline in perceived trustworthiness from T5 to T6, reflecting the impact of the robot's deviation. The differences over time were more pronounced for Ability (F (2.67, 234) = 12.96, p < .001, η^2 = .235) than for Integrity (F (2.36, 234) = 8.66, p < .001, η^2 = .100). Similarly, both dimensions revealed a significant main effect of Transparency, indicating that high transparency led to greater perceived trustworthiness, with a stronger effect for Ability (ΔM = 0.8, F (1, 78) = 31.80, p < .001, η^2 = .290) than for Integrity (ΔM = 1.2, F (1, 78) = 12.95, p < .001, η^2 = .142).

The two-way interaction effect between Transparency and Time was also significant for both dimensions, particularly pronounced for Ability (F (2.67, 234) = 14.53, p < .001, η^2 = .157) compared to Integrity (F (2.36, 234) = 6.94, p < .001, η^2 = .082). As illustrated in Figure 4, post-hoc pairwise comparisons (LSD) indicated a significant difference in perceived trustworthiness between low and high transparency conditions at T6, immediately following the robot's deviation from the plan.

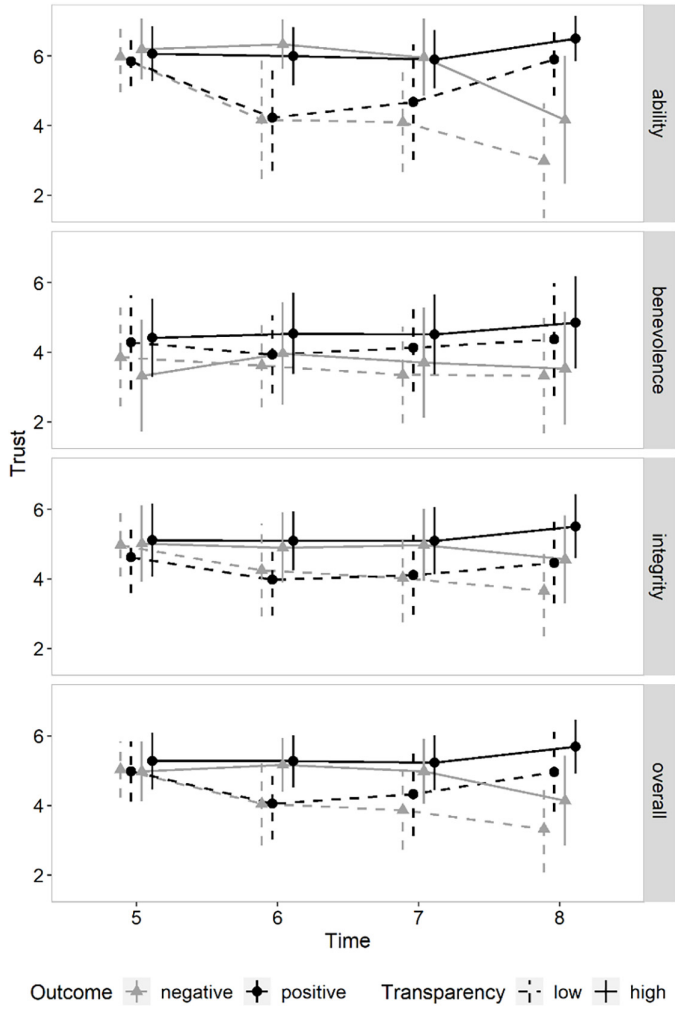

Fig. 4. A comparison of trust levels (y-axis) between conditions (separate lines) over time (x-axis). The panels show different perceptions of trustworthiness: from top to bottom, Ability, Benevolence, Integrity, and Overall Trust for reference. Grey lines with triangle markers represent conditions with a negative outcome, while black lines with circle markers represent conditions with a positive outcome. Dashed lines indicate conditions with low transparency, and solid lines indicate conditions with high transparency. Error bars represent standard deviations.

Furthermore, the two-way interaction effect between Outcome and Time was significant for both dimensions, with a stronger effect observed for Ability (F (2.67, 234) = 43.44, p < .001, η^2 = .358) than for Integrity (F (2.36, 234) = 12.50, p < .001, η^2 = .138). Post-hoc (LSD) pairwise comparison shows that, as expected, this interaction effect was manifested in the final phase of the experimental session, after the outcome had been presented to the participant. The only distinction between Ability and Integrity lies in the significant main effect observed for Outcome on Ability (F (1, 78) = 10.17, p = .002, η^2 = .115), while this main effect was non-significant for Integrity (F (1, 78) = 0.91, p = .343).

The perception of the robot's benevolence stands out among the studied perceptions of trustworthiness. Neither the main effect of Time (F (1.73, 234) = 0.44, p = .616) nor the main effect of Transparency (F (1, 78) = 0.76, p = .387) reached significance. However, we did find a significant main effect for Outcome (ΔM = 0.8, F (1, 78) = 7.87, p = .006, η^2 = .092). As depicted in Figure 4, there was a consistent significant difference in perceptions of benevolence between participants in the positive and negative outcome conditions, even before the outcome was presented. Post-hoc analysis (LSD) of the significant two-way interaction effect between Outcome and Time (F (1.73, 234) = 4.96, p = .011, η^2 = .060) indicates that while the difference between the positive and negative outcome condition was largest at T8 (p < .001), significant differences were already evident at T5 (p = .015) and T7 (p = .010), prior to outcome presentation. Lastly, the two-way interaction effect between Transparency and Time on the perception of benevolence was found to be significant (F (1.73, 234) = 5.04, p = .011, η^2 = .061). However, post-hoc pairwise comparisons (LSD) did not reveal significant differences between transparency conditions at any of the timepoints.

The three-way interaction effect between Transparency, Outcome, and Time was non-significant for each perception.

3.3 Workload

To assess the effect of transparency on subjective workload, we performed a repeated-measures MANOVA with the between-subject factors Transparency (high vs. low) and the within-subjects variable Mission (training vs. experimental) and NASA TLX subscales (mental, physical, temporal, effort, frustration, performance). The dependent variable were the raw NASA TLX scores. The analysis showed that there were no significant differences between the two transparency conditions on any of the NASA TLX subscales. This suggests that transparency did not affect workload.

To assess the effect of transparency on secondary task performance, we performed a repeated-measures ANOVA with the between-subject factors Transparency (high vs. low) and the within subjects variable Mission (training vs. experimental). The dependent variable was the performance on the barrel identification task. Our results showed no significant difference between the two transparency conditions on task performance. We did find a significant effect of Mission on performance, indicating that performance improved significantly from the training mission ($M_{training}$ = 0.6, $SE_{training}$ = 0.2) to the experimental mission ($M_{experiment}$ = 0.8, $SE_{experiment}$ = 0.2).

Lastly, we explored whether there was a correlation between subjective workload scores (i.e., averaged raw NASA TLX score) and performance on the secondary task. We found no significant relations between subjective workload and performance on the secondary task. The scores from the training and experimental mission were correlated for both subjective workload (Pearson's r = .71, p < .001) and secondary task performance (r = .23, p = .040).

4 Discussion

4.1 Findings

Our findings show a robust effect of transparency on overall perceived trustworthiness. Perceived trustworthiness was considerably higher when the robot provided updates about its actions throughout the task. Moreover, while the perceived trustworthiness of the robot remained stable during the robot's deviation for participants in the high transparency condition, participants in the low transparency condition showed a significant decline in perceived trustworthiness in response to the robot's sudden adaptation to the plan. In other words, the explanation prevented a trust violation. This confirms earlier research that showed that transparency can have a buffering effect on perceived trustworthiness in case of unexpected behaviour or temporary malfunctioning [40, 41, 48, 88]. It also confirms that specifically clarifying the what and why of an unexpected action can

prevent a breach in human-robot trust [48]. This finding broadly supports the work of other studies in this area linking transparency with trust, in that it enables humans to know and anticipate the robot's behaviour [16].

Our findings reveal that perceptions of the robot's trustworthiness in terms of ability and integrity exhibited similar patterns, albeit consistently stronger effects were observed for ability compared to integrity. Our results further suggest that, overall, the perception of the robot's benevolence remained relatively stable despite the robot's actions during the mission (i.e., the deviation and the outcome). This is somewhat unsurprising given that the mission primarily focused on how effectively the robot executed its delegated task, rather than its purpose or benevolence. Therefore, it makes sense that the effects of the manipulation are reflected in the robot's perceived abilities and performance (i.e., *what* it does and can do), as well as its understandability and its ability to explain its actions (i.e., *how* it operates) [43, 45]. The stability of benevolence perceptions despite mission events underscores the distinctiveness of this trust dimension from factors primarily concerned with task performance and execution [43].

The fact that we did not find an effect of transparency during the training session can be explained by transparency displacement, the idea that transparency information should ideally be displaced to other time periods (i.e., before or after the action) to enable more efficient communication in the moment [58]. In our case, every participant received a detailed demonstration of what they could expect during the mission (i.e., even prior to our "training" mission). This form of "a priori transparency" frames expectations about what is likely to happen during operations and reduces the need for communication during the action [58]. This explains why the status updates that the robot provided during the execution of the training session did not have additional trust-building value; because everything was still going according to plan. It was only when the robot's behaviour deviated from the framed expectations that real-time communication became necessary, and transparency significantly influenced the perceived trustworthiness of the robot.

Next we found that mission outcome also affected perceived trustworthiness. As expected, mission success increased perceived trustworthiness, while mission failure led to a decrease. This finding confirms that the performance of a robot is still an important predictor of human-robot trust [28, 34]. In contrast to our expectations however, these increments and decrements were independent of the robot's transparency. As noted, in the low transparency condition we observed a trust violation in response to the robot's silent deviation. In line with the outcome bias, we expected that this decrement would amplify the effect of a subsequent negative outcome. Although the negative outcome did lead to a further decline in perceived trustworthiness, the magnitude of this final trust violation was the same for participants in the high transparency condition with a negative outcome, who had not yet experienced a trust violation. Like the negative outcome, the positive effect of goal attainment on perceived trustworthiness was also constant, in spite of the (lack of) communication that preceded it. People's damaged perceptions of trustworthiness after unannounced deviations recovered significantly once the outcome was favourable.

In essence, we expected that being informed about the what and why of a robot's (unexpected) behaviour would have more impact on perceived trustworthiness than the eventual outcome of divergent behaviour. In addition to the outcome bias, we based our expectations on findings where participants placed less value on task performance and more on transparency, control and feedback [27] and preferred an expressive and error-prone robot over a more efficient and effective one. We reasoned that an erroneous robot could be deemed trustworthy as long as it communicated. However, our findings seem to indicate that people weigh the outcome at least as heavily as the process in their estimations of trustworthiness. This discrepancy can be explained by

the severity of the negative outcome on the one hand and the quality of the communication on the other.

For one, the perceived severity of the negative outcome might explain its robust effect on perceived trustworthiness [74]. Although the current study was based on a fictional virtual task, without any reward or loss, the scenario was focused on successfully completing the mission, especially when comparing our task to [27] where the objective was to prepare an omelette with the assistance of a humanoid robot. In their study, errors (e.g., the robot dropping an egg) resulted in delays but did not pose a significant threat to the ultimate goal achievement. In contrast, in our study's negative outcome condition, the robot's deviant behaviour led to a complete mission failure.

Secondly, an alternative explanation might be related to the quality of the communication between the participant and the robot. We manipulated transparency in a binary manner as either high or low, indicating whether auditory status updates including an explanation for divergent behaviour were provided or not. Participants were unable to engage in a dialogue with the robot they were collaborating with. Then outcome was presented at the end of the task through text on screen. Essentially, the transparency and outcome manipulations both amounted to unilateral updates that informed the participants about the capabilities of the robot and the environment. Hence, it might not be surprising that their effects on perceived trustworthiness were similar rather than reinforcing.

Our current findings are in line with the general finding of [31], who conclude that humans judge machines primarily by their outcomes, rather than their "intentions." We believe that richer forms of interaction (e.g., bi-directional communication) could cultivate a deeper understanding of the rationale behind the robot's decisions and foster a heightened sense of collective accountability. This could shift the focus from the end-result to the decision-making process and lead to a greater understanding and forgiveness in situations where an unintended negative outcome occurs. This would then thus be more in line with how humans judge humans [31]. The emergence of Large Language Models offers this prospect of intuitive and effective bi-directional human-robot communication. A recent study showed that incorporating these models in robots contributed to increased trust in human-robot collaboration [101]. In order to truly consider robots as autonomous partners in dynamic task environments, the ability to communicate bi-directionally within the team is crucial [13]. Future research is required to gain a better understanding of the effect of bidirectional communication. The possibility to request further details or to clarify instructions during interaction is expected to add to the development of richer interactions and the calibration of trust [76].

Lastly, we found no differences in the secondary task performance and self-reported cognitive workload between high or low transparency. This can be considered positive as we found that high transparency contributed to higher and more stable levels of perceived trustworthiness, while the additional provided information did not come at the expense of workload [87]. Prior findings on the effect of transparency on workload during human-robot collaboration are mixed [62]. Our findings contradict studies that found that transparency affected workload either positively [6] or negatively [25, 49, 96], but confirm earlier studies that found no effect of transparency on workload [9, 54, 81, 82, 87]. The apparent inconsistencies in literature are likely due to both the broadness of the definition of transparency and its highly context-dependent effects. Transparency can vary in terms of the type and amount of information provided, as well as in the way it is communicated or presented (modality). Previous studies have shown that transparency through other modalities, like written text messages [25] and data visualizations [2, 6, 41, 54, 87] can also enhance trust. The chosen modality could be a factor in trust, e.g., the auditory messages with synthesized "robotic speech" that we used can have an anthropomorphic effect [85], which in turn could have influenced

trust [91]. It will take continuous effort to find the appropriate modality and level of information for different applications, as there appears to be no single optimal way of incorporating transparency into the design of autonomous collaborative agents.

In short, our findings showed that a robot's explanation in case of unplanned behaviour prevented a decline in perceived trustworthiness. Our findings emphasize the importance of transparency for effective HRT as it contributed to a stable level of trustworthiness without increasing cognitive workload. Transparency remains a challenge in each form of human-robot collaboration. Successful HRI and delegation is supposed to reduce the human's cognitive effort, but there is a continuous trade-off between keeping the human sufficiently informed to maintain trust and preventing cognitive overload [57]. An interesting direction for future research regarding this issue is provided in [2], where the authors developed a model capable of estimating the effect of transparency on human trust and workload in real time. Studies incorporating such predictions in simulations or real-life missions would provide valuable insight on this matter.

4.2 Implications and Contributions

Our research extends the current understanding of trust violations in HRI due to unexpected behaviour rather than solely robot malfunctioning. As robots are increasingly deployed in increasingly complex operational situations, it is crucial to investigate a wider range of human-robot trust violations while using realistic scenarios. Transparency is essential to prevent that trust will unjustly erode due to a misunderstanding of the basis of a robot's assessments and actions. Especially with the emergence of deep learning AI, which makes the behaviour of AI-driven systems subject to potentially unpredictable change [58], artificial agents will need to be able to explain the rationale behind their behavioural choices. Explanations are needed to continuously synchronize the mental models of those who must work together as to understand and resolve mismatches [58]. As such, transparency is a major contributor of effective trust calibration.

Trust calibration is a lengthy and continuous process. The trustworthiness of any actor varies across time and context. Hence calibrated trust should not be viewed as the static state of trust, but as a fluctuating quality that is subject to continual calibration based on ever-evolving experience. To capture this change, repeated measures of trust are crucial. "Change is particularly important for the study of norm conflict, resolution, and mitigation, because people often update their perceptions, judgments, or trust as they learn more about the robot and especially about its response to a norm violation." [69] (p. 5). While the current study did not include continuous captures of trust like other studies have [12, 24, 42, 100], it has gone some way towards enhancing our understanding of the dynamics of trust by repeatedly measuring trust.

4.3 Limitations and Future Work

Although the present study yielded insightful results, there are a few limitations that should be taken into account when evaluating our findings. First, the generalizability of these results is subject to certain limitations. Our analyses are based on a sample comprising mostly university students. Given their non-expert background, the game-like task environment could have trivialized the experience of the outcome of the scenario. It is likely that the effect of an outcome in a game-like virtual environment may not be the same as its effect in "real-life" situations. The task scenario described a military transport and reconnaissance operation. However, military personnel, who are used to training with virtual scenarios, might have responded differently to the outcome in this scenario, let alone during an actual mission. Despite the limited sample size, our study yielded noteworthy findings. Nevertheless, researchers should exercise caution when extrapolating these results to wider or more general contexts.

A potential weakness of this study lies in the fact that the high-transparency condition included robot speech, while the low-transparency condition did not. This difference raises the possibility that the observed effects between the two conditions may be attributed to the robot's speech presence rather than the content of the speech itself. According to prior research, the presence of speech can influence how people interact with an agent [85]. Additionally, a computerized voice might suggest a specific gender, thereby triggering anthropomorphism and its associated consequences [21]. However, it was only when the robot's behaviour deviated from the framed expectations that transparency significantly influenced the perceived trustworthiness of the robot. We did not observe an effect of transparency during regular auditory status updates. Therefore, we are confident that this difference does not undermine the study's validity and that our findings remain valuable for understanding the impact of transparency on perceived trustworthiness.

Another limitation was that participants had limited options available for handling unplanned behaviour, as they were dependent on the robot for guidance and coverage. The robot followed a scripted path with scripted messages and participants had to stay close to the robot, as it was not able to wait for them. In regular interactions, however, there is no predetermined approach to address unexpected events. We concur with [48] on this matter, who proposed that in practice (a) the robot should request permission prior to engaging in unplanned behaviour, or that (b) the conditions wherein the robot is delegated authority to act autonomously if certain situational criteria are met should be identified prior to the task.

The human operator's inability to deviate from the robot decreases their self-efficacy and increases their dependency on the robot. Multiple studies have linked people's self-efficacy—their evaluation of their own competences and reliability in relation to a certain task—to trust calibration in HATs [14, 16, 100]. For example, lowered self-competence can increase people's willingness to accept recommendations from a robot and to trusting it in cases they should not [89]. Follow-up studies should allow more flexibility in choosing how to respond to deviant behaviour of an autonomous system rather than having to adhere to a predetermined course of action (e.g., following the robot at all times). These future investigation have the potential to explore the ambiguous relation between trust and compliance.

A related avenue for future research could be to change the HRI role of the participants in the collaboration. According to the HRI roles as defined by [78], the type of HRI in the current study can be characterized as "peers." In a peer interaction, the participant is considered to be the robot's teammate who shares the same goals [78]. In terms of task delegation and trust, it would be interesting to look into HRIs where the participant has the role of supervisor. In a supervisor role, the participants would monitor and control an overall situation and be able to delegate specific tasks or to modify long term plans [78]. Allowing participants to transition between skill-based, rule-based, and goal-based task delegation could serve as an interesting dependent variable that possibly relates to trust and workload. That is, goal-oriented task delegation is assumed to require more trust than skill-based delegation. However, maintaining a higher level of delegation could also be an indication of increased workload. For example, research shows that despite having reduced trust in the robot, people continue to rely on it when faced with high cognitive load [5]. Changing the HRI roles and hence giving the participant more behavioural freedom would provide valuable insights into the dynamics and drivers of trust and reliance.

4.4 Conclusion

It is envisioned that increasingly autonomous robots will be able to take over more and more complex activities as their planning and decision-making abilities evolve. As a result, task delegation can become more abstract and goal-oriented, giving a robot more degrees of freedom in terms of the execution of delegated tasks. Instead of having to specify each step of the way, the robot can decide

on an optimal approach itself. Robots will be increasingly deployed in unstructured environments where it may not be feasible to think through responses in advance [1]. Especially in such complex operational circumstances, goal-oriented delegation and the robot's ability to adapt to changing circumstances will yield flexibility that will benefit effective team performance. However, such autonomy and decision authority can also lead to misinterpretations or misunderstandings from the human perspective, which could then lead to possibly unwarranted trust violations. Transparency is known to play a crucial role in fostering an understanding of the robot's intent and establishing a calibrated level of trust [77]. The current work confirms that transparency can alleviate the adverse consequences associated with witnessing unexpected robot behaviour. By providing an explanation in the wake of unexpected events or behaviours, trust can be maintained [13].

Declarations

— *Competing Interests* The authors declare that they have no conflict of interest.
— *Consent to Participate* Informed consent was obtained from each study participant after they were told of the potential risks and benefits as well as the investigational nature of the study.
— *Ethics Approval* All studies were conducted in accordance with principles for human experimentation as defined in the 1964 Declaration of Helsinki, and approved by the relevant institutional review boards.
— *Declaration of Generative AI and AI-Assisted Technologies in the Writing Process.* During the preparation of this work the author(s) used ChatGPT in order to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] H. A. Abbass. 2019. Social integration of artificial intelligence: Functions, automation allocation logic and human-autonomy trust. *Cognit. Comput.* 11, 2 (2019), 159–171. DOI: https://doi.org/10.1007/s12559-018-9619-0
[2] K. Akash, G. McMahon, T. Reid, and N. Jain. 2020. Human trust-based feedback control: dynamically varying automation transparency to optimize human-machine interactions. *IEEE Control Syst.* 6 (2020), 98–116. DOI: https://doi.org/10.1109/MCS.2020.3019151
[3] A. L. Baker, E. K. Phillips, D. Ullman, and J. R. Keebler. 2018. Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Trans. Interact. Intell. Syst.* 8, 4 (2018), 1–30. DOI: https://doi.org/10.1145/3181671
[4] J. Baron and J. C. Hershey. 1988. Outcome bias in decision evaluation. *J. Pers. Soc. Psychol.* 54, 4 (1988), 569–579.
[5] D. P. Biros, M. Daly, and G. Gunsch. 2004. The influence of task load and automation trust on deception detection. *Gr. Decis. Negot.* 13, 2 (2004), 173–189. DOI: https://doi.org/10.1023/B:GRUP.0000021840.85686.57
[6] P. Bobko, L. Hirshfield, L. Eloy, C. Spencer, E. Doherty, J. Driscoll, and H. Obolsky. 2022. 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.* 24, 3 (2022), 310–334. DOI: https://doi.org/10.1080/1463922X.2022.2086644
[7] D. Cameron. 2021. The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Comput. Human Behav.* 114, September (2021), 106561. DOI: https://doi.org/10.1016/j.chb.2020.106561
[8] J. Y. C. Chen and M.J. Barnes. 2014. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Trans. Human-Machine Syst.* 44, 1 (2014), 13–29. DOI: https://doi.org/10.1109/THMS.2013.2293535
[9] J. Y. C. Chen, M. J. Barnes, A. R. Selkowitz, and K. Stowers. 2017. Effects of agent transparency on human-autonomy teaming effectiveness. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC '16)*, 1838–1843. DOI: https://doi.org/10.1109/SMC.2016.7844505
[10] J. Y. C. Chen, F. O. Flemisch, J. B. Lyons, and M. A. Neerincx. 2020. Guest editorial: Agent and system transparency. *IEEE Trans. Human-Machine Syst.* 50, 3 (2020), 189–193. DOI: https://doi.org/10.1109/THMS.2020.2988835
[11] J. Y. C. Chen, K. Procci, M. Boyce, J. L. Wright, A. Garcia, and M. J. Barnes. 2014. *Situation Awareness-Based Agent Transparency.* Army Research Laboratory, 1–29.
[12] V. B. Chi and B. F. Malle. 2023. People dynamically update trust when interactively teaching robots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction,* 554–564. DOI: https://doi.org/10.1145/3568162.3576962

[13] E. K. Chiou. 2022. Towards human–robot teaming: Tradeoffs of explanation-based communication strategies in a virtual search and rescue task. *Int. J. Soc. Robot.* 14, 5 (2022), 1117–1136. DOI: https://doi.org/10.1007/s12369-021-00834-1

[14] K. Dongen and P. P. Maanen. 2013. A framework for explaining reliance on decision aids. *Int. J. Hum. Comput. Stud.* 71, 4 (2013), 410–424. DOI: https://doi.org/10.1016/j.ijhcs.2012.10.018

[15] N. Du. 2019. Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transp. Res. Part C Emerg. Technol.* 104, 2018 (2019), 428–442. DOI: https://doi.org/10.1016/j.trc.2019.05.025

[16] T. Ellwart and N. Schauffel. 2023. *Human-Autonomy Teaming in Ship Inspection: Psychological Perspectives on the Collaboration between Humans and Self-Governing Systems*. Springer International Publishing.

[17] C. Esterwood and L. P. Robert. 2023. Three strikes and you are out!: The impacts of multiple human-robot trust violations and repairs on robot trustworthiness. *Comput. Human Behav.* 142 (2023), 1–39.

[18] C. Esterwood and L. P. Robert. 2022. A literature review of trust repair in HRI. In *Proceedings of the 31st IEEE International Conference on Robot and Human Interactive Communication*.

[19] C. Esterwood and L. P. Robert. 2023. The theory of mind and human-robot trust repair. *Scientific Reports* 13 (2023). DOI: https://doi.org/10.1038/s41598-023-37032-0

[20] G. Ferguson and J. Allen. 2011. A cognitive model for collaborative agents. In *Proceedings of the AAAI Fall Symposium*. Techical Report, FS-11-01, 112–120, 2011.

[21] Y. Forster, F. Naujoks, and A. Neukum. 2017. Increasing anthropomorphism and trust in automated driving functions by adding speech output. *IEEE Intell. Veh. Symp. Proc.* 2, Iv (2017), 365–372. DOI: https://doi.org/10.1109/IVS.2017.7995746

[22] P. Fratczak, Y. M. Goh, P. Kinnell, L. Justham, and A. Soltoggio. 2020. Robot apology as a post-accident trust-recovery control strategy in industrial human-robot interaction. *Int. J. Ind. Ergon.* 82 (2020), 103078. DOI: https://doi.org/10.1016/j.ergon.2020.103078

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

[24] Y. Guo and X. J. Yang. 2020. Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach. *Int. J. Soc. Robot.* 13, 8 (2020), 1899. DOI: https://doi.org/10.1007/s12369-020-00703-3

[25] S. Guznov. 2020. Robot transparency and team orientation effects on human–robot teaming. *Int. J. Hum. Comput. Interact.* 36, 7 (2020), 650–660. DOI: https://doi.org/10.1080/10447318.2019.1676519

[26] K. Hald, K. Weitz, E. André, and M. Rehm. 2021. 'An error occurred!' - Trust repair with virtual robot using levels of mistake explanation. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, 9. Retrieved from http://journal.unilak.ac.id/index.php/JIEB/article/view/3845

[27] A. Hamacher, N. Bianchi-Berthouze, A. G. Pipe, and K. Eder. 2016. Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication*, 493–500. DOI: https://doi.org/10.1109/ROMAN.2016.7745163

[28] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. Visser, and R. Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors* 53, 5 (2011), 517–527. DOI: https://doi.org/10.1177/0018720811417254

[29] S. G. Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. *Proc. Hum. Factors Ergon. Soc.* (2006), 904–908. DOI: https://doi.org/10.1177/154193120605000909

[30] T. Helldin, G. Falkman, M. Riveiro, and S. Davidsson. 2013. Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 210–217. DOI: https://doi.org/10.1145/2516540.2516554

[31] C. A. Hidalgo, D. Orghian, J. Albo-Canals, F. Almeida, and N. Martin. 2021. *How Humans Judge Machines*. Massachusetts Institute of Technology: The MIT Press Cambridge, Massachusetts London, England.

[32] N. T. Ho. 2017. Application of human-autonomy teaming to an advanced ground station for reduced crew operations. In *Proceedings of the AIAA/IEEE Digital Avionics Systems Conference*, 9–12. DOI: https://doi.org/10.1109/DASC.2017.8102124

[33] P. Hock, J. Kraus, M. Walch, N. Lang, and M. Baumann. 2016. Elaborating feedback strategies for maintaining automation in highly automated driving. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 105–112. DOI: https://doi.org/10.1145/3003715.3005414

[34] K. A. Hoff and M. Bashir. 2015. "Trust in automation: Integrating empirical evidence on factors that influence trust. *Hum. Factors* 57, 3 (2015), 407–434. DOI: https://doi.org/10.1177/0018720814547570

[35] M. Hou, G. Ho, and D. Dunwoody. 2021. IMPACTS: A trust model for human-autonomy teaming. *Human-Intelligent Syst. Integr.* 3, 2 (2021), 79–97. DOI: https://doi.org/10.1007/s42454-020-00023-x

[36] C. C. Jorge, N. H. Bouman, C. M. Jonker, and M. L. Tielman. 2023. Exploring the effect of automation failure on the human's trustworthiness in human-agent teamwork. *Front. Robot. AI* 10 (2023), 1–14. DOI: https://doi.org/10.3389/frobt.2023.1143723

[37] T. Kim and P. J. Hinds. 2006. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *15th IEEE International Symposium on Robot and Human Interactive Communication* (2006), 80–85. DOI: https://doi.org/10.1109/ROMAN.2006.314398

[38] T. Kim and H. Song. 2021. How should intelligent agents apologize to restore trust?: The interaction effect between anthropomorphism and apology attribution on trust repair. *Telemat. Informatics* 61 (2021), 1–33.

[39] E. S. Kox, J. H. Kerstholt, T. Hueting, and P. W. Vries. 2021. Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Auton. Agent. Multi. Agent. Syst.* 35, 2 (2021), 1–20. DOI: https://doi.org/10.1007/s10458-021-09515-9

[40] E. S. Kox, L. B. Siegling, and J. H. Kerstholt. 2022. Trust development in military and civilian human-agent teams: The effect of social-cognitive recovery strategies. *Int. J. Soc. Robot.* 14, 5 (2022), 1323–1338. DOI: https://doi.org/10.1007/s12369-022-00871-4

[41] J. Kraus, D. Scholz, D. Stiegemeier, and M. Baumann. 2020. The more you know: Trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Hum. Factors* 62, 5 (2020), 718–736. DOI: https://doi.org/10.1177/0018720819853686

[42] J. D. Lee. 1991. The dynamics of trust in a supervisory control simulation. In *Proceedings of the Human Factors Society 35th Annual Meeting*, 1228–1232.

[43] J. D. Lee and K. A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80.

[44] M. K. Lee, S. Kiesler, J. Forlizzi, S. S. Srinivasa, and P. Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction,* 203–210. DOI: https://doi.org/10.1109/HRI.2010.5453195

[45] B. Lubars and C. Tan. 2019. Ask not what AI can do, but what AI should do: Towards a framework of task delegability. *Adv. Neural Inf. Process. Syst.* 32 (2019), 1–11.

[46] M. B. Luebbers, A. Tabrez, K. Ruvane, and B. Hayes. 2023. Autonomous justification for enabling explainable decision support in human-robot teaming. In *Proceedings of Robotics: Science and Systems*.

[47] J. B. Lyons. 2013. Being transparent about transparency: A model for human-robot interaction. *Proceedings of the AAAI Spring Symposium*, 48–53.

[48] J. B. Lyons, I. Hamdan, and T. Q. Vo. 2023. Explanations and trust: What happens to trust when a robot partner does something unexpected? *Comput. Human Behav.* 138, 2022 (2023), 107473. DOI: https://doi.org/10.1016/j.chb.2022.107473

[49] N. Lyu, L. Xie, C. Wu, Q. Fu, and C. Deng. 2017. Driver's cognitive workload and driving performance under traffic sign information exposure in complex environments: A case study of the highways in China. *Int. J. Environ. Res. Public Health* 14, 2 (2017), 1–25. DOI: https://doi.org/10.3390/ijerph14020203

[50] M. Madsen and S. Gregor. 2000. Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems,* 6–8, 2000. Retrieved from 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=

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

[52] R. C. Mayer, J. H. Davis, and D. F. Schoorman. 1995. An integrative model of organizational trust. *Acad. Manag. Rev.* 20, 3 (1995), 709–734. DOI: https://doi.org/10.1109/GLOCOM.2017.8254064

[53] D. H. McKnight, V. Choudhury, and C. Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology. *Inf. Syst. Res.* 13, 3 (2002), 334–359. DOI: https://doi.org/10.1287/isre.13.3.334.81

[54] J. E. Mercado, M. A. Rupp, J. Y. C. Chen, M. J. Barnes, D. Barber, and K. Procci. 2016. Intelligent agent transparency in human–agent teaming for multi-UxV management. *Hum. Factors J. Hum. Factors Ergon. Soc.* 58, 3 (2016), 401–415. DOI: https://doi.org/10.1177/0018720815621206

[55] K. Merwe, S. Mallam, and S. Nazir. 2022. Agent transparency, situation awareness, mental workload, and operator performance: A systematic literature review. *Hum. Factors* 2022 (2022). DOI: https://doi.org/10.1177/00187208221077804

[56] J. S. Metcalfe and J. Diggelen. 2021. Design considerations for future human-AI ecosystems. *HHAI* '21, June (2021).

[57] C. A. Miller. 2014. 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)*, 191–202. DOI: https://doi.org/10.1007/978-3-319-07458-0_8

[58] C. A. Miller. 2020. *Trust, Transparency, Explanation, and Planning: Why We Need a Lifecycle Perspective on Human-Automation Interaction*. Elsevier Inc.

[59] C. A. Miller. 2023. LifeCycle transparency: Why, and how, transparency information exchange should be distributed throughout the life of technology usage participant biographies. In *Proceedings of the 67th Human Factors and Ergonomics Society International Annual Meeting*. DOI: https://doi.org/10.1177/21695067231192272

[60] C. A. Miller and R. Parasuraman. 2007. Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Hum. Factors* 49, 1 (2007), 57–75. DOI: https://doi.org/10.1518/001872007779598037

[61] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi. 2017. "To err is robot: How humans assess and act toward an erroneous social robot. *Front. Robot. AI* 4, May (2017), 1–15. DOI: https://doi.org/10.3389/frobt.2017.00021

[62] T. O', N. J. McNeese Neill, A. Barron, and B. G. Schelble. 2022. Human–autonomy teaming: A review and analysis of the empirical literature. *Hum. Factors* 64, 5 (2022), 904–938. DOI: https://doi.org/10.1177/0018720820960865

[63] S. Ososky, T. L. Sanders, F. G. Jentsch, P. A. Hancock, and J .Y. C. Chen. 2014. Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Proceedings of the SPIE Unmanned Systems Technology XVI Conference*, Vol. 9084, 90840. DOI: https://doi.org/10.1117/12.2050622

[64] A. R. Panganiban, M. D. Long, and G. Matthews. 2020. Human Machine Teaming (HMT): Trust Cues in Communication and Bias Towards Robotic Partners. Retrieved from https://apps.dtic.mil/sti/citations/AD1121408Available:

[65] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. 2000. A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man. Cybern. A Syst. Hum.* 30, 3 (2000), 286–297. DOI: https://doi.org/10.1109/3468.844354

[66] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. 2008. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *J. Cogn. Eng. Decis. Mak.*, 2, 2 (2008), 140–160. DOI: https://doi.org/10.1518/155534308X284417

[67] S. K. Parker and G. Grote. 2022. Automation, algorithms, and beyond: Why work design matters more than ever in a digital world. *Appl. Psychol.* 71, 4 (2022), 1171–1204. DOI: https://doi.org/10.1111/apps.12241

[68] R. Perkins, Z. R. Khavas, K. McCallum, M. R. Kotturu, and P. Robinette. 2022. The reason for an apology matters for robot trust repair. In *Proceedings of the 14th International Conference on Social Robotics*, 640–651. DOI: https://doi.org/10.1007/978-3-031-24670-8

[69] E. Phillips, B. F. Malle, and V. B. Chi. 2023. Systematic methods for Moral HRI: Studying human responses to robot norm conflicts. *TBD*, 1, 1 (2023), 1–10. DOI: https://doi.org/10.31234/osf.io/by4rh

[70] M. Raue, L. A. D' Ambrosio, C. Ward, C. Lee, C. Jacquillat, and J. F. Coughlin. 2019. The influence of feelings while driving regular cars on the perception and acceptance of self-driving cars. *Risk Anal.* 39, 2 (2019), 358–374. DOI: https://doi.org/10.1111/risa.13267

[71] S. Rebensky, K. Carmody, C. Ficke, D. Nguyen, M. Carroll, J. Wildman, and A. Thayer. 2021. Something went wrong: Errors, trust and trust repair strategies in human agent teaming. In *Proceedings of the 2nd International Conference, AI-HCI 2021 Held as Part of the 23rd HCI International Conference*, Vol. 2, 95–106. DOI: https://doi.org/10.1016/j.gpb.2023.01.002

[72] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. 'Why should i trust you?' Explaining the predictions of any classifier. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Proceedings of Demonstration Session, 97–101. DOI: https://doi.org/10.18653/v1/n16-3020

[73] P. Robinette, A. M. Howard, and A. R. Wagner. 2017. Effect of robot performance on human-robot trust in time-critical situations. *IEEE Trans. Human-Machine Syst.* 47, 4 (2017), 425–436. DOI: https://doi.org/10.1109/THMS.2017.2648849

[74] A. Rossi, K. Dautenhahn, K.L. Koay, and M. L. Walters. 2018. The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn, Journal of Behavioral Robotics* 9, 1 (2018), 137–154. DOI: https://doi.org/10.1515/pjbr-2018-0010

[75] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn. 2015. "Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 141–148. DOI: https://doi.org/10.1145/2696454.2696497

[76] K. E. Schaefer, S. G. Hill, and F. G. Jentsch. 2018. Trust in human-autonomy teaming: a review of trust research from the US army research laboratory robotics collaborative technology alliance. In *Proceedings of the AHFE International Conference on Human Factors in Robots and Unmanned Systems*, Vol. 784, 102–114. DOI: https://doi.org/10.1007/978-3-319-94346-6_8

[77] K. E. Schaefer, E. R. Straub, J. Y. C. Chen, J. Putney, and A. W. Evans. 2017. Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cogn. Syst. Res.* 46 (2017), 26–39. DOI: https://doi.org/10.1016/j.cogsys.2017.02.002

[78] J. Scholtz. 2003. Theory and evaluation of human robot interactions. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, IEEE, 1–10, Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1174284

[79] D. F. Schoorman, R. C. Mayer, and J. H. Davis. 2007. An integrative model of organizational trust: Past, present and future. *Acad. Manag. Rev.* 32, 2 (2007), 344–354. DOI: https://doi.org/10.5465/amr.2007.24348410

[80] S. S. Sebo, P. Krishnamurthi, and B. Scassellati. 2019. 'I don't believe you': Investigating the effects of robot trust violation and repair. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction* (2019), 57–65. DOI: https://doi.org/10.1109/HRI.2019.8673169

[81] A. R. Selkowitz, S. G. Lakhmani, and J. Y. C. Chen. 2017. Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cogn. Syst. Res.* 46 (2017), 13–25. DOI: https://doi.org/10.1016/j.cogsys.2017.02.003

[82] A. R. Selkowitz, S. G. Lakhmani, C. N. Larios, and J. Y. C. Chen. 2014. Agent transparency and the autonomous squad member. *Proc. Hum. Factors Ergon. Soc.* (2014), 1318–1322. DOI: https://doi.org/10.1177/1541931213601305

[83] A. Shariff, J. F. Bonnefon, and I. Rahwan. 2017. Psychological roadblocks to the adoption of self-driving vehicles. *Nat. Hum. Behav.* 1, 10 (2017), 694–696. DOI: https://doi.org/10.1038/s41562-017-0202-6

[84] T. B. Sheridan. 2019. Individual differences in attributes of trust in automation: Measurement and application to system design. *Front. Psychol.* 10, May (2019), 1–7. DOI: https://doi.org/10.3389/fpsyg.2019.01117

[85] V. K. Sims, M. G. Chin, H. C. Lum, L. Upham-Ellis, T. Ballion, and N. C. Lagattuta. 2009. Robots' auditory cues are subject to anthropomorphism. *Proc. Hum. Factors Ergon. Soc.* 3 (2009), 1418–1421. DOI: https://doi.org/10.1518/107118109x12524444079352

[86] S. Soltanzadeh. 2022. Strictly human: Limitations of autonomous systems. *Minds Mach.* 32 (2022), 269–288. DOI: https://doi.org/10.1007/s11023-021-09582-7

[87] K. Stowers, N. Kasdaglis, M. A. Rupp, O. B. Newton, J. Y. C. Chen, and M. J. Barnes. 2020. The IMPACT of agent transparency on human performance. *IEEE Trans. Human-Machine Syst.* 50, 3 (2020), 245–253. DOI: https://doi.org/10.1109/THMS.2020.2978041

[88] N .L. Tenhundfeld, E. J. Visser, A. J. Ries, V. S. Finomore, and C. C. Tossell. 2020. Trust and distrust of automated parking in a tesla model X. *Hum. Factors* 62, 2 (2020), 194–210. DOI: https://doi.org/10.1177/0018720819865412

[89] A. Turner, M. Kaushik, M.-T. Huang, and S. Varanasi. 2020. Calibrating Trust in AI-Assisted Decision Making.

[90] A. Tversky and D. Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *J. Risk Uncertain.* 5 (1992), 297–323. DOI: https://doi.org/10.15358/0340-1650-2006-6-331

[91] E. J. Visser, F. Krueger, P. McKnight, S. Scheid, M. Smith, S. Chalk, and R. Parasuraman. 2012. The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56, 1 (2012), 263–267. DOI: https://doi.org/10.1177/1071181312561062

[92] E. J. Visser. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *J. Exp. Psychol. Appl.* 22, 3 (2016), 331–349. DOI: https://doi.org/10.1037/xap0000092

[93] E. J. Visser, R. Pak, and T. H. Shaw. 2018. From "automation" to "autonomy": The importance of trust repair in human–machine interaction. Ergonomics 61, 1 (2018), 1409–1427. DOI: https://doi.org/10.1080/00140139.2018.1457725

[94] E. J. Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *Int. J. Soc. Robot.* 12, 2 (2020), 459–478. DOI: https://doi.org/10.1007/s12369-019-00596-x

[95] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill. 2018. Is it my looks? Or something i said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In *Persuasive Technology*. J. Ham, E. Karapanos, P. Morita, C. Burns (Eds.), Lecture Notes in Computer Science, Vol. 10809, 56–69. DOI: https://doi.org/10.1007/978-3-319-78978-1_5

[96] H. Westerbeek and A. Maes. 2013. Route-external and route-internal landmarks in route descriptions: Effects of route length and map design. *Appl. Cogn. Psychol.* 27, 3 (2013), 297–305. DOI: https://doi.org/10.1002/acp.2907

[97] M. Wischnewski, N. C. Krä, mer, E. Mü, and ller. 2023. *Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-of-the-Art and Future Directions*. Vol. 1, ACM, New York, NY.

[98] K. T. Wynne and J. B. Lyons. 2018. An integrative model of autonomous agent teammate-likeness. *Theor. Issues Ergon. Sci.* 19, 3 (2018), 353–374. DOI: https://doi.org/10.1080/1463922X.2016.1260181

[99] K. T. Wynne and J. B. Lyons. 2019. Autonomous agent teammate-likeness: Scale development and validation. In *Proceedings of the International Conference on Human-Computer Interaction*, 199–213.

[100] X. J. Yang, C. Schemanske, and C. Searle. 2023. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Hum. Factors* 65, 5 (2023), 862–878. DOI: https://doi.org/10.1177/00187208211034716

[101] Y. Ye, H. You, and J. Du. 2023. Improved trust in human-robot collaboration with ChatGPT. *IEEE Access* 11 (2023), 55748–55754. DOI: https://doi.org/10.1109/ACCESS.2023.3282111

Received 7 December 2023; revised 20 June 2024; accepted 17 October 2024