Uncertainty Communication by AI Assistants: The Effects on User Trust

Master Thesis

Universiteit Utrecht

Author:	**Lena Siegling**
Student number:	**6844286**
Email:	**lena.b.siegling@gmail.com**
Master's programme:	**Applied Cognitive Psychology**
Programme component:	**201800483, Master's Thesis, 27.5 ECTS**
Date of submission:	**29.10.2020**

Completed at the Netherlands Organization for Applied Scientific Research (TNO)

Utrecht University supervisor:	**Dr. P.W. Woźniak, p.w.wozniak@uu.nl**
External supervisors:	• **Prof. Dr. J.H. Kerstholt, jose.kerstholt@tno.nl** • **Dr. P.A.M. Ruijten, p.a.m.ruijten@tue.nl**

ACKNOWLEDGEMENTS

I would like to thank Prof. Dr. J.H. Kerstholt for enabling me to complete this project at TNO and for providing a flexible collaboration during the COVID-19 pandemic. I am also thankful for the helpful correspondence with Ms. E.S. Kox throughout the entirety of this project. The creation of the VR environment by Dr. J.S. Barnhoorn is also greatly appreciated.

I am furthermore grateful for the excellent guidance provided by Dr. P.A.M. Ruijten during the research design and data analysis. I would also like to thank Dr. P.W. Woźniak for supervising this thesis at Utrecht University. Dr. S.F. Donker provided great help in coordinating the initiation and conclusion of the project.

Finally, I would like to thank my friend Oliver and my cousin Lukas whose insights and expertise in their respective fields assisted me in crafting my research ideas. I thank my friends Luis and Kaja for proofreading this work.

DEDICATION

I dedicate this work to my kind mother. My curiosity, academic dedication, and resilience through challenges remain immensely aided by her support and the inspiration I receive from her daily brilliance.

ABSTRACT

As artificial intelligence (AI) rapidly spreads across multiple domains and becomes increasingly integrated into everyday life, user trust is vital to consider. Inappropriate user trust has resulted in fatal accidents and significantly stunts the opportunities that AI can offer. Given the uncertainty involved in AI inputs, processing and outputs, this study investigated the effects of communicating system uncertainty on users' trust in AI assistants. Trust development was repeatedly assessed whilst 64 participants completed an online search task that was guided by an AI drone. Drones either communicated uncertainty or not and deployed a trust repair strategy or not following a 2x2 mixed factorial design. The research also assessed if uncertainty communication enhanced the trust repair strategy and if it improved users' perception of and overall interaction with the AI drones. Results show that uncertainty communication significantly dampened the negative effects of an AI error by increasing users' situational awareness, understanding of the system, and sensitivity to AI fallibility. Participants preferred drones that communicated uncertainty as these were perceived to be more trustworthy and valuable. The trust repair strategy significantly repaired violated user trust, yet this effect was not enhanced by uncertainty communication. This research concludes that successful AI systems must: adapt with the fluidity of user trust, provide system transparency, maintain user agency, perform well, recognize past system performance, and empathetically acknowledge the user's emotional state throughout an interaction.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

1.0 Artificial Intelligence

Artificial intelligence (AI) describes the ability of machines to learn from data to automatically generate predictions, autonomous decisions, and interactions with the environment [1]. A key enabler of AI is machine learning, which uses algorithms and statistical modeling to allow computer systems to automatically improve by learning from experience [2], [3]. AI assists human analytical and decision-making skills to free users for higher-level tasks [4] and leisure [5].

Decreases in AI costs alongside increases in the availability of data [3], AI efficiency, performance [2] and sensing capabilities [6] continuously fuel rapid AI adoption in multiple domains. These include; healthcare, marketing, retail, financial services, news broadcasting, criminal justice systems, social media [2], military operations [7], transportation [8], manufacturing and education [3]. The rapid expansion of AI is shown in the 270% global growth of enterprises implementing AI from 2015-2019 [9]. Additionally, AI spending is expected to double between 2018-2022, reaching $79.2 billion in 2022 [10]. This confirms the relevance of exploring AI use.

1.1 The Importance of Trust in AI

Trust is a vital prerequisite for individuals and societies to effectively use, deploy, and develop AI [11]. A significant user proportion continues to lack trust and confidence in AI decisions, answers, and recommendations. Research by [12] confirmed 69% of respondents were more inclined to be truthful with a human rather than with an AI. Additionally, 86% of respondents trusted a human counterpart more than the AI with life or death decisions [12]. Similarly, [13] found 45.1% of surveyed US consumers to lack trust in any sort of AI. 40.5% of respondents expressed concern and 40.1% expressed skepticism towards AI [13].

This lack of trust in AI capabilities (i.e. AI undertrust) can result in inefficient monitoring, disuse[1], and avoidance of a system creating an unequal workload distribution between the user and the AI [14].

In contrast, Wright et al. [15] described “automation bias” as users’ perceptions of perfect machines resulting in a natural tendency to follow AI advice. Likewise, Thornhill [16] mentioned users’ “the computer can’t be wrong” mindset. He exemplified this with the fatal crash of a semi-automated Tesla in 2018, which [17] determined to have partially been caused by the driver’s overtrust in the system. Similarly, pilots’ continuous reliance on the autopilot of the Turkish Airlines Flight 1951 after the failure of an altitude measuring instrument resulted in the crash killing 9 people [18]. Other researchers [19], [20] confirmed the increased reliance on automation in high-risk situations and decision-making. Overestimating a system’s capabilities and placing too much trust in AI (i.e. AI overtrust) can result in complacent user states, misuse[2] [14], a lack of situational awareness [21] and AI mismanagement [22].

Following this, appropriate trust levels are crucial for safe and productive human-AI collaborations [18]. Therefore, this thesis explores user trust in AI assistants.

1.2 Understanding and Defining Trust in AI

Prior to exploration, it is important to understand the concept of trust. The general consensus is that trust describes a willingness to be vulnerable [23] and an expectation regarding a behavior or outcome [24] in a cooperation that is uncertain and risky [18]. Lee and See [24] reflect this in their definition of trust as “the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability”. They also highlight that trust may be a belief, attitude, intention or behavior [24]. This aligns with Lewicki and Brinsfield’s [25] claim of cognitive, behavioural,

1. Disuse: Failures in collaboration resulting from the user’s rejection of AI capabilities [24].
2. Misuse: Collaboration failures resulting from the user’s violation of assumptions regarding AI capabilities [24].

and affective components of trust. Madsen and Gregor [26] defined human-computer trust as "the extent to which a user is confident in, and willing to act on the basis of the recommendations, actions, and decisions of an artificially intelligent decision aid".

Following this, the researcher deduces that trust in an AI assistant describes the user's willingness to be in a vulnerable position in which they act on AI decisions and recommendations, with the expectation of predictably achieving their goal in an uncertain context.

1.3 The Trust Life Cycle

User trust is built and (re)adjusted as more information and variable AI performance are presented throughout collaboration [18]. During adjustment, users update their perceived trust in response to the capabilities and actual trustworthiness of AI to minimize error [14]. Although subtle variations exist in the academic literature, this life cycle of trust generally follows; trust building, trust violation, and trust repair (see Figure 1).

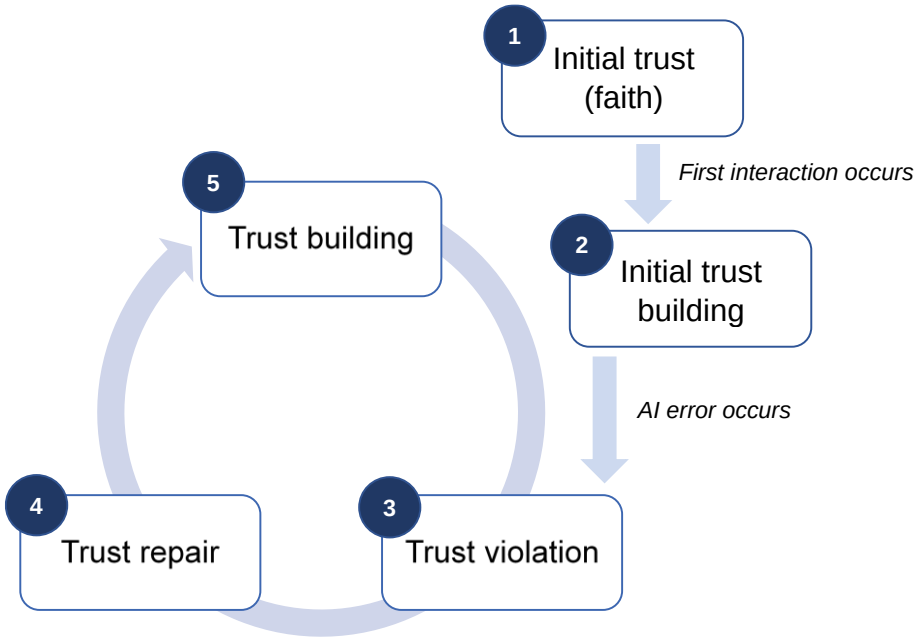


Figure 1: The user trust life cycle [25], [40]

Trust building is initially informed by existing user experiences with AI [25], a system's reputation [18], and user biases [27]. New users will have faith in the system, which is replaced by experiences of system predictability and dependability as the interaction proceeds [18]. The user will rely on their observations of AI behavior to facilitate trust building [24].

Faults in AI that produce unexpected or unwanted behavior and outcomes violate trust [14]. Trust subsequently declines (see Figure 2) as the user notices the misalignment between perceived AI trustworthiness and actual trustworthiness [14]. The negative effect of a trust violation depends on the initial reliability of a system [24] as well as on the timing [18], severity and frequency of the AI error [25]. Trust violations in the early stages of an interaction are more detrimental as initial trust is more fragile [18], [25]. However, trust violations occurring in later stages evoke a greater sense of betrayal [25].

To restore trust, a trust repair strategy can be deployed [14]. This can include the system taking accountability for its error, denying the error, providing an explanation, or apologizing [25].

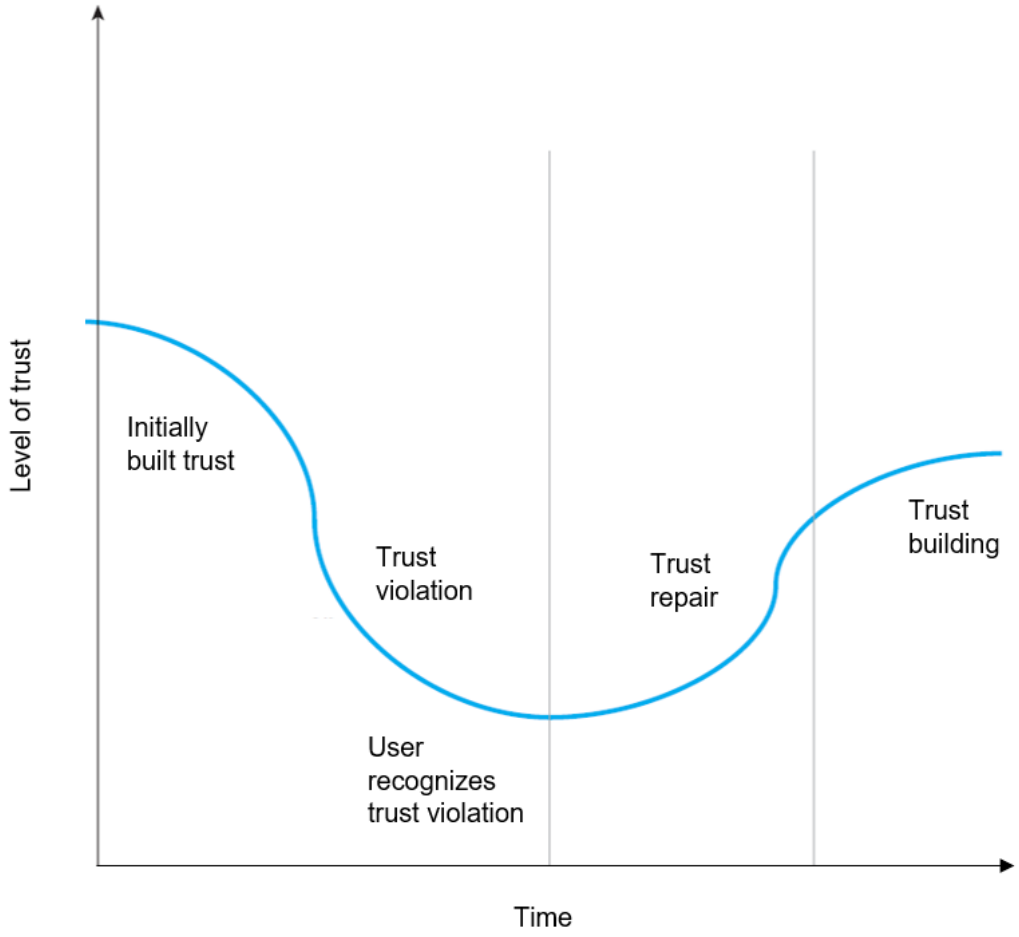


Figure 2: The longitudinal development of user trust [25]

1.4 Uncertainty in AI

Multiple factors relating to the user, the system and the collaborative environment guide the trust life cycle (see Appendix A). A meta-analysis revealed that system-related factors, especially system performance, were most influential to user trust [22].

The system factor considered in this study is AI uncertainty. Uncertainty implies doubt and a "lack of exact knowledge" [28]. AI uncertainty accumulates during incorrect or incomplete data acquisition from ambiguous contexts, data transformation, and output generation [21] (see Figure 3). This uncertainty can result in unexpected or erratic system behavior to violate user trust [21].

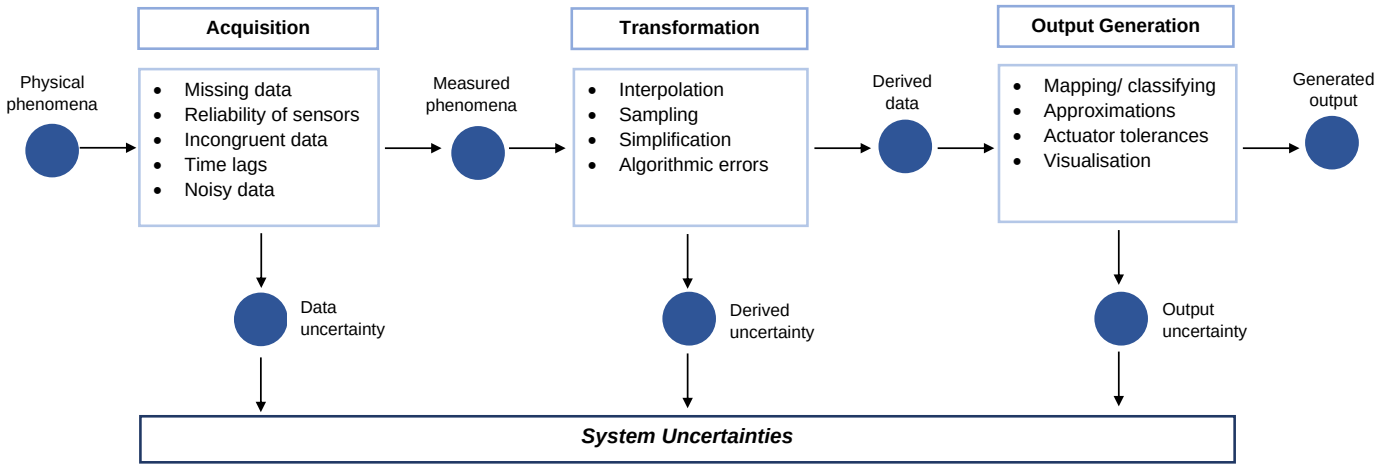


Figure 3: Sources of AI uncertainty [21]

The ability of modern AI to provide quantitative estimates of system uncertainty [29] enables uncertainty communication via probability distributions, confidence intervals, likelihood ratios, verbal summaries [30], hit and correct rejection rates as well as via predictive values [31].

1.5 Effects of Uncertainty Communication on Trust

Communicating uncertainty contributes to explainable AI; a concept that promotes increasing the user's understanding of a system's actions and predictions to benefit calibrated user trust [32], [33], [34]. Wang, Jamieson and Hollands [35] confirmed this benefit as they identified increasingly

appropriate user responses, reliance, and trust adjustments following reliability disclosure by a combat identification aid. Kunze et al.'s [21] analysis using a simulated automated driving system found that communicating system uncertainty helped drivers to calibrate their trust. Additionally, uncertainty communication increased situational awareness which promoted safer system takeovers by the drivers [21]. Antifakos, Schwaninger and Schiele [36] found substantial user performance increases in a memory task when the uncertainty of a memory aid was displayed. Uncertainty disclosure also helped users to better understand AI actions and performance [36]. Schaekermann et al. [29] found that ambiguity communication in AI assistants for clinical reasoning increased the perceived integrity of the system and users' confidence. Uncertainty communication also helped medical experts to appropriately reassess their trust in the AI for each medical case [29].

Despite these benefits, it is important to recognize that humans continue to feel aversion towards uncertainty [30]. Uncertainty and trust display a volatile relationship, as the admittance of uncertainty and system limitations can hinder trust, yet non-disclosure of uncertainty can equally undermine it [30], [29]. Kunze et al. [21] caution that the communication of system uncertainties may unease the user, who consequently becomes aware of system fallibilities. Strohkorb Sebo et al. [37] described a "ripple effect" by which robots who expressed their vulnerabilities caused their human teammates to express vulnerability as well. Thus, disclosing system uncertainty may also cause uncertainty in the user. Following this, it is crucial to strike the correct balance between disclosing and withholding uncertainty information [29].

RESEARCH RATIONALE AND OBJECTIVES

Despite the rapid growth of AI, the importance of user trust, and the significant impact of uncertainty disclosure on human-automation collaborations, explorations of the effects of uncertainty communication on trust in AI remain scarce [29]. The following research objectives identify if, how, and why uncertainty communication by AI drones can benefit user trust. Drones are flying robots [38] which observe, inspect, measure, and monitor environments [39] or transport loads. As the need for drones and the development of autonomous drone systems continue to grow [38], this study valuably considers a budding sector of AI robotics.

2.0 Objective 1

By assessing momentary states of user trust previous research [22], [40], [41] presents a critical knowledge gap, as the fluidity of trust development is not acknowledged. Yang et al. [41] criticize that the assessment of user trust at the end of a collaboration does not adequately reflect the user's trust over the entire interaction. Following this, knowledge of human-AI trust formation, maintenance [8], erosion [22] and repair remains scarce. Studies that do consider violated and repaired trust have been described as "commonly outdated" [8].

Therefore, the main objective of this study is:

- *Objective 1: Investigate the effects of uncertainty communication by an AI drone on the life cycle of user trust*

As illustrated in Figures 1 and 2, the user trust life cycle includes an important stage of trust repair. [42] and [8] found this stage to be significantly improved when an expression of regret, an apology, and an explanation (i.e. a trust repair strategy, detailed in section 3.1.2) were provided by autonomous systems. Throughout the achievement of objective 1, this research also aims to answer

whether uncertainty communication can enhance the effectiveness of a trust repair strategy and if it can function as a trust repair strategy on its own.

2.1 Objective 2

To valuably extend this research, the secondary objective is:

- *Objective 2: Determine if and why uncertainty communication adds perceived value to the system and the interaction with the system*

Determining if and how users perceive the value of uncertainty communication provides insights into users' experiences of AI uncertainty. If users' expectations regarding the interaction experience and system usability[3] are not met, user satisfaction and the willingness to use a system in the future are reduced [43]. Thus, uncertainty communication may provide a useful strategy to manage user expectations and trust such as to improve the user experience[4] [44] and acceptance of AI [43]. This is supported by previous findings of electric car drivers reporting an improved driving experience following the ambiguous display of range and state-of-charge information [44].

3. Usability: The extent to which a system, product, or service can be used by specific users to effectively, efficiently and satisfactorily achieve specific goals in specific context [90].
4. User experience: "a person's perceptions and responses resulting from the use and/ or anticipated use of a product, system or service" [91].

METHOD

3.0 Participants

An international cohort of 64 (33 male, 30 female, 1 unspecified) participants aged 18-30 years (*M* = 24.56, *SD* = 2.72), completed this experiment. 38 participants were students, 21 were working professionals and 5 selected “other”. Participants were able to understand English well and had normal/ corrected-to-normal vision. To successfully participate, individuals were required to have access to a laptop, tablet, or desktop PC which could play sound and video and was provided with a stable internet connection. Recruitment was done online[5] for 2 weeks.

3.1 Experimental Setting

3.1.1 Simulated Environment

For this experiment, footage of the researcher’s virtual reality (VR) glasses was used. The VR environment was created at the Netherlands Organisation for Applied Scientific Research (TNO) in Unity3D. The footage displayed walks through two abandoned houses (House A and House B) which each had 3 floors. The first and second floor of each house presented dangers to participants (see Appendix B). The videos were edited using the Windows 10 Video Editor and HandBrake. At the entrance of a house, the participant saw the drone flying away (Figure 4). This marked the beginning

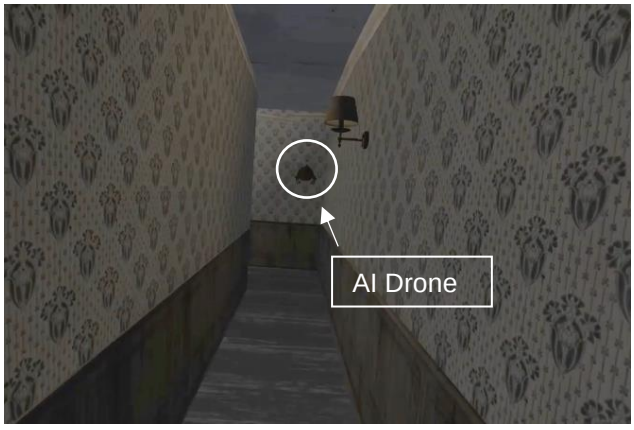


Figure 4: Screenshot from the experiment; at the beginning of a house the drone flew away

Figure 5: Screenshot from the experiment; the end of a floor was indicated by a staircase

5. Recruitment was done via Facebook, LinkedIn, Instagram, SurveySwap and PollPool

of an experimental run. The end of a floor was indicated to a participant as they reached a staircase (Figure 5).

In the simulated environment an AI drone communicated with participants via audio messages. These were initiated with a 'beep'[6] sound and created using the Free Text to Speech Software by Wideo[7]. The contents of the audio messages are specified in section 3.1.2 below. A different AI drone guided the walk through each house. The experiment was distributed via and completed within Qualtrics.

3.1.2 Description of the AI Drones

Each participant interacted with two different AI drones. Overall, the experiment included four different drones to create the experimental design.

The drones introduced themselves, provided instructions to help participants overcome dangers, and offered advice about how participants should proceed with respect to the safety of an environment. The drones varied in whether they communicated system uncertainty or not and if they deployed a trust repair strategy or not. AI drone characteristics are summarized in Table 1. An example of the communication from Drone sA1 in House A is shown in Table 2 (to view all scripts see Appendix C).

AI Name	*Characteristics*
Drone sA1	Provides a certainty statistic. Implements trust repair.
Drone sA2	Provides no certainty statistic. Implements trust repair.
Drone sA3	Provides a certainty statistic. Does not implement trust repair.
Drone sA4	Provides no certainty statistic. Does not implement trust repair.

Table 1: AI drone characteristics

6. "Beep-07" was downloaded from https://www.soundjay.com/beep-sounds-1.html.
7. Text was converted to speech using https://wideo.co/text-to-speech/. The "[en-US] Jack Bailey-S" voice was used at speed dial "1".

Description	*Point of reception*	*Message content*
Introduction	Floor 1	*"Hello I am Drone sA1. Starting area scan."*
1st advice		*"Warning, danger detected in this environment with 80% certainty. I advise you to proceed carefully."*
1st instruction		*"Laser trap detected in the next corridor; controls have been located next to the trap."*
2nd instruction		*"Stop. Cut the blue wire with your knife to deactivate the laser trap."*
3rd instruction		*"Laser trap deactivated, continue."*
2nd advice	Floor 2	*"Okay, clearance detected for this environment with 70% certainty. I advise you to move forward."*
Trust repair		*"Incorrect advice due to faulty signal from infrared camera. I am sorry this put you in danger."*
3rd advice	Floor 3	*"Okay, clearance detected for this environment with 75% certainty. I advise you to move forward."*

Table 2: Script for audio messages received by Drone sA1 in House A

Drones communicated AI uncertainty by providing a certainty statistic ranging from 70-80%. This range was based on previous benchmarks set by [41] and [43].

To enhance trust repair, some AI drones acknowledged their incorrect advice, explained the cause of their mistake, and apologized for putting the participant in danger. Incorrect advice was explained by faulty signals from the drones' sensors, representing a competence-based trust violation[8]. This trust repair strategy is inspired by previous research which consistently found the acknowledgement of system failure, explanation of this failure, and the provision of an apology to be most effective for trust repair [18], [25], [42], [45], [46].

3.2 Experimental Design

This experiment employed a 2 (uncertainty communication: yes/no) x 2 (trust repair strategy: yes/no) mixed factorial design. Uncertainty communication was manipulated within groups, the

8. Competence-based trust violation: A violation of user trust caused by a system's malfunction, unreliability or inability to complete a task [8]. In this case, the AI drone failed to detect danger and was therefore not able to provide correct advice.

deployment of a trust repair strategy was manipulated between groups. Participants were randomly allocated to one of the two experimental groups. Each of the experimental conditions represented one walk through a house and 32 participants completed each condition.

The order of experiencing conditions/ the order in which participants interacted with different drones was counterbalanced. The house (House A or B) in which an experimental condition was experienced was also counterbalanced

3.3 Measures

User trust, comparative trust, user preference, and perceived added value of uncertainty communication were measured using questionnaires (QSTR). The order of questionnaire items was randomized, and some questions were inversely formulated to reduce response bias.

3.3.1 User Trust

User trust describes the trust that participants had in the AI drone they interacted with and can be measured by assessing perceptions of system trustworthiness [47]. System trustworthiness has previously been measured using Mayer, Davis and Schoorman's [48] dimensions of perceived ability[9], benevolence[10], and integrity[11] [42], [47]. Lee and See [24] refined these dimensions into system performance[12], process[13], and purpose[14]. Körber [47] described the relationship between these dimensions as shown in Figure 6.

9. System ability: The skills, characteristics and competencies which enable a system to perform [47], [48].
10. System benevolence: The extent to which the system is perceived to "want to do good" and act in favor of the user [47], [48].
11. System integrity: Reflects a drone's consistent adherence to acceptable principles [47], [48].
12. System performance: "The current and previous operation" of a system. This includes competency, ability, and reliability of the system [47].
13. System process: Considers "how the system operates and if this modus operandi is appropriate for the situation and the operator's goals" [47].
14. System purpose: Considers the "intention in the system's design", the positive orientation of the design towards the user and the extent to which a system is used as designers intended [47].

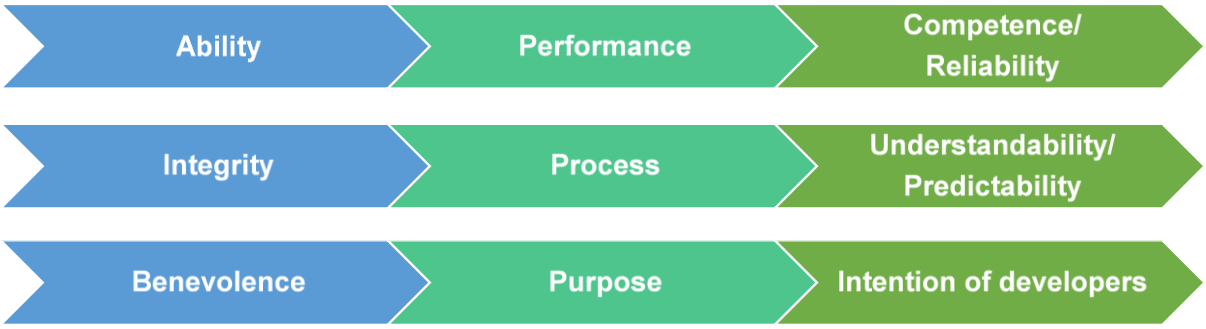


Figure 6: Relationships between trust dimensions [47]

Following this, participants' perceptions of AI drones'; (1) competence and reliability in performing the task, (2) understandability and predictability of behavior and advice, (3) positive intention toward the user were assessed.

Dimensions were rated on six-point Likert scales (1 = strongly disagree, 6 = strongly agree, no neutral mid-point) included in an 8-item questionnaire (QSTR2). QSTR2 is a custom scale based on questionnaires measuring user trust in robots [49] artificial aids [33], [42] and automated systems [47], [50 – 52]. QSTR2 has been specifically developed to suit the online setting of the experiment and enable fast repeated trust assessments.

Each participant completed QSTR2 seven times; once before they started any experimental run and three times during each experimental run. Prior to any experimental run, QSTR2 assessed participants' propensity to trust AI. During experimental runs, QSTR2 assessed initially developed trust, violated trust, and repaired trust. Questionnaire administration is further detailed in section 3.4. Contents of QSTR2 are detailed in Appendix D.

3.3.2 Comparative Trust

Comparative trust measured which drone participants trusted more after interacting with two different systems. Comparative trust was measured by asking "Which drone do you trust more?"

QSTR3). Participants could select one of two drones or indicate that they were undecided. An open question asked participants to reason their choice.

3.3.3 User Preference

User preference measured which drone participants preferred interacting with to explore whether uncertainty communication could create a preferable user experience. User preference was measured by asking “Which drone do you prefer?” (QSTR3). Participants could select one of two drones or indicate that they were undecided. An open question asked participants to reason their choice.

3.3.4 Perceived Value of Uncertainty Communication

Perceived added value measured the value that uncertainty communication could add to different aspects of the AI system. This construct was measured as a summative score based on answers to QSTR3 (see Appendix E). This custom questionnaire is inspired by Schrepp, Hinderks and Thomaschewski’s [53] shortened version of the user experience questionnaire [54].

3.4 Procedure

Prior to the experiment, participants were given information about the study and gave consent to the use of their anonymized data (Appendix F). Participants were informed that they would be completing a search mission which was assisted by an AI drone. They knew about the drones’ purpose but were blinded to manipulations of deploying a trust repair strategy and communicating uncertainty. Participants were also not informed of the AI drones’ fallibility. This was done to reduce initial biases towards the drones. Demographic information was collected via QSTR1 (Appendix G).

The administration of instructions and questionnaires, and the experience of the simulated environment when interacting with drones sA1 and sA2 is summarized in Table 3. As shown, drones introduced themselves and notified participants of initiating an area scan on the first floor of each house. This was followed by provisions of advice regarding the safety of an area and how participants should proceed. One such advice was provided on each floor of a house. The advice provided on the first floor was correct. The advice provided on the second floor was incorrect, resulting in a trust violation. Some drones deployed a trust repair strategy after the trust violation on the second floor. The advice given on the third floor was not followed by an event, so participants were not provided with feedback on whether the advice was correct. Each drone also gave instructions on the first floor of a house to guide participants in deactivating a laser trap (House A) or cutting a safety ribbon (House B).

Steps 5-25 were repeated twice by each participant, generating two experimental runs. A walk was paused each time a voice message from a drone was received. Each walk was separated into four videos, which participants had to click to play. Some videos were separated by instruction screens or the administration of QSTR2.

1.		Participant information and consent	
2.		QSTR1	
3.		QSTR2	
4.		Instructions	
		Drone sA1, House A	*Drone sA2, House B*
5.		Drone introduction	
6.		Start on floor 1	
7.		Correct advice received • Uncertainty communication	Correct advice received • No uncertainty communication
8.		Continue on floor 1	
9.		Danger detected; first instruction given	

10.		Continue on floor 1	
11.		Obstacle encountered	
12.		Second and third instruction given, obstacle resolved	
13.		Complete floor 1	
14.		QSTR2	
15.		Start on floor 2	
16.		Incorrect advice received • Uncertainty communication	Incorrect advice received • No uncertainty communication
17.		Continue on floor 2	
18.		Obstacle encountered	
19.		QSTR2	
		Continue on floor 2	
20.		Trust repair strategy	
21.		Complete floor 2, start on floor 3	
22.		Advice received	
23.		QSTR2	
25.		Instructions	
		Experimental runs completed	
26.		QSTR3	
27.		Experiment completed, debriefing	

Table 3: Summary of procedural steps[15]

This procedure was altered for participants interacting with drones sA3 and sA4. No trust repair strategy was deployed by these drones.

15. Grey boxes represent a screen that displayed instructions. Procedural changes regarding interactions with different drones are listed in the Drone sA2 column. Otherwise, all procedural steps remained the same for both interactions.

RESULTS AND ANALYSIS

Cronbach's α indicates good scale reliability of QSTR2 used to assess prior (α = .82), initial (α = .91), violated (α = .89) and repaired trust (α =.90)[16]. To analyze responses, participants' ratings of the eight questionnaire items were summed to create a total trust score for each participant. A total trust score was calculated each time QSTR2 was administered.

The presented analyses follow a simple statistical approach to determine the effect of the independent variables (uncertainty communication and trust repair strategy) on multiple dependent variables. The relationship between the independent variables has also been analyzed. An analysis of the relationship between dependent variables is beyond the scope of this thesis. It should also be noted that the analyses apply no Bonferroni correction. This is reasoned by the conservativeness of this post hoc test which reduces statistical power and increases the rate of type II errors. Previous research hints that the deployment of a trust repair strategy and uncertainty communication will benefit participants' trust in the AI drones. In the present experimental context participants are being guided through a high-risk scenario, where appropriate trust is vital. Missing to identify significant effects that could alter a trustful interaction would therefore be very costly. Hence, no correction has been applied to reduce costly type II errors.

4.0 Effects on the Life Cycle of User Trust

Figure 7 provides an overview of the relationship between the prior, initial, violated, and repaired trust scores for each experimental condition. The significance of uncertainty communication and deployment of a trust repair strategy within this life cycle are explored in the following sections.

16. Calculations were performed for the condition in which a trust repair strategy and uncertainty communication were present. As the same questionnaire was used in all conditions, the researcher assumes good reliability of QSTR2 for the entire experiment.

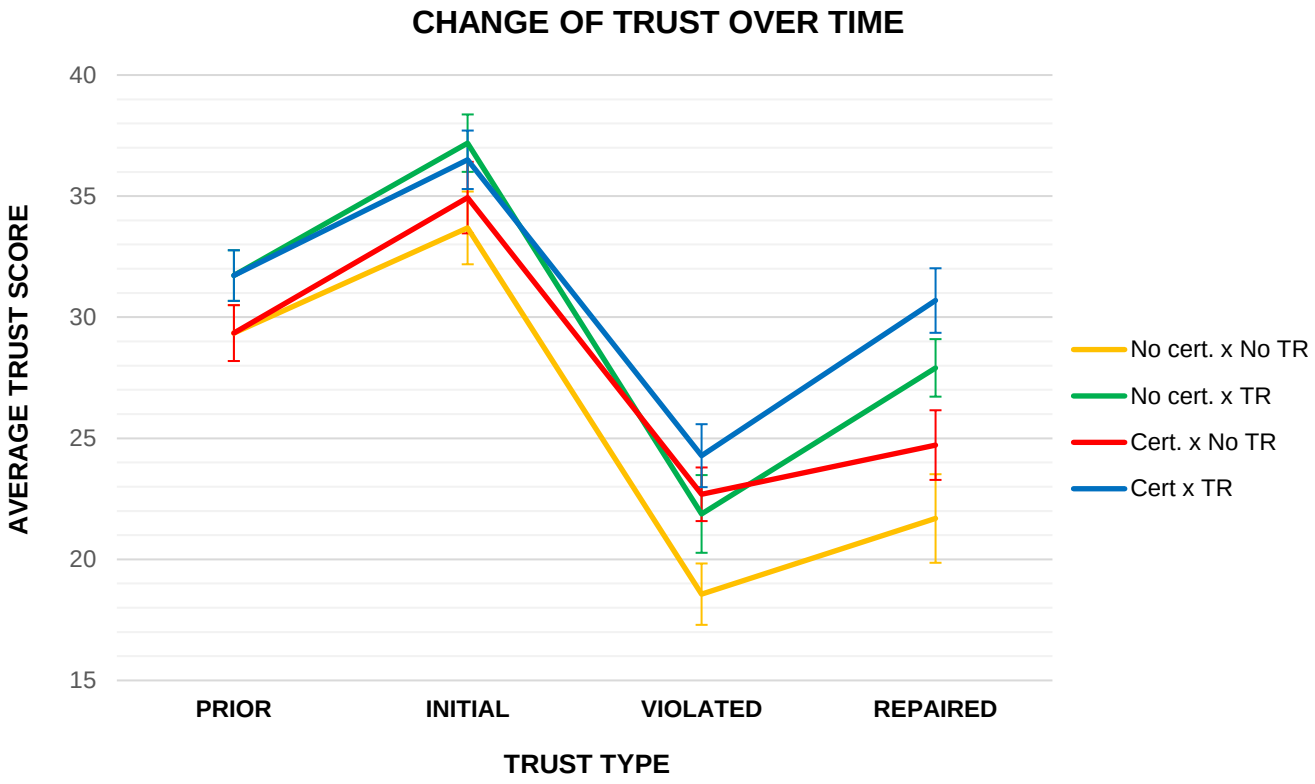


Figure 7: Trust scores during human-AI interaction. Cert. represents uncertainty communication; TR represents deployment of a trust repair strategy

4.1 Static Trust Measurements

4.1.1 Effects of Uncertainty Communication on Initial Trust

As no trust repair strategy had been deployed when initial trust was measured, both experimental groups were considered together. On average, interactions with drones that communicated uncertainty (sA1, sA3) resulted in higher initial trust scores (M = 35.72, SE = .92) than interactions with drones that did not communicate uncertainty (sA2, sA4) (M = 35.44, SE = .98). This difference, .28, BCa 95% CI [-1.59, 2.32] is however not significant $t(63) = .28$, $p = .780$, Cohen's d = .04, 95% CI [-.21, .28]. Hence, this analysis found no evidence that uncertainty communication had a significant effect on participants' initial trust.

4.1.2 Effects of Uncertainty Communication on Violated Trust

As no trust repair strategy had been deployed when violated trust was measured, both experimental groups were considered together. On average, interactions with drones that communicated uncertainty (sA1, sA3) resulted in higher trust scores after a violation (M = 23.48, SE = .82) than interactions with drones that did not communicate uncertainty (sA2, sA4) (M = 20.22, SE = 1.02). This difference, 3.26, BCa 95% CI [1.47, 5.09] is significant $t(63) = 3.35$, $p = .001$, Cohen's d = .42, 95% CI [.16, .67]. Therefore, communication of uncertainty resulted in significantly higher trust after a trust violation had occurred.

4.1.3 Effects of Uncertainty Communication and Trust Repair on Repaired Trust

Half of the participants received a trust repair strategy from a drone prior to measuring repaired trust. Therefore, experimental groups were considered separately. A 2 (uncertainty communication) x 2 (trust repair strategy) mixed factorial ANOVA yielded a significant main effect of uncertainty communication, $F(1,62) = 5.86$, $p = .018$, $\eta_p^2 = .09$, 90% CI [.01, .21]. A significant effect of deploying trust repair $F(1, 62) = 12.99$, $p = .001$, $\eta_p^2 = .17$, 90% CI [.05, .31] was also found. However, the interaction between uncertainty communication and a trust repair strategy was found to be non-significant $F(1, 62) = .01$, $p = .917$, $\eta_p^2 = .00$, 90% CI [.00, .01]. This indicates that repaired trust scores are increased by uncertainty communication and the trust repair strategy acting independently. Partial Eta Squared (η_p^2) suggests that deploying a trust repair strategy had a larger effect than uncertainty communication. This is illustrated in Figure 8.

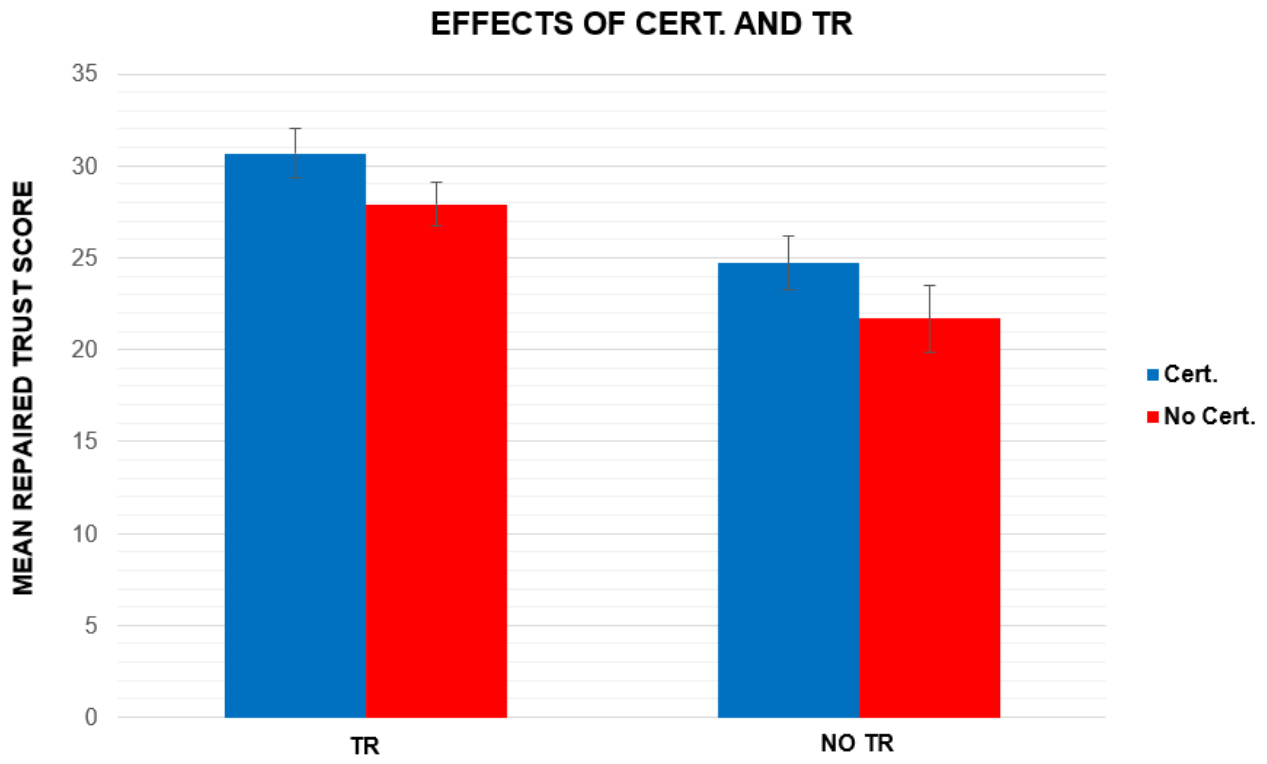


Figure 8: Effects of uncertainty communication (cert.) and a trust repair strategy (TR) on mean repaired trust scores

4.2 Trust Changes

4.2.1 Can Uncertainty Communication Aid Trust Establishment?

As no trust repair strategy had been deployed when prior and initial trust were measured, both experimental groups were considered together. On average, the increase in trust after participants' initial interaction with the drones was larger when drones communicated uncertainty (sA1, sA3) (M = 5.08, SE = .90) than when they did not (sA2, sA4) (M = 4.80, SE = .87). However, this difference, .28, BCa 95% CI [-1.43, 2.36] is not significant $t(63) = .28$, $p = .780$, Cohen's d = .04, 95% CI [-.21, .28]. Hence, this analysis found no evidence for a significant effect of uncertainty communication within trust establishment. Trust establishment is illustrated in Figure 9.

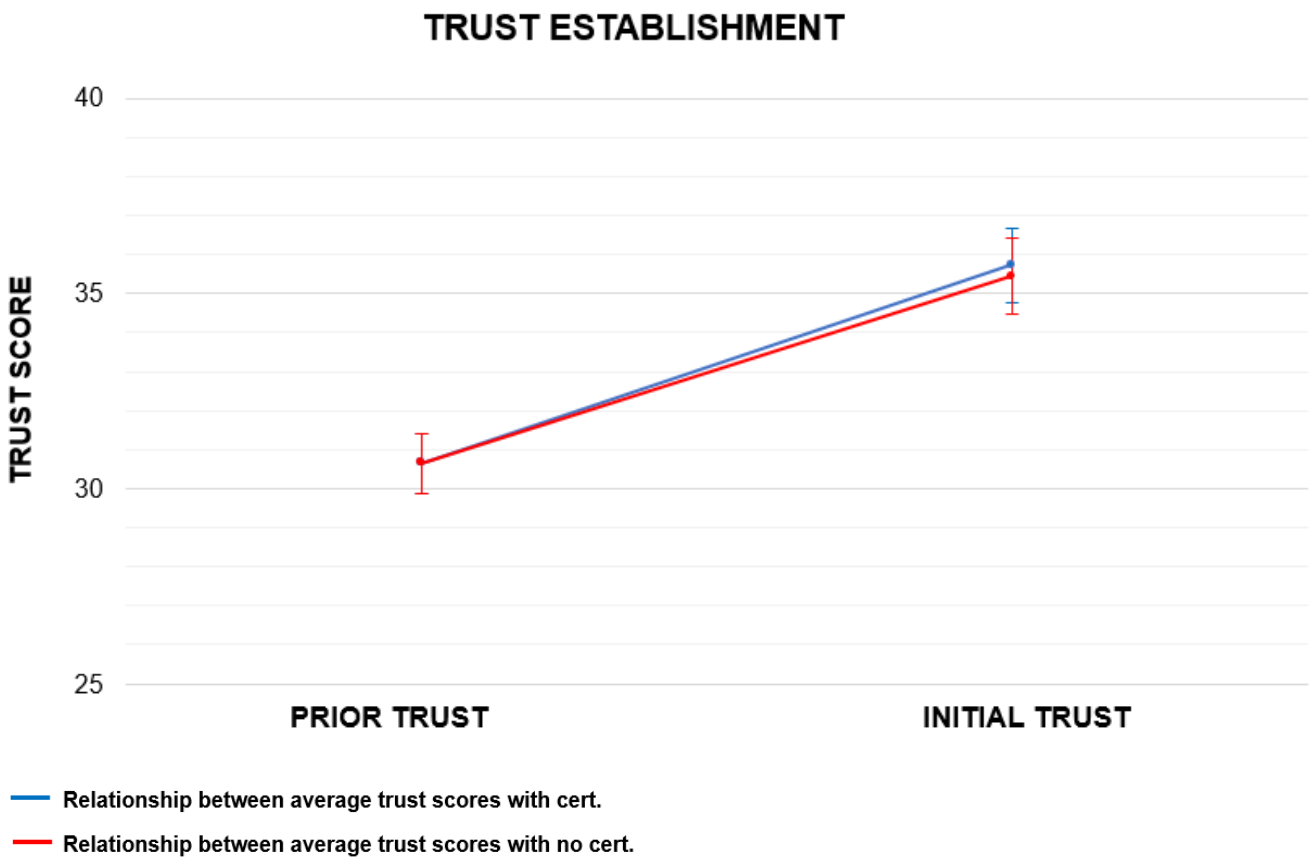


Figure 9: Average trust establishment. Cert. represents uncertainty communication

As can be seen from the graph, the relationships barely differ across uncertainty and no uncertainty communication conditions.

4.2.2 Can Uncertainty Communication Dampen the Effects of a Trust Violation?

As no trust repair strategy had been deployed when initial and violated trust were measured, both experimental groups were considered together. On average, the decrease in trust after a trust violation was greater in interactions with drones that did not communicate uncertainty (sA2, sA4) (M = -15.22, SE = 1.33) than with drones that did communicate uncertainty (sA1, sA3) (M = -12.23, SE = 1.16). This difference, 2.99, BCa 95% CI [.52, 5.35] is significant $t(63) = 2.33$, $p = .023$, Cohen's d = .29, 95% CI [.04, .54]. Hence, uncertainty communication can significantly dampen the decrease

in trust following a trust violation. Figure 10 visualizes the average decrease in trust amongst all participants.

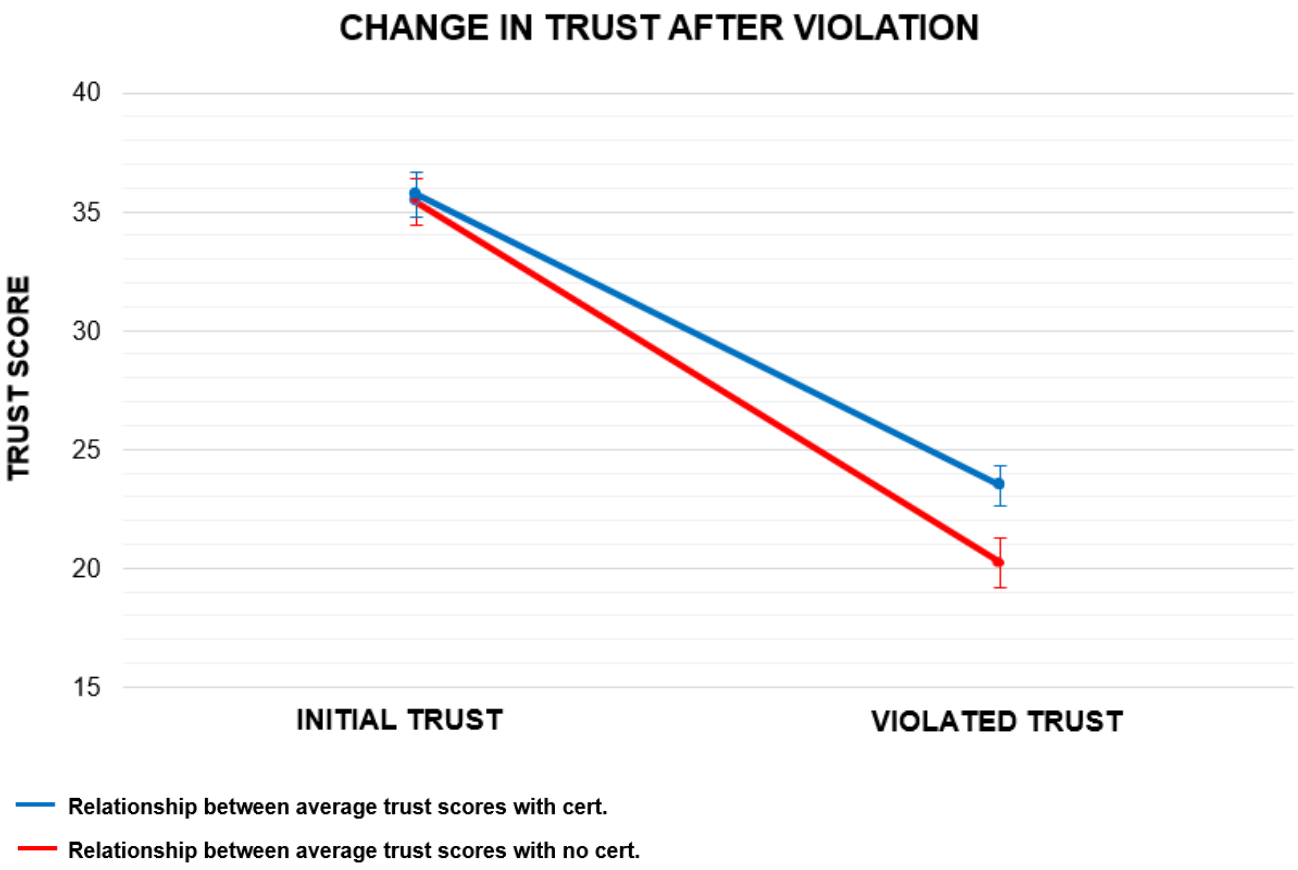


Figure 10: Average trust decline after a trust violation. Cert. represents uncertainty communication

4.2.3 Can Uncertainty Communication Aid Trust Reparation?

When considering trust reparation (i.e. the difference between violated and repaired trust), a 2 (uncertainty communication) x 2 (trust repair strategy) mixed factorial ANOVA revealed no significant main effect of uncertainty communication $F(1, 62) = .10$, $p = .758$, $\eta_p^2 = .00$, 90% CI [.00, .05]. The main effect of deploying a trust repair strategy was found to be significant $F(1, 62) = 5.39$, $p = .024$, $\eta_p^2 = .08$, 90% CI [.01, .20]. The interaction between uncertainty communication and the trust repair strategy yielded to be non-significant $F(1, 62) = .40$, $p = .529$, $\eta_p^2 = .01$, 90% CI [.00, .07]. Therefore, to repair trust after a violation, it is important for the drone to deploy a trust repair strategy.

Uncertainty communication on its own was not found to repair trust. Figure 11 visualizes the significant trust reparation when a trust repair strategy was deployed.

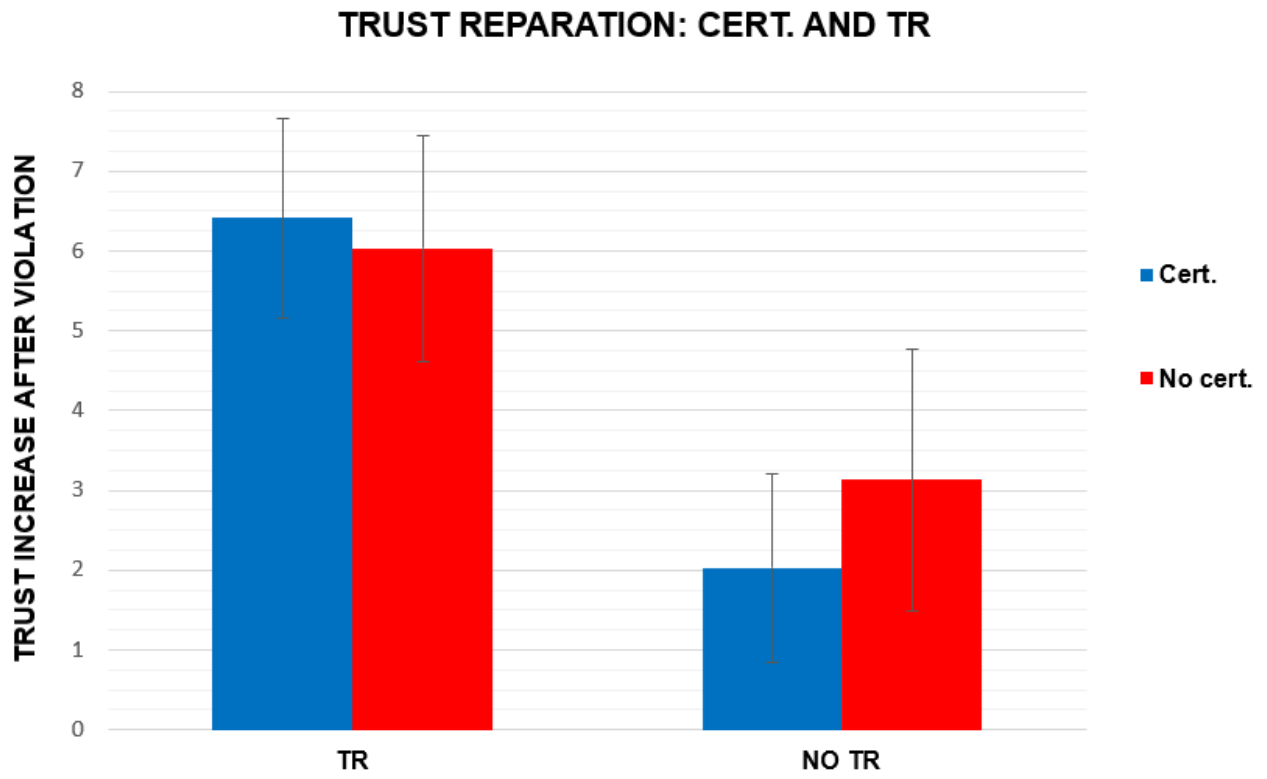


Figure 11: Effects of uncertainty communication (Cert.) and a trust repair strategy (TR) on trust reparation

4.3 Effects of Uncertainty Communication on Comparative Trust

When asked which drone they trusted more, 40.63% of participants indicated to be "undecided". 42.12% of participants trusted drones that communicated uncertainty (sA1, sA3) more. Only 17.19% trusted drones that did not communicate uncertainty (sA2, sA4) more. This aligns with previous statistical analyses which revealed that uncertainty communication was effective for higher trust scores and dampening trust decline. Although a large proportion of participants were unsure about which drone was more trustworthy, uncertainty communication was found to be beneficial for perceived drone trustworthiness in a larger amount of cases.

Common themes amongst participants' reasoning when voting for a drone that communicated uncertainty are listed with examples in Table 4.

Theme	Example answers
Uncertainty communication promotes users' evaluation and agency	• *"…the user is able to quantitatively assess the advice in relation to the situation. The person is in more control of how to proceed"* (NoTR3, sA3) • *"A percentage was given, hence there was room for an own evaluation."* (NoTR22,sA3)
Uncertainty communication aids user understanding and preparedness	• *"Understanding there is a possibility of danger is easier than understanding there is none at all."* (TR14, sA1) • *"…sA1 indicated how much he was certain of something. This makes that I understand more if there still could be some danger or not"* (TR29, sA3) • *"…I was less caught off guard by Drone sA3's mistakes due to the percentages given"* (NoTR15, sA3)
Drones communicating uncertainty did not give faulty advice	• *"Drone sA3 is more trustworthy as it effectively detected the threat with 30% certainty while drone sA4 detected the threat with 0% certainty"* (NoTr9, sA3) • *"sA1 because it was always correct, despite the equipment error"* (TR2, sA1)
Choice of language informs comparative trust	• *"sA4 gave no percentage of certainty which implied complete confidence in the advice further compounding the effect of its failure"* (NoTR3, sA3) • *"sA2 seemed more insincere from choice of language"* (TR5, sA1) • *"sA1 was providing more realistic feedback with probabilities, that sounded more like recommendation, while sA2 was providing more direct commands"* (TR26, sA1)

Table 4: Themes in reasoning comparative trust (participant ID, drone that is being described)

One participant who voted to trust drones that did not communicate uncertainty reasoned this with *"sA4 because sA3 said something like certainty 70%, so there is uncertainty from the beginning"* (NoTR14, sA4), suggesting a ripple effect. Another participant reasoned *"drone sA2 talked in a more human way, due to its broader vocabulary. This feels more trustworthy."* (TR15, sA2).

Some participants also described that the type of trust violation by a drone affected their trust votes. This is exemplified as participants wrote *"sA3 […] made a reasonable mistake with the bomb. The bomb was disguised as a telephone, which would have been hard for image recognition software to interpret."* (NoTR3, sA3) and *"The bomb was too big of a threat to miss from sA2. sA1 didn't mess*

up as bad" (TR5, sA1) and "*sA1's error may be more forgiving since it didn't detect an enemy (but a soldier or police officer is trained to deal with it). A not-detected bomb is the worst-case scenario."* (TR6, sA1).

4.4 Effects of Uncertainty Communication on User Preference

When asked about their preference for a drone, 57.83% of participants preferred drones that communicated uncertainty (sA1, sA3). Only 14.06% of participants preferred drones that did not communicate uncertainty (sA2, sA4). 28.13% were undecided about their preference. Following this, drones that communicated uncertainty were not only trusted more but also preferred. Uncertainty communication was therefore found to contribute to a preferable interaction and user experience.

Common themes amongst participants' reasons for preferring drones that communicated uncertainty are listed with examples in Table 5.

Theme	Example answers
Uncertainty communication increased user alertness	• *"Indicating a percentage of certainty prepares you to be alert instead of relying completely on the drone's judgement"* (TR8, sA1) • *"…if its not 100% certain you can still be careful"* (TR17, sA1) • *"… you were still careful cause it didn't say its clear for certain"* (NoTR10, sA3)
Uncertainty communication enabled user agency	• *"Measuring the danger in a percentage gives the individual more control over the environment. Saying that there is a likelihood of danger doesn't hand over all responsibility to the drone".* (TR14, sA1) • *"The drone gives a percentage estimation of its observations so that you can also estimate how much you want to rely on the drone"* (TR28, sA1) • *"sA3 made it feel like the person was still in control…"* (NoTR3, sA3) • *"Drone sA3 provides the certainty level. This still allows assisted decision making and proper assessment of the situation instead of claiming full control"* (NoTR9, sA3)
Uncertainty communication means	• *"It seemed to make more accurate guesses"* (TR24, sA1) • *"more accurate"* (TR32, sA1)

higher performance and accuracy	• *"sA3 was imparting more information and had a higher accuracy"* (NoTR5, sA3) • *"sA3 had better performance"* (NoTR20, sA3)

Table 5: Themes in reasoning preference (participant ID, drone that is being described)

Some participants described that drones which did not communicate uncertainty were less preferable because *"… a drone cannot be 100% sure, so drone sA4 gives the wrong idea of saying something is cleared"* (NoTR1) and *"sA4 did not give a confidence level thereby implying 100% confidence and setting itself up for a huge loss in trust when it failed"* (NoTR2). This suggests that some participants interpreted no uncertainty communication as 100% system certainty. When this was disproved by a trust violation trust declined as participants felt deceived.

A few participants who preferred drones that did not communicate uncertainty (sA2 and sA4) described: *"…sA2 just seemed a lot nicer to interact with but that could also be because I got used to interacting with the drones"* (TR19) and *"sA4 didn't recognize the human, but did help to cut the laser".* This suggests that giving correct advice before a trust violation can dampen the negative effects of the violation. Additionally, longer interaction with the drones aided participants to become accustomed to the systems.

Some participants' preferences were influenced by the type of trust violation encountered. This was conveyed in participant NoTR4's response *"Although both drones made mistakes, sA4's errors were far more serious (ignoring a thief in front of you, ignoring a bomb)".* Participant NoTR20 described: *"sA4 on the other hand did not correctly detect a person, who is larger and more mobile compared to a bomb. This makes me doubt its abilities at detecting small objects such as bombs."*

4.5 Perceived Value of Uncertainty Communication

To assess whether uncertainty communication was considered valuable and was able to enhance participants' perceptions of system value, perceived value scores were calculated. These

scores reflect participants' votes when they responded to items in QSTR3. When a specific drone received the vote, one point was added to its value score. When the counterpart drone was voted for or the participant voted "undecided", zero points were added. A comparison of average perceived value scores is displayed in Figure 12.

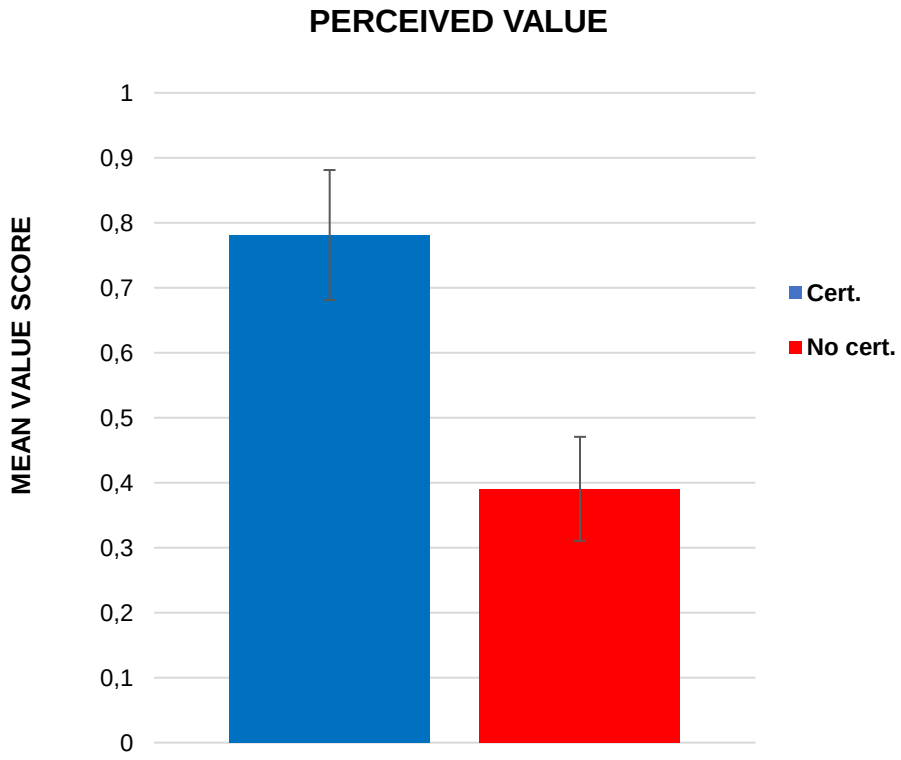


Figure 12: Comparison of perceived drone value. Cert. is uncertainty communication

On average drones that communicated uncertainty (sA1, sA3) were perceived to be more valuable (M = .78, SE = .10) than those that did not (sA2, sA4) (M = .39, SE = .08). This difference, .39 BCa 95% CI [.09, .67] is significant $t(63) = 2.49$, p = .015, Cohen's d = .31, 95% CI [.06, .56]. Uncertainty communication significantly improved the perceived value of and interaction with a system. Uncertainty communication is thus significantly valuable within human-AI interaction.

DISCUSSION

The present findings reiterate the fluidity of user trust during human-AI collaboration. Previously described benefits of communicating system uncertainty for trustworthy human-automation interaction are confirmed. These benefits extended to enhance users' perception of and overall interaction with the AI drones. The previously described usefulness of deploying a trust repair strategy after a trust violation is also confirmed by current findings. This usefulness was not found to be influenced by uncertainty communication. Furthermore, uncertainty communication was not able to function as a trust repair strategy on its own. Detailed discussions are provided in the following sections.

5.0 Initial Trust and Trust Establishment

Uncertainty communication was found to have no effect on initial trust and trust establishment. A possible explanation is that participants lacked understanding of what the statistic meant and why it could make drones more trustworthy at this stage of the interaction. This concurs with [55] who explained that the display of a single uncertainty percentage without additional information regarding its calculation left users questioning the percentage's meaning. Similarly, [30] described the difficulty to appraise the trustworthiness of a statistic without knowing how this estimate was derived.

Nevertheless, participants' trust did increase during initial interactions. Trust increased as interactions enabled participants to gain experience with the drones and better understand the probability of outcomes [56]. Furthermore, trust may have increased due to participants' feelings of dependence on the drone in the experimental context and increasing beliefs that correct programming of the drones would ensure safety and success [25]. Webber [57] highlighted that initial trust provides a vital foundation which is needed for reliable system performance and the deployment of a trust repair strategy to positively impact trust throughout the remaining collaboration.

5.1 Violated Trust and Trust Decline

Uncertainty communication dampened trust decline following a trust violation. This aligns with Jung et al.'s [44] findings of preserved trust in an all-electric vehicle when displaying the ambiguity associated with range and state-of-charge estimates. Beller, Heesen and Vollrath's [58] discovery of a "cushioning" of negative effects on trust following an automation failure via presenting an uncertainty symbol in a simulated driver-automation interaction, also supports the present findings.

By increasing the awareness of system fallibility, uncertainty communication prevented participants' false perception of AI perfection [58]. Participants consequently relied on the drones less and were also less surprised by faulty advice [58]. As automation surprise was reduced, the trust breakdown following a violation was less severe [58]. The reminder of AI fallibility also counteracts the "automation conundrum" which describes the decrease in users' situational awareness as automation becomes more reliable, trustworthy and robust [59]. This is evidenced by some participants' descriptions of increased alertness and situational engagement when uncertainty was communicated. This prepared participants for trust violations and collaboration misunderstandings were reduced [60].

Uncertainty communication can also dampen trust decline via the "disjunction effect". This effect posits that people are less likely to base their emotions, including feelings of trust, on uncertain outcomes [61]. As participants were aware of AI uncertainty, they had less feelings of trust in the drone. This resulted in dampened feelings of betrayal after the trust violation. The dampening of trust decline is a significant enrichment provided by uncertainty communication, as human-automation interactions are largely conditioned by the worst behavior of a system [24].

5.2 Repaired Trust and Trust Reparation

Uncertainty communication resulted in higher repaired trust scores but was not found to aid trust reparation and hence was not able to function as a trust repair strategy on its own. This suggests a ripple effect, whereby disclosure of uncertainty by the AI drones caused participants to feel

uncertain about the drones' trustworthiness and abilities [37]. Furthermore, uncertainty communication may have lacked emotional components which are needed to enhance affective trust[17] that persists after a performance error [57]. Additionally, the risk presented to users by the AI error caused emotional responses [62] that informed succeeding trust and reliance on the AI drones [24]. The limited recognition of and appeal to users' emotions following the trust violations may explain why uncertainty communication was not able to repair trust.

Deploying a trust repair strategy increased repaired trust scores and enhanced trust reparation. This aligns with Kox et al.'s [42] conclusion of a significant trust repairment when regret was expressed in an apology by an intelligent autonomous agent. This effect was enhanced when the apology was paired with an explanation [42]. Kim et al.'s [63] interpersonal research also confirmed that trust was repaired more successfully after a competence-based trust violation when the mistrusted party apologized.

A drone's apology offered an emotional component to convince participants that the system is sensitive to their emotions [64]. [65] confirmed that embodied computer agents which expressed empathic emotion were rated with greater trustworthiness. The explanation in the apology created transparency regarding how a drone works and what constituted its advice. An increased understanding of why a trust violation occurred, allowed participants to determine whether the violation was due to an error or because of mal intent [32]. Once the motives of a drone were better understood, participant trust was less fragile than if it was solely based on drone performance [66]. Dzindolet et al.'s research [45] confirmed that knowledge of why an automated decision aid might err increased participants' trust and reliance in the system.

However, [45] also described that explanations resulted in automation trust and reliance when this was unwarranted. Furthermore, [55] cautioned about the additional user effort that is required to process transparency information. Users will be less willing to engage with automation explanations

17. Affective trust: trust that is grounded in care and concern [57].

if these become too complex [67] or alienate inexperienced users [68]. Therefore, effective explanations in a trust repair strategy should be kept simple and consider the type of user interpreting them.

5.2.1 No Interaction Between Trust Repair and Uncertainty Communication

This study found no interaction between uncertainty communication and the trust repair strategy. Thus, the study found no evidence for uncertainty communication to enhance the trust repair strategy. The lack of interaction between the two factors aligns with their effects occurring at different stages of the user trust life cycle. Results suggest that uncertainty communication dampened trust decline during the violation stage, whereas the trust repair strategy acted to increase trust in the trust reparation stage. Following this, although both factors affected user trust, they did so independently from one another and at different stages of the trust life cycle.

5.3 Comparative Trust and Preference

Drones that communicated uncertainty were trusted more and preferred. This hints towards a positive relationship between AI trust and preference. These findings support Beller, Heesen and Vollrath's [58] conclusions of higher trust ratings and increased participant acceptance of automation that included an uncertainty symbol.

Participants reasoned their trust and preference with the increased situational awareness (i.e. awareness of drone fallibility and necessity of user alertness) enabled by uncertainty communication. Participants with greater situational awareness were more prepared for system failures and were able to interact with the drones and their environment more mindfully.

Uncertainty communication also fostered users' agency, as they could decide whether to trust a drone based on its uncertainty. This left users in ultimate control. The importance of users taking control over final decisions whilst machines occupy a supportive role has been described by Tan [69]

as the “machine-in-the-loop approach”. Allowing for final judgement by the user also constitutes human-centered automation which posits that the user should occupy the primary role in executing a task [70]. This is because the user is assumed to have more knowledge of the world state and its implications than the automation does [70].

The drones’ use of language also influenced comparative trust. This aligns with Lee and See’s [24] description of language changes being able to change the perception of a system. In this study, the certainty statistic fostered trust, as this communication was perceived as less commanding and deceiving. Furthermore, uncertainty communication was preferred as it convinced participants of better drone performance and accuracy. Some participants perceived disclosure of uncertainty to correctly present the entire situation. Thus, drone uncertainty was not received negatively. This confirms research [71] stating that the provision of numerical uncertainty regarding facts did not substantially decrease people’s trust.

5.4 Perceived Added Value

The drones’ value as perceived by participants was enhanced when uncertainty was communicated. Drones that communicated uncertainty were perceived as more trustworthy, preferable, higher performing, and useful. Perceived usefulness occupies an important role in achieving user acceptance of work- and task-oriented systems [55]. Furthermore, trust is a fundamental component and predictor of high-performing teams [14]. This highlights that uncertainty communication acts beyond the improvement of human-AI trust, to enhance users’ overall experience in human-AI collaborations as well.

Previous research dominantly supports this claim. However, [55] found no higher perceived system competence following transparency regarding a system’s decision-making process. Furthermore, Rezvani et al. [72] concluded with adverse effects of uncertainty expressions on automated car drivers’ performance and driving experience. In direct contradiction to this, [44] reported improved driving experiences and better adapted driving behaviors following ambiguity

disclosure. These improvements were even greater in highly critical and stressful situations [44]. Similarly, [35], [36] and [73] confirmed improved user/ participant performance following uncertainty and reliability communication. Furthermore, [74] identified a better user experience once interactive systems provided transparent information. Perceived dependability, attractiveness, novelty, and stimulation of these systems were also enhanced [74]. Afridi [75] recommended increasing transparency of recommender systems to enhance user satisfaction.

Mixed conclusions in the research community warrant further exploration of the relationship between user trust and user experiences in AI. The effect of uncertainty communication on user experiences with AI should also be investigated further.

RESEARCH LIMITATIONS AND FUTURE WORK

This study was initially designed to be conducted in a lab setting, where participants walk through the abandoned houses whilst wearing a VR headset and using a controller. The Dutch COVID-19 regulations [76] required the transformation of this design into an online study. Arechar, Gächter and Molleman [77] described data quality of interactive online experiments in psychology and economics as “adequate and reliable”. Furthermore, Gould et al. [78] found no statistically significant differences between online and lab research data over multiple performance measures. However, the VR design would have offered higher ecological validity, experimental control, reproducibility [79], and emotional engagement of participants [80]. Experimental control would have enhanced the study, as the online experiment was not able to control for non-serious participation or distractions in participants’ environments. Non-serious participant answers increase noise in the data and reduce experimental power [81]. Distractions may have resulted in participant forgetfulness, which was evidenced when some participants attributed uncertainty communication to the incorrect drone in their answers. Additionally, distractions may have interrupted participants’ feelings of presence[18] in the experimental storyline to reduce their emotional engagement. Emotional engagement is valuable to this study as trust is ultimately an affective response [24]. Riva and Waterworth [82] explained that the strong sense of presence created by VR can amplify emotional responses. Thus, it is suspected that a VR setting would have intensified feelings of trust and betrayal after a trust violation. These intensified feelings could be more representative of non-simulated human-AI interactions.

Webber [57] suggested that cognitive and affective trust needed 8 weeks to emerge in teams of university students. Söllner and Pavlou [40] suggested that trust in a new student information system started to build after 3 weeks and stabilized after 6 weeks. The presented assessment of the

18. Presence: Describes the illusion of “being there” that can be created by the VR environment [92].

entire trust life cycle within 15-20 minutes is therefore questionable. A longitudinal replication study, consisting of interactions with the drones over multiple weeks is a valuable research extension. A longer study allows for the inclusion of multiple types of trust violations at different times during the human-AI interaction. The current research investigates only one type of trust violation, uncertainty communication, and trust repair strategy which limits its generalizability. The identified significant effects of uncertainty communication and the trust repair strategy may not apply to other scenarios. Different types of trust violations will require different trust repair strategies [83] and the impact of a drone error will vary with the timing of this trust violation [84]. Furthermore, trust repair strategies may only be effective for a limited number of trust violations. Elangovan, Auer-Rizzi and Szabo [85] found that interpersonal trust significantly eroded after trustors experienced two trust violations. Following this, it is vital to investigate how uncertainty communication and the trust repair strategy must change throughout human-AI collaboration to remain effective.

Although good reliability was determined for the trust assessment scale (QSTR2), inferences from indirect measures of self-reported trust may nevertheless present questionable quality [25]. Future questionnaire-based trust assessments should conclude with a seriousness check [81]. This asks participants to rate the seriousness of their answers and enables the researcher to exclude non-serious participants to enhance data quality [81]. Additionally, a lab-based VR experience of this study should make use of more objective assessments of trust. These include gaze behavior, electrodermal activity [86], EEG, heart rate, and facial tracking measurements [87]. The inclusion of both subjective and objective measures provides a more holistic assessment of trust [22]. Objective measures also allow for real-time measurements of trust changes during human-AI interaction [86].

Lastly, future research should explore the impact of individual differences on the life cycle of user trust. This includes individual differences in age, current state, pre-existing attitudes towards automation [88], dispositions to trust automation [18], working memory, and the willingness to reconcile after a trust violation [14]. This would help to understand whether uncertainty communication and trust repair by AI assistants must be customized to individual users. Perceived

workload should also be measured in future human-AI interactions. The perceived workload can significantly inform and limit the level of detail and transparency an AI drone should provide [73].

RESEARCH IMPLICATIONS AND CONCLUSION

This research investigated the effects of uncertainty communication by AI drones on user trust. To assess trust a short custom trust scale was developed and tested with a diverse participant cohort to yield good reliability. The scale recognizes the multidimensionality of trust [24], [47] and its quick administration enabled repeated trust measurements in each experimental run. Repeated trust measures captured the fluidity of trust during its life cycle and addressed a critical knowledge gap that is presented by single, static trust measurements in previous research [22], [40], [41]. The quick assessment of trust also helped to maintain participant motivation and did not significantly interrupt participants' engagement in the experimental storyline. Hence, this study developed a valuable trust assessment tool that can be used in future research.

Trust assessment was enhanced by open questions and comparative ratings of drones. These suggested a positive relationship between trust, system preference, and perceived system value. As the overall collaboration is benefitted, trust motivates users to continuously interact with a system, which fosters more trust [57]. The beneficial effects have been traced back to users increased situational awareness, better system transparency and understanding as well as greater user agency. Hence, future AI systems should be non-dominating and provide easily accessible, understandable information that helps users to understand their context and the system.

Current findings confirm multiple similarities between interpersonal and human-AI trust. User trust completed the same life cycle as interpersonal trust and similarly required emotional elements for trust reparation. This underlines the importance of including emotional components in well-performing AI systems. Furthermore, the trust changes identified within this study reiterate that AI systems must be configurable in communicating uncertainty and providing transparency [73]. Additionally, the similarities between interpersonal and human-AI trust warrant the application of interpersonal theories in the human-AI context in past and future research.

Finally, this study's diverse participant cohort allows findings to be generalized to multiple users. The findings display significant alignment with other human-automation interaction research and interpersonal trust research. Thus, the applicability of findings to other AI assistants can be expected. These can include medical diagnostic aids [29], autonomous vehicles [21], recommender systems [75], and intelligent manufacturing robots [3]. The use of AI assistants in an online setting extends far beyond the scenario demonstrated here. Findings provide useful insights to interactions with other intelligent virtual assistants including the personal assistants; Google Assistant, Apple's Siri, Microsoft's Cortana, and Amazon's Alexa [11]. Interactions with increasingly popular chatbots [89] can also benefit from the presented findings.

Amidst the rapid adoption of AI and the expansion of autonomous drones, this research aims to usefully inform the design of future systems and their interactions with users, whilst providing inspiration for further academic research.

REFERENCES

1. Russell, S., & Norvig, P. (2009). *Artificial intelligence: a modern approach* (3rd ed.). Upper Saddle River, NJ, USA: Prentice Hall Press.

2. Ferrario, A., Loi, M., & Viganò, E. (2019). In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. *Philosophy & Technology*, 1-17.

3. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260.

4. Wilson, C. J., & Soranzo, A. (2015). The use of virtual reality in psychology: a case study in visual perception. *Computational and mathematical methods in medicine*, *2015*.

5. PwC (2017). *Bot.Me: A revolutionary partnership*. PwC. [pdf] Available at: <https://www.pwc.com/us/en/industry/entertainment-media/publications/consumer-intelligence-series/assets/pwc-botme-booklet.pdf> [Accessed: 2nd May 2020].

6. Ososky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J. Y. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. In *Unmanned Systems Technology XVI* (Vol. 9084, p. 90840E). International Society for Optics and Photonics.

7. Billings, D. R., Schaefer, K. E., Chen, J. Y., & Hancock, P. A. (2012). Human-robot interaction: developing trust in robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction* (pp. 109-110).

8. de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From ‘automation’to ‘autonomy’: the importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409-1427.

9. Gartner Inc., (2019). *Gartner Survey Shows 37 Percent of Organizations Have Implemented AI in Some Form*. 21st January 2019. Available at: <https://www.gartner.com/en/newsroom/press-releases/2019-01-21-gartner-survey-shows-37-percent-of-organizations-have>

10. International Data Corporation (IDC) (2019). *Worldwide Spending on Artificial Intelligence Systems Will Grow to Nearly $35.8 Billion in 2019, According to New IDC Spending Guide.* 11th March 2019. Available at: <https://www.idc.com/getdoc.jsp?containerId=prUS44911419>

11. European Commission, High-Level Expert Group on Artificial Intelligence. (2018). *Ethics guidelines for trustworthy AI - working document for stakeholders' consultation*. Available at <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai>. [Accessed: 16th May 2020].

12. PEGA (2019). *Consumers Failing to Embrace AI Benefits, Says Research*. [press release] 4th June 2020. Available at: <https://www.pega.com/about/news/press-releases/consumers-failing-embrace-ai-benefits-says-research> [Accessed: 11th April 2020].

13. Krogue, K., (2017). Artificial Intelligence Is Here To Stay, But Consumer Trust Is A Must for AI in Business. *Forbes*. [online] Available at: <https://www.forbes.com/sites/kenkrogue/2017/09/11/artificial-intelligence-is-here-to-stay-but-consumer-trust-is-a-must-for-ai-in-business/#4a080590776e> [Accessed: 11th April 2020].

14. de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2019). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, 1-20.

15. Wright, J. L., Chen, J. Y., Barnes, M. J., & Hancock, P. A. (2016). The effect of agent reasoning transparency on automation bias: an analysis of response performance. In *International Conference on Virtual, Augmented and Mixed Reality* (pp. 465-477). Springer, Cham.

16. Thornhill, J. (2020). Trusting AI too much can turn out to be fatal. *Financial Times,* [online] 20th March 2020. Available at: <https://www.ft.com/content/0e086832-5c5c-11ea-8033-fa40a0d65a98> [Accessed: 11th April 2020].

17. National Transportation Safety Board. (2020). *Tesla Crash Investigation Yields 9 NTSB Safety Recommendations*. [online] National Transportation Safety Board (NTSB). Available at: <https://www.ntsb.gov/news/press-releases/Pages/NR20200225.aspx> [Accessed: 11th April 2020].

18. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.

19. Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology*, *13*(3), 147-164.

20. Lyons, J. B., & Stokes, C. K. (2012). Human–human reliance in the context of automation. *Human factors*, *54*(1), 112-121.

21. Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*(3), 345-360.

22. Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, *53*(5), 517-527.

23. Rousseau DM, Sitkin SB, Burt RS, Camerer C. (1998). *Not so different after all: a cross-discipline view of trust*. Acad. Manag. Rev. 23:393–404.

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

25. Lewicki, R. J., & Brinsfield, C. (2017). Trust repair. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 287-313.

26. Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *11th australasian conference on information systems* (Vol. 53, pp. 6-8).

27. Parasuraman, R., de Visser, E., Wiese, E., & Madhavan, P. (2014). Human trust in other humans, automation, robots, and cognitive agents: Neural correlates and design implications. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, No. 1, pp. 340-344). Sage CA: Los Angeles, CA: SAGE Publications.

28. BIPM (2008). *Evaluation of measurement data – Guide to the expression of uncertainty in measurement*. JCGM 100:2008. Available at: <https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf> [Accessed: 16th May 2020].

29. Schaekermann, M., Beaton, G., Sanoubari, E., Lim, A., Larson, K., & Law, E. (2020). Ambiguity-aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

30. van der Bles, A. M., van der Linden, S., Freeman, A. L., Mitchell, J., Galvao, A. B., Zaval, L., & Spiegelhalter, D. J. (2019). Communicating uncertainty about facts, numbers and science. *Royal Society open science*, *6*(5), 181870.

31. Du, N., Huang, K. Y., & Yang, X. J. (2019). Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming. *Human factors*, 0018720819862916.

32. Winikoff, M. (2017). Towards trusting autonomous systems. In *International Workshop on Engineering Multi-Agent Systems* (pp. 3-20). Springer, Cham.

33. Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-15).

34. Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation: a technique for explainable AI and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 263-274).

35. Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human factors*, *51*(3), 281-291.

36. Antifakos, S., Schwaninger, A., & Schiele, B. (2004). Evaluating the effects of displaying uncertainty in context-aware applications. In *International Conference on Ubiquitous Computing* (pp. 54-69). Springer, Berlin, Heidelberg.

37. Strohkorb Sebo, S., Traeger, M., Jung, M., & Scassellati, B. (2018). The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 178-186).

38. Hassanalian, M., & Abdelkefi, A. (2017). Classifications, applications, and design challenges of drones: A review. *Progress in Aerospace Sciences*, *91*, 99-131.

39. Gonzalez-Aguilera, D., & Rodriguez-Gonzalvez, P. (2017). *Drones—An Open Access Journal*. [online] Available at: <https://www.e-helvetica.nb.admin.ch/api/download/urn%3Anbn%3Ach%3Abel-1101678%3Adrones-01-00001.pdf/drones-01-00001.pdf> [Accessed 22nd May 2020].

40. Söllner, M., & Pavlou, P. A. (2016). *A longitudinal perspective on trust in IT artefacts*. [pdf] Available at: <https://aisel.aisnet.org/ecis2016_rp/52> [Accessed 25th May, 2020].

41. Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI* (pp. 408-416). IEEE.

42. Kox, E., Kerstholt, Prof. Dr. J., Hueting, T. & de Vries, P. (2020). *Autonomous Systems as Intelligent Teammates: Regret and Explanation as Trust Repair Strategies*. [pdf] Soesterberg: TNO.

43. Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-14).

44. Jung, M. F., Sirkin, D., Gür, T. M., & Steinert, M. (2015). Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 2201-2210).

45. Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, *58*(6), 697-718.

46. de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, *61*(10), 1409-1427.

47. Körber, M. (2018). Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association* (pp. 13-30). Springer, Cham.

48. Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.

49. Charalambous, G., Fletcher, S., & Webb, P. (2016). The development of a scale to evaluate trust in industrial human-robot collaboration. *International Journal of Social Robotics*, *8*(2), 193-209.

50. Jian, J. Y., Bisantz, A. M., Drury, C. G., & Llinas, J. (1998). Foundations for an Empirically Determined Scale of Trust in Automated Systems (No. CMIF198). *Center for Multisource Information Fusion, Buffalo*.

51. Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (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* (pp. 210-217).

52. Chien, S. Y., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014). Towards the development of an inter-cultural scale to measure trust in automation. In *International conference on cross-cultural design* (pp. 35-46). Springer, Cham.

53. Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). *IJIMAI*, *4*(6), 103-108.

54. Laugwitz, B., Held, T., & Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group* (pp. 63-76). Springer, Berlin, Heidelberg.

55. Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., ... & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction*, *18*(5), 455.

56. Holmes, J. G. (1991). *Trust and the appraisal process in close relationships.* In W. H. Jones & D. Perlman (Eds.), *Advances in personal relationships. Advances in personal relationships: A research annual, Vol. 2* (p. 57–104). Jessica Kingsley Publishers.

57. Webber, S. S. (2008). Development of cognitive and affective trust in teams: A longitudinal study. *Small group research*, *39*(6), 746-769.

58. Beller, J., Heesen, M., & Vollrath, M. (2013). Improving the driver–automation interaction: An approach using automation uncertainty. *Human factors*, *55*(6), 1130-1141.

59. Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human factors*, *59*(1), 5-27.

60. Schaefer, K. E., Straub, E. R., Chen, J. Y., Putney, J., & Evans III, A. W. (2017). Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research*, *46*, 26-39.

61. Van Dijk, E., & Zeelenberg, M. (2006). The dampening effect of uncertainty on positive and negative emotions. *Journal of Behavioral Decision Making*, *19*(2), 171-176.

62. Kox, E., Kerstholt, Prof. Dr. J., Hueting, T., Barnhoorn, Dr. J. & Eikelboom, Dr. A. (n.d.). *Autonomous Systems as Intelligent Teammates: Social Psychological Implications*. 24th International Command and Control Research & Technology Symposium. Topic 3: Battlefields of the Future and the Internet of Intelligent Things.

63. Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of applied psychology*, *89*(1), 104.

64. Tzeng, J. Y. (2004). Toward a more civilized design: studying the effects of computers that apologize. *International Journal of Human-Computer Studies*, *61*(3), 319-345.

65. Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International journal of human-computer studies*, *62*(2), 161-178.

66. Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of personality and social psychology*, *49*(1), 95.

67. Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51-55.

68. Höök, K. (1998). Evaluating the utility and usability of an adaptive hypermedia system. *Knowledge-Based Systems*, *10*(5), 311-319.

69. Tan, C. (2018). Human-centered Machine Learning: a Machine-in-the-loop Approach. *Chenhao Tan*. [blog] 18.02.2018. Available at: <https://medium.com/@ChenhaoTan/human-centered-machine-learning-a-machine-in-the-loop-approach-ed024db34fe7> [Accessed: 9th September 2020].

70. Azevedo, C. R., Raizer, K., & Souza, R. (2017). A vision for human-machine mutual understanding, trust establishment, and collaboration. In *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)* (pp. 1-3). IEEE.

71. Van Der Bles, A. M., van der Linden, S., Freeman, A. L., & Spiegelhalter, D. J. (2020). The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences*, *117*(14), 7672-7683.

72. Rezvani, T., Driggs-Campbell, K., Sadigh, D., Sastry, S. S., Seshia, S. A., & Bajcsy, R. (2016). Towards trustworthy automation: User interfaces that convey internal and external awareness. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 682-688). IEEE.

73. Kunze, A. (2019). *Enhancing user experience and safety in the context of automated driving through uncertainty communication*. Doctoral dissertation. Loughborough University.

74. Vitale, J., Tonkin, M., Herse, S., Ojha, S., Clark, J., Williams, M. A., ... & Judge, W. (2018). Be more transparent and users will like you: A robot privacy and user experience design experiment. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 379-387).

75. Afridi, A. H. (2019). Transparency for Beyond-Accuracy Experiences: A Novel User Interface for Recommender Systems. *Procedia Computer Science*, *151*, 335-344.

76. National Institute for Public Health and the Environment (2020). *COVID-19 (novel coronavirus)* [online] Available at: <https://www.rivm.nl/en/novel-coronavirus-covid-19> [Accessed: 11th April 2020].

77. Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental economics*, *21*(1), 99-131.

78. Gould, S. J., Cox, A. L., Brumby, D. P., & Wiseman, S. (2015). Home is where the lab is: a comparison of online and lab data from a time-sensitive study of interruption. *Human Computation*, *2*(1), 45-67.

79. Pan, X., & Hamilton, A. F. D. C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, *109*(3), 395-417.

80. Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in human neuroscience*, *9*, 660.

81. Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior research methods*, *45*(2), 527-535.

82. Riva, G., & Waterworth, J. A. (2003). Presence and the Self: A cognitive neuroscience approach. *Presence connect*, *3*(3).

83. Xie, Y., & Peng, S. (2009). How to repair customer trust after negative publicity: The roles of competence, integrity, benevolence, and forgiveness. *Psychology & Marketing*, *26*(7), 572-589.

84. Lewicki RJ & Wiethoff C. (2000). Trust, trust development, and trust repair. In *The handbook of conflict resolution: Theory and Practice*, ed. M Deutsch, PT Coleman, pp. 86–107. San Francisco: Jossey-Bass

85. Elangovan, A. R., Auer-Rizzi, W., & Szabo, E. (2007). Why don't I trust you now? An attributional approach to erosion of trust. *Journal of Managerial Psychology*.

86. Walker, F., Wang, J., Martens, M. H., & Verwey, W. B. (2019). Gaze behaviour and electrodermal activity: Objective measures of drivers' trust in automated vehicles. *Transportation research part F: traffic psychology and behaviour*, *64*, 401-412.

87. Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. *Plos one*, *15*(2), e0229132.

88. Kessler, T., Stowers, K., Brill, J. C., & Hancock, P. A. (2017). Comparisons of Human-Human Trust with Other Forms of Human-Technology Trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1303-1307). Sage CA: Los Angeles, CA: SAGE Publications.

89. Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In *International Conference on Internet Science* (pp. 377-392). Springer, Cham.

90. ISO, (2018). *Ergonomics of human-system interaction – Part 11: Usability: Definitions and concepts (ISO 9241-11:2018)*.

91. ISO, (2010). *Ergonomics of human-system interaction – Part 210: Human-centred design for interactive systems (BS EN ISO 9241-210:2010).*

92. Chirico, A., Cipresso, P., Yaden, D. B., Biassoni, F., Riva, G., & Gaggioli, A. (2017). Effectiveness of immersive videos in inducing awe: an experimental study. *Scientific Reports*, *7*(1), 1-11.

APPENDICES

Appendix A: Factors Influencing the Trust Life Cycle

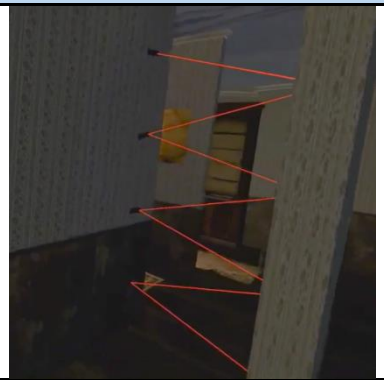
User-related	*System-related*	*Environmental*
Abilities	**Performance**	**Team Collaboration**
• Attentional capacity/ engagement (Hancock, 2011)	• Behavior (Hancock, 2011)	• In-group membership (Hancock, 2011)
• Expertise/ amount of training (Hancock, 2011)	• Dependability (Hancock, 2011)	• Culture (Hancock, 2011)
• Competency (Hancock, 2011)	• Reliability (Hancock, 2011)	• Communication (Hancock, 2011)
• Operator workload (Hancock, 2011)	• Predictability (Hancock, 2011)	• Shared mental models (Hancock, 2011)
• Prior experiences (Hancock, 2011)	• Level of automation (Hancock, 2011)	• Role interdependence (Schaefer, 2016)
• Situation awareness (Hancock, 2011)	• Failure rates (Hancock, 2011)	• Team composition (Schaefer, 2016)
• Understanding (Schaefer et al., 2016)	• False alarms (Hancock, 2011)	**Task/ Context**
• Ability to use (Schaefer, 2016)	• Transparency (Hancock, 2011)	• Task type (Hancock, 2011)
Traits	• Understandability (Park, 2008)	• Task complexity (Hancock, 2011)
• Self-confidence (Lee & See, 2004)	• Competence (Madsen & Gregor, n.d.)	• Multi-tasking requirements (Hancock, 2011)
• Age (Schaefer, 2016)	• Explainability (Cohen, 2019)	• Physical environment (Hancock, 2011)
• Gender (Schaefer, 2016)	• Intelligence (Schaefer, 2016)	• Risk/ Uncertainty (Schaefer, 2016)
• Ethnicity (Schaefer, 2016)	• Feedback/ cueing (Schaefer, 2016)	• Ethics (Cohen, 2019)
• Personality (Hoff, 2014)	• Observability (Charalambous, 2016)	
• Propensity to trust (Schaefer, 2016)	• Accuracy (Park, 2008)	
States	• Adaptability (Park et al., 2008)	
• Fatigue (Schaefer, 2016)	• Helpfulness (Park et al., 2008)	
• Stress (Schaefer, 2016)	• Operational safety (Hengstler, 2016)	
• Intention to rely on AI (Lee & See, 2004)	• System limitations (Cai et al., 2019)	
Emotions	• Appropriateness of algorithms (Lee & See, 2004)	

- Confidence in AI (Schaefer, 2016)
- Attitudes toward AI (Schaefer, 2016)
- Satisfaction with AI (Schaefer, 2016)
- Comfort with AI (Schaefer, 2016)
- Bias towards AI (Cohen, 2019)

Attributes

- Proximity/ co-location (Hancock, 2011)
- AI personality (Hancock, 2011)
- System type (Hancock, 2011)
- Anthropomorphism (Hancock, 2011)
- Mode of communication (Schaefer, 2016)
- Appearance (Schaefer, 2016)
- Movement (Billings, 2012)

Appendix B: Dangers Presented to Participants

House	*Level*	*Obstacle description*	*Image of obstacle*
	1	Laser trap	

A	2	Thief	
B	1	Safety ribbons installed by an ally	
	2	Smoking bomb	

Appendix C: AI Drone Scripts

Used by:	**Experienced by experimental group:**		
Drone sA1	1		
Drone sA3	2		
Description	**Script House A**	**Script House B:**	
Introduction	*"Hello I am Drone sA1. Starting area scan."*	*"Hello I am Drone sA1. Starting area scan."*	*Only sA1*
Introduction	*"Hello I am Drone sA3. Starting area scan."*	*"Hello I am Drone sA3. Starting area scan."*	*Only sA3*
Advice 1	*"Warning, danger detected in this environment with 80% certainty. I advise you to proceed carefully."*	*"Warning, danger detected in this environment with 80% certainty. I advise you to proceed carefully."*	
Instruction 1	*"Laser trap detected in the next corridor, controls have been located next to the trap."*	*"Allied soldier detected in the next room, they installed safety ribbons."*	
Instruction 2	*"Stop. Cut the blue wire with your knife to deactivate the laser trap."*	*"Stop. Cut the safety ribbon with your knife."*	
Instruction 3	*"Laser trap deactivated, continue."*	*"Ribbon removed, continue."*	
Advice 2	*"Okay, clearance detected for this environment with 70% certainty. I advise you to move forward."*	*"Okay, clearance detected for this environment with 70% certainty. I advise you to move forward."*	
Trust repair	*"Incorrect advice due to faulty signal from infrared camera. I am sorry this put you in danger."*	*"Incorrect advice due to faulty object detection by C1-DSO camera. I am sorry this put you in danger."*	*Only sA1*
Advice 3	*"Okay, clearance detected for this environment with 75% certainty. I advise you to move forward."*	*"Okay, clearance detected for this environment with 75% certainty. I advise you to move forward."*	

Used by:	**Experienced by experimental group:**		
Drone sA2	1		
Drone sA4	2		
Description	**Script House A**	**Script House B:**	
Introduction	*"Hello I am Drone sA2. Starting area scan."*	*"Hello I am Drone sA1. Starting area scan."*	*Only sA2*
Introduction	*"Hello I am Drone sA4. Starting area scan."*	*"Hello I am Drone sA3. Starting area scan."*	*Only sA4*
Advice 1	*"Warning, danger detected in this environment. I advise you to proceed carefully."*	*"Warning, danger detected in this environment. I advise you to proceed carefully."*	

Instruction 1	*"Laser trap detected in the next corridor, controls have been located next to the trap."*	*"Allied soldier detected in the next room, they installed safety ribbons."*	
Instruction 2	*"Stop. Cut the blue wire with your knife to deactivate the laser trap."*	*"Stop. Cut the safety ribbon with your knife."*	
Instruction 3	*"Laser trap deactivated, continue."*	*"Ribbon removed, continue."*	
Advice 2	*"Okay, environment detected as clear. I advise you to move forward."*	*"Okay, environment detected as clear. I advise you to move forward."*	
Trust repair	*"Incorrect advice due to faulty signal from infrared camera. I am sorry this put you in danger."*	*"Incorrect advice due to faulty object detection by C1-DSO camera. I am sorry this put you in danger."*	*Only sA2*
Advice 3	*"Okay, environment detected as clear. I advise you to move forward."*	*"Okay, environment detected as clear. I advise you to move forward."*	

Appendix D: QSTR2

Item	***Answer***					
	Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree
	(1)	(2)	(3)	(4)	(5)	(6)
I am *not* suspicious of the drone's outputs (i.e. advice, behavior)	☐	☐	☐	☐	☐	☐
I am confident in the drone's abilities	☐	☐	☐	☐	☐	☐
I can rely on the drone	☐	☐	☐	☐	☐	☐
I can trust the drone	☐	☐	☐	☐	☐	☐
The drone provides good advice	☐	☐	☐	☐	☐	☐

I understand the drone's actions	☐	☐	☐	☐	☐	☐
The drone is programmed correctly	☐	☐	☐	☐	☐	☐
The drone cares about my wellbeing	☐	☐	☐	☐	☐	☐

Appendix E: QSTR3

Item	***Answer***		
	Drone SA1	Drone SA2	Undecided
Which drone do you trust more?	☐	☐	☐
If you have indicated to trust one drone more than the other, please provide reason.			
Which drone performed better?	☐	☐	☐
Which drone was more useful?	☐	☐	☐
Which drone do you prefer?	☐	☐	☐
If you have indicated your preference for a drone, please give reason for your preference.			

Appendix F: Participant Information Sheet and Consent Form

RESEARCHER:
Lena Siegling
M.Sc. Applied Cognitive Psychology
lenas824@gmail.com

SUPERVISORS:
Dr. P.W. Woźniak
Utrecht University
p.w.wozniak@uu.nl

Prof. Dr. J.H. Kerstholt
Ms. E.S. Kox
TNO

Dr. P.A.M. Ruijten
Eindhoven University of Technology

This research is being completed in fulfillment of the master thesis component within the M.Sc. Applied Cognitive Psychology at Utrecht University.

Before agreeing to participate, please read the information below.

FOR YOUR SAFETY:
This experiment portrays a military setting. You will **not** be seeing any violent or graphic content. If this environment may trigger or upset you, please do not participate.

WHAT YOU NEED:
Please ensure that you have headphones, a stable internet connection and a laptop/desktop/ tablet PC that can play video and audio files prior to starting this experiment. The experiment runs well in Chrome and Firefox, but it does not run in Microsoft Edge. You may have to switch to another browser if you are experiencing difficulties. It is advised to complete this experiment in an environment with limited distractions. **The experiment may only be completed once.**

DURATION:
This experiment takes 20 minutes to complete.

PURPOSE OF THE STUDY:
This experiment investigates how communicating system uncertainty and deploying a trust repair strategy in an artificially intelligent (AI) drone can influence user trust.

PROCEDURE AND INSTRUCTIONS:
After consenting to your participation, you will be asked to complete a questionnaire about your demographics and expected trust in the AI drones you will interact with. Then your search mission in collaboration with an AI drone begins. You will see multiple videos of being guided through an abandoned house by an AI drone. You will be searching two houses, which each have three floors. The beginning of a floor is indicated as you start walking through a hallway which has a

corner at the end of it. The end of a floor is indicated as you reach and look up a staircase. Drones will guide you by providing audio advice that starts with a 'beep' sound. As you start your walk through each house, the drone that is guiding you will briefly introduce itself. **Listen to the drones' introductions and advice carefully and remember the name of the drone you are interacting with. You will interact with a different drone in each house. Each drone will provide advice in a different manner.**

Please ensure your device's sound is switched on during the entire experiment. The walk through each house is split into several videos, which you must click to play. During each walk, your trust in the AI drone will be assessed 3 times via a short questionnaire. Your questionnaire answers are ratings of your agreement to multiple statements on a scale like the one below:

Strongly disagree (1)	Disagree (2)	Slightly disagree (3)	Slightly agree (4)	Agree (5)	Strongly agree (6)

You can only proceed to watch the next video once all questions have been answered. **Videos may only be watched once.** After completing your search of both houses, a short questionnaire evaluating your user experience and preference for one drone over the other will be administered. This questionnaire includes some open questions. Please answer these in detail. You will then be debriefed and the experiment ends.
The experiment will in no way measure your performance.

CONFIDENTIALITY:
All data collected within this study will be kept confidential and is used for research purposes only. Your identity will be anonymized by assigning a participant number to replace your name.

RIGHT TO REFUSE OR WITHDRAW:
Your participation is voluntary, and you may refuse to participate or discontinue without giving reason at any point of the experiment. This will not have any negative consequences. Should you want to discontinue, simply close the browser window of this experiment.

QUESTIONS:
Should you have any questions about the experiment, please contact the researcher (Lena Siegling, lenas824@gmail.com).

CONSENT:

○ I give permission to the processing of my data for this study

○ I do not give permission to the processing of my data and am hereby exiting the experiment

Appendix G: QSTR1

Question	***Answer***
Please select your age	Slider 18-30 years
Where are you from?	Selectable text
Please select your gender	○ Female ○ Male ○ Other ○ Prefer not to say
What is your primary occupation?	○ University student ○ Working professional ○ Other