Detecting and Overcoming Trust Miscalibration in Real Time Using an Eye-tracking Based Technique

by

Yidu Lu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2020

Doctoral Committee:

Professor Nadine Sarter, Chair
Professor Richard Gonzalez
Associate Professor Bernard Martin
Assistant Professor Xi Jessie Yang

Yidu Lu

luyd@umich.edu

ORCID iD: 0000-0003-4261-0053

DEDICATION

This dissertation is dedicated to the readers.

ACKNOWLEDGEMENTS

I would like to express my thanks to many people who made it possible for me to complete this dissertation work.

First and foremost, I would like to thank my advisor, Dr. Nadine Sarter, for all the support, guidance, and encouragement throughout these past five years. I am really honored to be advised by such a great advisor. And I am very grateful for the time she devoted to help me grow as a researcher step by step. She always guided me to think critically, innovatively, and pay attention to details for high quality research. Thank you for all the trust, the patience, the insightful discussions, the challenging questions, and the valuable suggestions. They made these five years a wonderful journey for me.

I would also like to thank the members of my dissertation committee: Dr. Richard Gonzalez, Dr. Bernard Martin, and Dr. Xi Jessie Yang, for serving on my committee and providing constructive feedback throughout this dissertation work.

I am very grateful to the financial support from University of Michigan Department of Industrial and Operations Engineering, the Federal Aviation Administration, the Rackham Graduate School, and the Toyota Research Institute. Their support enabled me to conduct my research work and share my findings through academic conferences.

I was very fortunate to have the support from the staff members in the Department of Industrial and Operations Engineering. I would like to thank Christopher Konrad, Olof (Mint) Minto, and Charles Woolley for their continued technical assistance throughout my experiments,

thank Matt Irelan, Tina Picano Sroka, Teresa Maldonado and Wanda Dobberstein for all kinds of support that made everything on this journey smooth and enjoyable.

I would like to thank my lab mates in the THInC lab – Brandon Pitts, Julie Prinet, Yuzhi Wan, Keven Lieberman, and Karanvir Panesar – who provided invaluable support over these five years. I would also like to thank my friends I met in Ann Arbor, special thanks go to Ke Liu, Na Du and Albert Fu for bringing me lots of fun during this journey.

Lastly, I would like to thank my parents and my boyfriend for their unconditional love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

The introduction of automation technologies to various application domains has resulted in improved efficiency and precision of operations. However, it has also created challenges. One problem that has recently received considerable attention is trust miscalibration, i.e., a mismatch between a person's trust in automation and the actual capabilities and reliability of the system. Trust miscalibration can refer to too much or too little trust in a system which, in turn, leads to misuse (e.g., overreliance) or disuse (e.g., slow adoption) of automation. Avoiding these undesirable outcomes requires a better understanding of, and support for trust calibration – the focus of the proposed research. To date, most studies on trust suffer from a number of limitations, including highly intrusive techniques for measuring trust, a limited understanding of important factors affecting the process of trust development and a lack of effective countermeasures to deal with trust miscalibration.

To address these shortcomings, the goals of this dissertation were to (1) develop an eye-tracking based technique to infer trust levels and variations in real time, (2) identify how the magnitude and duration of changes in system reliability affect the process of trust evolution and calibration, and (3) develop and evaluate the effectiveness of a real-time intervention (an audio alert) for supporting trust calibration. To this end, a series of empirical studies were conducted in the context of an Unmanned Arial Vehicle (UAV) control simulation. Participants were required to perform two tasks in parallel: a tracking task and a target detection task, the latter with the assistance of an imperfectly reliable automated system. Subjective trust measures, eye movements, behavioral data, and performance outcome data were recorded.

As expected, participants in the first study monitored low-reliability UAVs more closely, as indicated by a set of eye-tracking metrics. Variations in their monitoring behavior aligned with their subjective trust ratings, suggesting that eye tracking is indeed a promising less intrusive technique for inferring trust in real time. The second experiment showed that participants were sensitive to four types of UAV reliability changes that differed with respect to magnitude and duration. Both duration and, even more so, magnitude affected participants' trust calibration and recovery. The large and long reliability drop in system performance had the most severe negative impact on trust and target detection performance. To prevent performance breakdowns due to trust miscalibration, an eye-tracking based trust inference system using a k-Nearest Neighbor algorithm was developed and used in Experiment 3 to trigger audio alerts in case of a divergence between system reliability and participant trust in the system. The audio alert was successful in improving trust calibration and contributed to faster trust recovery following a period of low system performance. However, these improvements did not translate into better performance on the target detection task.

The findings from this dissertation add to the knowledge base in trust in automation. At a methodological level, a new nonintrusive was developed and validated. At a conceptual level, a better understanding of the effects of variations in the magnitude and duration of system reliability changes on trust and performance was gained. And at an applied level, a candidate countermeasure to trust miscalibration was designed and tested. Ultimately, this research helps prevent catastrophic consequences due to inappropriate reliance on automation, and thus contributes to safer operations in a wide range of application domains.

Chapter 1

Introduction

The introduction of automation technologies to various application domains has resulted in improved efficiency and precision of operations (Breton & Bossé, 2003). However, it has also created challenges. One important problem that has recently received considerable attention is trust miscalibration, i.e., a mismatch between a person's trust in automation and the actual capabilities and reliability of the system (Lee & See, 2004). Trust miscalibration can refer to too much or too little trust in a system which, in turn, leads to misuse or disuse of automation (Parasuraman & Riley, 1997). Misuse involving overreliance on technology has contributed to incidents and accidents in high-risk domains such as aviation, military operations and nuclear power plants. For instance, in 2013, Asiana flight 214 struck the sea wall during its approach to San Francisco International Airport. Three passengers were fatally injured in this accident. According to the NTSB investigation, pilots' overreliance on the automation was one important factor contributing to this accident. The pilots 'trusted' that the automation would provide speed protection at all times and therefore did not intervene when, in fact, the auto-throttle system had entered the so-called HOLD mode in which it no longer prevented the airspeed from falling below a minimum value. In contrast, a lack of trust has led to the slow adoption, and even complete rejection of highly automated systems. For example, in the domain of healthcare, automated infection prevention surveillance was found to help reduce 61% of the time infection control professionals (ICPs) spent on surveillance activities; yet, only 23%~56% of surveyed facilities adopted the automated surveillance system. A lack of understanding of how the system worked made it difficult to for ICPs to establish trust and rely on the system (Hebden, 2015). Avoiding these undesirable outcomes requires a better understanding of, and support for trust calibration – the focus of the proposed research. To date, most studies on trust suffer from a number of limitations, including disruptive and biased techniques for measuring trust, a limited understanding of important factors affecting the process of trust development and a lack of effective countermeasures to deal with trust miscalibration. To address these shortcomings, this research **addressed** three main objectives:

1) Develop an eye tracking based technique for inferring trust in real time in a nonintrusive manner
2) Study how the degree and duration of variations in system reliability affect the calibration and levels of trust in automated systems over time
3) Evaluate the effectiveness of real-time audio alerts for improving trust calibration and proper use of automation

This chapter will provide an overview of studies on trust in human-automation interaction and identify gaps in the literature. Eye tracking will then be introduced as a promising way to infer trust in technology in real time. Finally, to overcome the issue of trust miscalibration, a potential countermeasure using both training and audio alerts will be described.

Trust in automation

Human-automation interaction

Automation can be defined as “the execution by a machine agent (usually a computer) of a function that was previously carried out by a human” (Parasuraman & Riley, 1997). The development and wide application of automation systems has created numerous benefits: automation is not vulnerable to cognitive limitations such as fatigue, stress and bias; physical restrictions in extreme working environment are no longer an issue; automation has also improved operation precision and work efficiency remarkably (Parasuraman & Miller, 2004). At the same time, the inappropriate use of automated systems – both misuse and disuse of modern technologies - has led to unexpected or undesirable outcomes, including incidents and accidents (Parasuraman & Riley, 1997). Misuse refers to operators over-relying on “imperfectly reliable” automation while disuse happens when operators neglect or underutilize automation even when it works well. Several studies have shown that trust in automation is an important factor affecting automation usage (Hoff & Bashir, 2015; Lee & Moray, 1994; Lee & See, 2004).

Trust

Trust is a concept that was originally developed and studied in the field of social psychology but soon gained interest in other research areas, such as sociology, economics, political science and, most recently, ergonomics. Barber (1983) defined interpersonal trust along three dimensions: persistence of natural and moral laws (expectation of constancy), technically competent performance (skill-, rule-, knowledge-based abilities that a trustor expects from a trustee), and fiduciary responsibility (trustees’ duties to place others’ interests before their own). Later, Rempel et al. (1985) focused on the dynamic nature of human-human trust and proposed predictability (dominating trust in early stages), dependability (dominating trust later) and faith (orienting to future trust) as three main aspects of interpersonal trust. Muir (1987, 1994) was the first to systematically study trust between humans and machines. Based on the assumption that trust in machines resembles trust in humans, she integrated the above six dimensions to build a model of human-machine trust. However, it was soon noted that there are indeed important differences between human-human and human-machine trust (Madhavan & Wiegmann, 2007b; Meyer & Lee, 2013). For example, Lee and See (2004) who define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" proposed that one major difference is that "automation lacks intentionality", which implies that the machine does not exhibit features like loyalty and dependability, features that form the basis for the development of human-human trust. The authors built a dynamic model of trust and reliance on automation. This model identifies three critical factors affecting trust development: 1) the closed-loop dynamics of trust and reliance, 2) the importance of context on trust and on mediating the effect of trust on reliance, and 3) the role of information display on developing appropriate trust. According to this model, trust, along with other attitudes (such as self-confidence or perceived risk) determines people's intention to rely on automation. This intention, combined with factors such as time constraints, affects how people actually rely on the automation. The actual interaction between people and automation, in turn, affects how much people trust the system, because trust is largely dependent on the observation of actual automation behavior. A study by Jian et al. (2000) highlighted that people are more likely to distrust a machine, compared to human beings. They explained that this difference is, in part, due to the fact that interpersonal trust tends to build up over a long period of time while there is often little time for humans to develop trust in automation (Lee & Moray, 1994).

Ideally, operators' trust in automation should be calibrated, i.e., it should be proportional to the trustworthiness or reliability of the system. In reality, however, there is often "a mismatch between the perceived and the actual system performance and capabilities", an issue called trust miscalibration (Lee & Moray, 1994; Muir, 1987). Miscalibration relates to two important dimensions of trust, its resolution and its specificity (Lee & See, 2004, See Figure 1.1). Resolution refers to how well the range of automation capabilities is reflected in the range of operator trust (Cohen, Parasuraman, & Freeman, 1998). Specificity includes both functional and temporal specificity. Functional specificity describes the degree to which trust is associated with a particular function of the system. Temporal specificity, the focus in this research, refers to "changes in trust as a function of the situation or over time" (Lee & See, 2004). Both trust resolution and trust specificity highlight the dynamic nature and context-sensitivity of trust in automation. However, most studies to date have treated trust as a steady-state, rather than time-variant variable (Yang, Unhelkar, Li, & Shah, 2017). These studies have made important contributions by identifying a wide range of factors that affect trust levels (Hoff & Bashir, 2015; Schaefer, Chen, Szalma, & Hancock, 2016) but supporting trust calibration will require a better understanding of the process of trust formation and change (Khastgir, Birrell, Dhadyalla, & Jennings, 2017; G. H. Walker, Stanton, & Salmon, 2016; Yang, Wickens, & Hölttä-Otto, 2016; Yu et al., 2017).

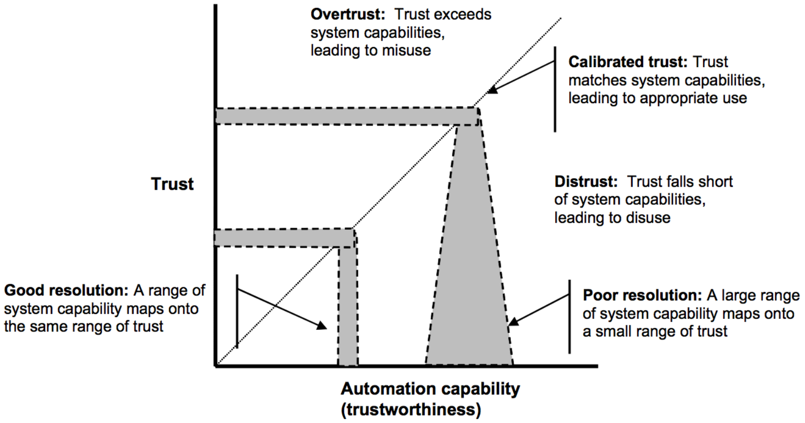

Overtrust: Trust exceeds system capabilities, leading to misuse
Calibrated trust: Trust matches system capabilities, leading to appropriate use
Trust
Distrust: Trust falls short of system capabilities, leading to disuse
Good resolution: A range of system capability maps onto the same range of trust
Poor resolution: A large range of system capability maps onto a small range of trust
Automation capability (trustworthiness)

Figure 1.1. The relationship among trust calibration, trust resolution, and automation capability/trustworthiness (Lee & See, 2004)

Factors affecting trust

Trust in human-machine teams can be affected by many factors. A literature review of 30 trust-related studies divided these factors into three categories: human-related, environment-related and automation-related (Schaefer et al., 2016). Human-related factors can be divided further into traits (e.g., age, gender, ethnicity etc.), emotive factors (e.g., attitudes towards automation, comfort/confidence/satisfaction with automation), states (e.g., stress, fatigue), and cognitive factors (e.g., ability to use automation, expectation). Factors related to the environment are team collaboration (e.g., role interdependence, team composition) and task/context (e.g., risk/uncertainty, physical environment). Automation-related factors include the features and capability of automation, such as its appearance, reliability, feedback and behavior. The latter were found to be the main contributors to the development of trust in human-machine teams, followed by human-related factors.

Among **automation-related** factors, system reliability has received the most attention and appears to be the most critical factor in shaping operators' trust in a system (Lee & Moray, 1994; Lee & See, 2004; Schaefer et al., 2016). Not surprisingly, highly reliable automation gains more trust than low reliability automation (Bailey & Scerbo, 2007; Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003; Madhavan, Wiegmann, & Lacson, 2006; Walliser, de Visser, & Shaw, 2016). Note that operators' standard for considering a system reliable and their tolerance for automation error varies as a function of the potential consequences of a failure. For example, if a life saving device, such as robotic medical device, is not 100% reliable, the device is considered untrustworthy while other less safety critical technologies may be trusted even when they are only 70% reliable. Operators' trust is affected not only by the actual reliability of a system but also by how reliable they perceive the system to be. Perceived reliability can differ from actual reliability when operators do not notice mistakes or when a system performs as designed but not as expected by the operator (e.g., Christopher D Wickens, 1995). Finally, the timing of poor system performance also matters. If a system fails early on during human-machine interaction, the negative effect on trust tends to be much more severe compared to a performance breakdown happening at a later stage (Desai et al., 2012; Robinette, Howard, & Wagner, 2017).

While the effects of system reliability on trust have been studied quite extensively, research in this area still involves some notable limitations. In most studies, system reliability was varied as a between-subject factor (e.g., Dzindolet, Pierce, Beck, Dawe, & Anderson, 2001; Lee & Moray, 1994), and trust was measured only one time - at the end of the experiment. Very few studies have examined how changes in system reliability- in particular, the magnitude and duration of system reliability changes - affect longer-term trust development at an individual level. Another shortcoming of most studies is that they focused on the effects of automation failure on trust decrement (Ezer, Fisk, & Rogers, 2008; Wiegmann, Rich, & Zhang, 2001); little is known about the trust recovery process once system performance returns to a high level. Finally, it remains unclear how system reliability interacts with other factors, such as priming, a process where exposure to one stimulus influences the response to subsequent stimuli, without the individual necessarily being aware of this effect (Tulving & Schacter, 1990).

In addition to automation-related factors, **human-related factors** can affect trust development. Among them, culture (Huerta, Glandon, & Petrides, 2012), age (Ho, Wheatley, & Scialfa, 2005), and trust propensity (Merritt, Heimbaugh, LaChapell, & Lee, 2013) contribute to dispositional trust (an individual's enduring tendency to trust automation) (Hoff & Bashir, 2015), while expectation (Mayer, Sanchez, Fisk, & Rogers, 2006), experience with a system (Sanchez, Rogers, Fisk, & Rovira, 2014; Yuviler-Gavish & Gopher, 2011), and understanding of a system (Manzey, Reichenbach, & Onnasch, 2012) contribute to learned trust, which is known as an operator's evaluation of the automation based on past experience or current interaction (Hoff & Bashir, 2015). The present research considers two human-related factors, expectation and training, because they are more easily manipulated in real world environments than factors that affect dispositional trust (Mayer, 2008).

Expectation is a factor that is embedded in the definition of trust in automation. Muir (1994) extended the definition of interpersonal trust to human/machine relationships and proposed that trust development depended on the realization of three types of expectations: 1) persistence (predictable behavior), 2) technical competence (to work properly, to be competent), 3) fiduciary responsibility (morality considerations). When operators interact with an automated system, the initial expectation for automation capabilities was compared to the actual automation performance, and then contributed to the dynamic trust evolution process (Merritt & Ilgen, 2008). Some research addressed the importance of delivering true capabilities and limitations of the automated system, so that operators can have the right expectations and are more likely to gain well-calibrated trust in automation (Khastgir, Birrell, Dhadyalla, & Jennings, 2018). Furthermore, experiment results from several studies showed that operators with high automation expectation (the expectation that automation is trustworthy) were more sensitive to changes (both increases and decreases) in automation reliability than participants with low automation expectation (Mandell, 2018; Pop, Shrewsbury, & Durso, 2015).

Providing training to operators in advance can alter operators' trust and reliance patterns. On the one hand, training increases operators' familiarity with the automation and promote automation use. For example, in a driving study, participants who were both informed about and trained on how to use an auto-parking system showed higher trust compared to participants who were only told about how to employ the auto-parking feature (Tenhundfeld et al., 2019). Training can serve also as a preventive intervention to mitigate the negative effects of automation misuse. For example, if operators receive training about when an automated system would not function well before they started using it, they would show less trust in that certain situation but unquestioningly accept the automation decision aid in a general situation (Masalonis, 2003). Another study revealed that operators who were exposed to potential automation failures during training were less likely to have complacency in the automation (Bahner, Hüper, & Manzey, 2008). These research findings suggested that training can either promote trust or suspend trust. However, it is still not clear what should be conveyed and taught in the training in order to best support trust calibration.

Trust calibration

Ideally, operators' trust in automation should be calibrated, i.e., it should be proportional to the trustworthiness or reliability of the system. To date, attempts to promote appropriate trust in automation fall into two main categories: 1) influencing trust formation early on, through training and instruction, in a top-down fashion and 2) providing alerts and real-time feedback on system performance throughout human-machine interaction, and thus shape trust in a bottom-up fashion.

One promising means of affecting trust formation early on appears to be priming. Priming refers to a process where exposure to one stimulus influences the response to subsequent stimuli, without the individual necessarily being aware of this effect (Tulving & Schacter, 1990). Some studies have confirmed that shaping an operator's general attitude towards automation through priming alters the trust formation process during actual automation usage (Abe & Richardson, 2006; Ezer, Fisk, & Rogers, 2007; Lacson, Wiegmann, & Madhavan, 2005; Pop et al., 2015). For instance, different framings of why the system made error can lead to different levels of trust even though the aid's reliability was the same among different groups (Bisantz & Seong, 2001). Participants in the sabotage group (where they were told that the automation was subject to intentional attack from the enemy) were more likely to disuse automation compared to the control group. Clare et al. (2015) also found that priming is an effective way of manipulating operators' trust levels. However, their results highlight that the effect of priming may last for a limited time only. Research has also shown that the specific content of priming needs to be tailored to each automated system and its application environment in order to best benefit operators. In an automated driving study, one group of participants received trust-promoted introduction, and another group received trust-lowered introduction. Even in the risky environment where take-over is needed to avoid collision, participants in the trust-promoted group still trusted the automation and collided with the obstacle when driving the autonomous vehicle (SAE level 3), while no collision happened in the trust-lowered group (Körber, Baseler, & Bengler, 2018), suggesting that trust-lowered introduction was more appropriate for trust calibration in this case.

Another means of shaping trust in and reliance on automation is to provide real-time continually updated feedback about system reliability and trustworthiness. For instance, in an experiment conducted by McGuirl and Sarter (2006), continuously updated system confidence information for a neural network based decision support system (DSS) was provided to pilots who experienced and had to respond quickly to two types of icing conditions in a flight simulator. The information helped pilots decide when to trust and rely on the DSS's diagnosis of aircraft icing, versus when to engage in the icing identification task themselves. This resulted in less misuse of the system and better performance (i.e., fewer stalls), compared to when no confidence information was available. In another study on trust in and reliance on an automated combat identification (CID) system (Wang, Jamieson, & Hollands, 2009), reliability information for the system was presented to participants. Results indicated that this information did contribute to more appropriate reliance on the CID system. However, participants' performance was improved only in the 80% reliability condition, not in the 67% reliability condition, suggesting that the effectiveness of this approach may depend on the actual level of automation reliability.

Kraus and his colleagues (Kraus, Scholz, Stiegemeier, & Baumann, 2019) integrated past research findings on trust development and proposed a model to illustrate the dynamic trust calibration process, as shown in Figure 1.2. Their model highlights the fact that "trust is calibrated in accordance with information provided both prior to and during the interaction with an automated system". Very few studies have combined these two means of supporting trust calibration, a shortcoming that was addressed in the present research.

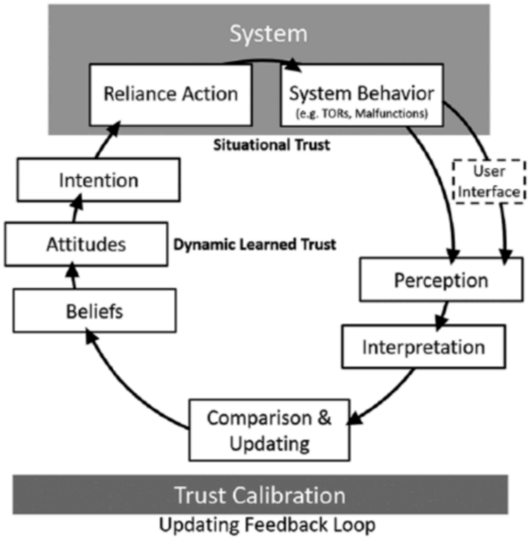


Figure 1.2. Dynamic trust calibration model

Trust measurement

To date, several measures of trust have been developed and employed. These measures fall into three main categories: subjective ratings, behavioral data and physiological data.

Subjective ratings

Many trust-related studies rely heavily on subjective rating scales (e.g., Bagheri & Jamieson, 2004; Barg-Walkow & Rogers, 2016; Merritt & Ilgen, 2008), and many different scales have been developed (e.g., Chavaillaz, Wastell, & Sauer, 2016; Madsen & Gregor, 2000; Manzey et al., 2012; Sauer, Chavaillaz, & Wastell, 2016). In a systematic review of empirical research on trust in automation over the past ten years (Hoff and Bashir, 2015), the "Checklist of Trust between People and Automation (CTPA)" (Jian, Bisantz, & Drury, 2000), a 12-item questionnaire with 7-point Likert scale ratings, was found to be the most commonly used self-reporting scale. Sample items include "The system is reliable", "The system behaves in an underhanded manner", and "I am confident in the system". CTPA was developed and tested through a series of experiments and is considered a valid method for inferring human trust in automation (Clare, Cummings, & Repenning, 2015). Still, some researchers created their own questionnaires to better serve their specific research purposes and settings (e.g., Manzey et al., 2012; Merritt, Lee, Unnerstall, & Huber, 2015).

Subjective rating scales have the advantage of being easy to implement and employ. When properly constructed, subjective ratings can be reliable and valid indicators of people's trust (Moray & Inagaki, 1999). However, responses may not be comparable across different people. For example, a self-reported trust rating of '60' may not reflect the same actual level of trust for different participants. Another shortcoming of rating scales is that they tend to be employed only once, often at the end of an experiment (Bagheri & Jamieson, 2004; Pop et al., 2015; Rovira, McGarry, & Parasuraman, 2007). This approach fails to capture how trust evolves over time due to priming or variations in system performance. Some studies have tried to measure trust changes over time by repeatedly asking for subjective ratings at short intervals throughout an experiment (Clare et al., 2015; Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013; Holliday, Wilson, & Stumpf, 2016; Yang et al., 2017). This technique can capture the temporal evolution of trust to some extent, through successive 'snapshots', but it is quite disruptive of task performance and thus undesirable for use in simulation studies and in real-world environments. Another problem with repeatedly asking for trust ratings in a laboratory setting is that participants may speculate that system reliability has changed whenever they are asked for a rating, which leads to biased responses.

Behavioral data

Some studies try to infer trust levels from observable operator/participant behavior. This approach is based on the assumption that trust, which is considered an attitude, forms the basis for intentions (Lee & See, 2004) and, ultimately, mediates behavior (Dzindolet et al., 2003; Dzindolet, Pierce, Beck, & Dawe, 2002). Reliance and compliance are two behavioral measures that have been widely used in past trust research (Chancey, Bliss, Yamani, & Handley, 2017). Compliance refers to an operator taking action in accordance with a system-generated warning while reliance denotes the absence of an operator action when the system indicates an intact system or a normal situation (Dixon, Wickens, & McCarley, 2007; Meyer, 2001, 2004). In addition to these generic behaviors, some studies have employed more domain- or task-specific measures. For instance, in studies related to autonomous vehicles, when a driver reverts to manual control of the vehicle, this is assumed to be an indicator of a loss of, or reduction in trust (Brown & Noy, 2004). Similarly, when operators interact with adaptable automation, switching to lower levels of autonomy may signal low or reduced trust (Manzey et al., 2012). As with subjective ratings, behavioral measures tend to be outcome-oriented and discrete. This makes it impractical to trace moment-to-moment changes in trust which is important for studying trust calibration, including both resolution and specificity.

Most research on trust combines subjective ratings with behavioral data. In some studies, these measures were found to be positively correlated (Dzindolet et al., 2003; Muir & Moray, 1996). However, other experiments obtained inconsistent findings using both techniques (Rovira et al., 2007; Satterfield, Baldwin, de Visser, & Shaw, 2017). For instance, Miller et al. (2016) conducted an experiment to measure drivers' trust in an autonomous vehicle. They recorded both drivers' behavior (taking manual control of the car) and their trust ratings. The experiment showed that drivers intervened with the behavior of the autonomous car even when they expressed that, subjectively, they trusted the system. The current research will investigate how various trust-related factors may lead to such disassociations between subjective ratings and behavioral data.

Physiological data

Physiological measures such as electroencephalography (EEG), galvanic skin response (GSR), and functional magnetic resonance imaging (fMRI) have been used widely to assess mental workload (Haapalainen, Kim, Forlizzi, & Dey, 2010) and affective states (Lichtenstein, Oehme, Kupschick, & Jürgensohn, 2008). More recently, studies have highlighted the promise of inferring trust from these physiological signals (de Visser et al., 2018; Jung, Dong, & Lee, 2019). For example, observational error positivity (oPe), a type of neural marker that can be obtained from EEG, was found to be positively related to subjective trust ratings (de Visser et al., 2018). Research has also tried to build models of human trust in modern technologies using EEG and GSR signals (Ajenaghughrure, Sousa, Kosunen, & Lamas, 2019; Hu, Akash, Jain, & Reid, 2016). In both studies, it was found that, once participants' trust reached a stable status, the model could achieve a mean accuracy of over 70% to predict people's trust level in real time. Though physiological data can be collected and analyzed in real-time, setting up for data collection can be rather difficult and time consuming. For example, when collecting EEG data, participants have to wear a cap with embedded electrodes that need to be filled with an abrasive gel, a set-up that is quite intrusive and uncomfortable.

In summary, a more process-oriented, unobtrusive, real-time technique for measuring trust is needed in order to better understand trust development and promote trust calibration. The next section will describe eye tracking as the proposed method to achieve these goals.

Eye tracking: a promising means of tracing trust development in real time

Eye tracking fundamentals

An eye tracker is a sensor device that can trace where people are looking in real time using infrared light. There are two main types of eye trackers on the market: desktop-mounted (screen-based) and head-mounted (wearable). In the current research, a wearable device, Tobii Pro Glasses 2, was used to collect participants' eye movement data in an unobtrusive way. Participants are wearing this eye tracker like a regular pair of glasses and are thus able to move their heads freely during an experiment. Absolute points of gaze (POGs) are obtained from the raw eye movement data (Majaranta & Bulling, 2014). They indicate the spatial locations people are looking at in a scene, in the form of x, y coordinates and a corresponding timestamp.

The raw eye movement gaze points recorded by an eye tracker can be used to determine two basic elements in eye tracking research: fixations and saccades. Fixations refer to those times when the eyes remain relatively stable and take in new information from the area the eyes are focusing on (Jacob, 1995; Rayner, 1995, 2009). A fixation is usually composed of multiple gaze points. Saccades refer to rapid eye movements between fixations (Jacob, 1995; Jacob & Karn, 2003; Salvucci & Goldberg, 2000). It is assumed that people do not acquire information during a saccade. Fixations and saccades tend to be studied in relation to areas of interest (AOI), which are specific regions defined by researchers based on the purpose of the study or based on tasks, as shown in Figure 1.3.

Several techniques have been developed to identify fixations and saccades (Salvucci, 1999). The fixation filter used by Tobii Pro Analyzer is based on the Velocity-Threshold identification fixation filter (I-VT) algorithm developed by Salvucci & Goldberg (2000). This I-VT filter is an algorithm that classifies fixations and saccades based on the velocity of eye shifts. Specifically, Tobii Pro Analyzer uses a threshold of 100 degrees/second to identify fixations. This means that if the velocity between two consecutive gaze points is smaller than 100 degrees/second, they are considered to belong to the same fixation. A moving velocity between two points above this threshold are categorized as saccades. The filter also merges adjacent fixation points (with a visual angle smaller than 0.5 degree) and discards short fixations that last for less than 60 ms to reduce noise and disturbances in classification (Olsen, 2012b).

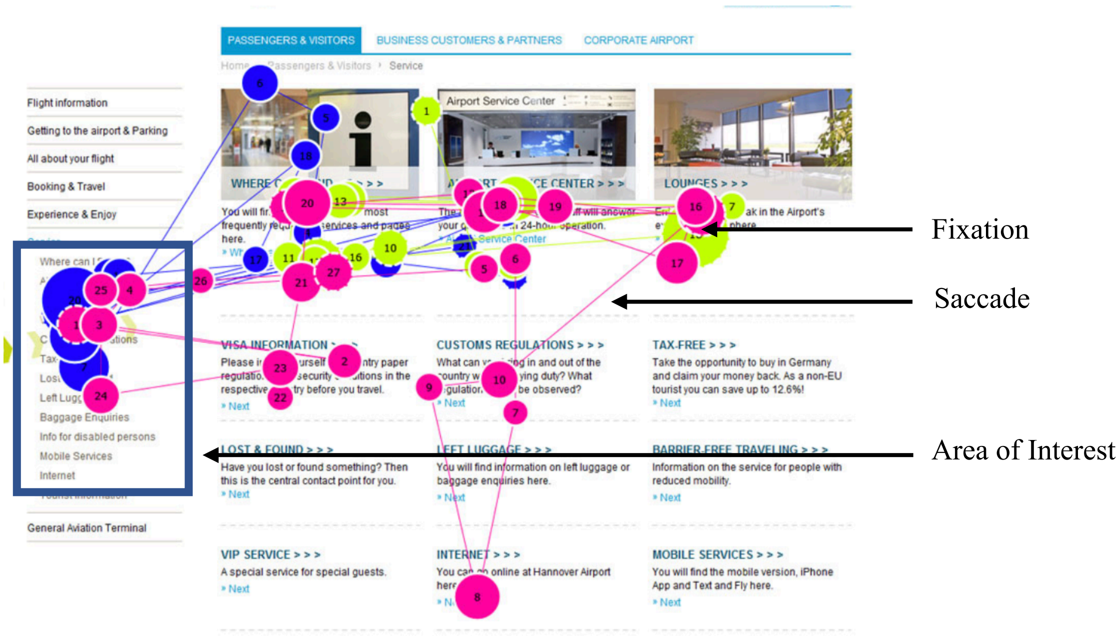


Figure 1.3. Illustration of fixations, saccades and an Area of Interest (AOI)

Eye tracking applications and benefits

Eye tracking has been employed successfully in a variety of disciplines including psychology, neuroscience, marketing, human factors and computer science (Duchowski, 2002). In the area of human factors, eye movement data have been used to gain insight into people's perception and cognitive processes, such as decision making (Vachon & Tremblay, 2014), cognitive workload (Marshall, 2002; Peißl, Wickens, & Baruah, 2018) and situation awareness (Barnard & Lai, 2010; Moore & Gugerty, 2010). For example, eye tracking metrics including scan path and eye fixations help predict the time spent on making a decision, and pupil dilation can be used as an indicator of the decision-making quality (Vachon & Tremblay, 2014). Pupil size and blink rate have been used widely to assess cognitive workload (Niezgoda, Tarnowski, Kruszewski, & Kamiński, 2015; Tsai, Viirre, Strychacz, Chase, & Jung, 2007). In aviation, research has shown that fixation rates and dwell times are well suited for measuring level 1 situation awareness (low-level perceptual processes) and scanning entropy (also referred to as 'Nearest Neighbor Index') appears to be a good indicator of level 3 situation awareness (high-level cognitive process) (Van De Merwe, Van Dijk, & Zon, 2012).

One of the main benefits of using eye tracking is its non-invasive nature (Rayner, 1998) which makes it possible to employ eye tracking in simulation and field studies (Ji, Zhu, & Lan, 2004; Kircher, Ahlstrom, & Kircher, 2009). Another advantage is that eye tracking data can be collected and processed data in real-time. This is a prerequisite for identifying and counteracting problematic mental states (such as inappropriate levels of workload or trust) in a timely fashion. For example, in a study focusing on attentional narrowing (an involuntary reduction in the attentional scope), eye tracking was used to detect the onset of this state in a multitasking environment in real time. A display adaptation was then triggered to broaden participants' attention field and mitigate the negative effects of attentional narrowing (Prinet, 2016).

Eye tracking, attention management, trust and automation usage

Eye tracking is considered by many to be the most direct way to measure people's attention allocation (e.g., Shinar, 2008). The underlying assumption is that a person's direction of gaze is an approximation of or closely related to what information their attention is focused on (Hoffman & Subramaniam, 1995; Shepherd, Findlay, & Hockey, 1986). This assumption is commonly referred to as the "eye–mind hypothesis" (Just & Carpenter, 1980). Although some have argued that attention allocation and eye movements may be completely independent of each other (Klein & Pontefract, 1994; Posner, 1980), Rayner (2009) showed that, at least in the context of performing tasks such as visual search or reading, it is quite difficult and impractical to have eye movements and attention focus on different locations or content.

Eye tracking has been widely used in the study of attention allocation and automation usage (Diez et al., 2001). For example, eye tracking data combined with behavioral data were used to identify shortcomings in pilots' understanding and monitoring of highly automated commercial flight deck systems (Sarter, Mumaw, & Wickens, 2007). Eye tracking has also been employed to study complacency and overreliance on automation due to overtrust in the system. Complacency has been shown to result in operators monitoring the automation less often than optimal which can lead to breakdowns in joint system performance (Moray & Inagaki, 2000; Parasuraman & Manzey, 2010). For example, in a study on air traffic control, it was observed that air traffic controllers who failed to detect traffic conflicts showed significantly fewer fixations on the radar display when in automation mode, compared to manual mode; in contrast, fixation frequencies for controllers who reliably detected conflicts did not differ between the two modes of operation (Metzger & Parasuraman, 2005). This connection between monitoring behavior and performance was confirmed in several other studies (Bagheri & Jamieson, 2004; C. Wickens, Dixon, Goh, & Hammer, 2005).

While the relation between (visual) attention allocation, automation usage and performance has been studied extensively (e.g., Cullen, Rogers, & Fisk, 2013; Onnasch, Ruff, & Manzey, 2014), there is limited empirical evidence for the feasibility and validity of using eye tracking to infer human trust in automation even though trust has been shown to alter operators' attention and monitoring strategies (Bailey & Scerbo, 2007; Molloy & Parasuraman, 1996). For instance, Hergeth and colleagues (Hergeth, Lorenz, Vilimek, & Krems, 2016) collected eye tracking data to quantify drivers' trust in a highly automated car. Gaze behavior was measured in terms of the frequency of monitoring (defined as the number of fixations on the driving scene) during a particular non-driving-related task (NDRT) and the monitoring ratio (defined as the sum of fixation durations in a NDRT, scaled to the duration of that NDRT). Subjective trust ratings were collected also and compared with the eye tracking results. The findings support the assumption of a negative correlation between trust and monitoring frequency. In other words, participants indeed monitored automation-related displays more often when the system was trusted less, and vice versa.

As mentioned above, the benefits of using eye tracking for measuring trust are its non-intrusive nature (compared to repeated subjective trust ratings) and its ability to trace moment-to-moment changes in trust in real time. More research is needed, however, to establish which eye tracking metrics are particularly well suited for capturing trust levels and variations. Moacdieh and Sarter (2015) proposed three categories of eye tracking metrics that relate to different aspects of monitoring and information search: location/spread, directness and duration. Among these three categories, location/spread (e.g., number of fixations) and duration metrics (e.g., mean fixation duration) are likely best suited for inferring trust as low trust levels can be expected to lead to more frequent and longer fixations on agents and displays. Another open question is to what extent and how fast variations in trust translate into changes in people's attention allocation as expressed by changes in eye movement data.

Adaptation algorithm

Using eye tracking as a real-time, unobtrusive measure of trust is a prerequisite for detecting and counteracting trust miscalibration and avoid resulting performance breakdowns. Another requirement for achieving this goal is the development of an eye tracking based algorithm that can function as a classifier of a person's cognitive state. To date, several modeling, machine learning or deep learning techniques have been used for this purpose. These techniques include but are not limited to logistic regression, random forest, k-nearest neighbor, multilayer perceptron, and convolutional neural networks.

Eye tracking metrics extracted from raw eye movement data have proved to work well as inputs to these different modeling methods. For example, pupil diameter was used successfully to predict user performance in real-time using a random forest method (Buettner, Sauer, Maier, & Eckhardt, 2018). Scanpath length and fixation duration were able to predict the time taken to make a decision (in a simulated setting to classify an aircraft as hostile or not) using logistic regression (Vachon & Tremblay, 2014). In a study using eye tracking data to detect personality traits, seven classifying methods (AdaBoost, Decision Tree, Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine, and k-Nearest Neighbor) were applied to ten unique features extracted from the raw eye movement data. Naive Bayes turned out to achieve the highest accuracy in identifying personality traits (Berkovsky et al., 2019).

To date, most studies trying to infer people's trust in a system have used physiological data such as EEG or GSR as input to train a classification model. For instance, using features extracted from both EEG and GSR data, a study found that a general trust sensor model could be created with a quadratic discriminant classifier that achieved an average accuracy rate of 75% in predicting trust levels (Akash, Hu, Jain, & Reid, 2018). In another study, employing EEG signals, deep learning methods such as Convolutional Neural Networks (CNN) were able to assess trust calibration and outperform other machine learning methods such as Multi-Layer Perceptron (MLP) and Deep Neural Network (DNN) (Choo et al., 2019). However, to our knowledge, there is very limited research using eye tracking data as input and applying the above modeling methods to infer human operator's trust levels in real time.

Summary

Trust miscalibration remains a major challenge for safe and effective human-machine teaming and collaboration. The overall goal of this dissertation is therefore to develop an eye-tracking based technique to detect trust miscalibration and address the problem and thus prevent performance breakdowns in real time. To achieve this goal, a series of research activities and empirical studies were conducted that will be described in the following chapters. Chapter 2 reported on an experiment that established the feasibility and validity of using eye tracking to infer trust levels in real time. The study described in Chapter 3 examined how the magnitude and duration of variations in system reliability affect trust evolution and calibration, and it established which combination of those two factors affected trust calibration and performance the most. In Chapter 4, an eye-tracking based algorithm for classifying trust levels is developed which is then used in the experiment in Chapter 5 to trigger an audio alert in case of trust miscalibration. The alert serves to redirect participants' attention allocation and improve overall system performance. Finally, Chapter 6 summarizes the findings from this line of research and suggests potential future work to address limitations of the current effort.

References

Abe, G., & Richardson, J. (2006). Alarm timing, trust and driver expectation for forward collision warning systems. *Applied ergonomics, 37*(5), 577-586.

Ajenaghughrure, I. B., Sousa, S. C., Kosunen, I. J., & Lamas, D. (2019). *Predictive model to assess user trust: a psycho-physiological approach.* Paper presented at the Proceedings of the 10th Indian Conference on Human-Computer Interaction.

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using eeg and gsr. *ACM Transactions on Interactive Intelligent Systems (TiiS), 8*(4), 27.

Bagheri, N., & Jamieson, G. A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced "complacency.". *Human performance, situation awareness, and automation: Current research and trends*, 54-59.

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66*(9), 688-699.

Bailey, N., & Scerbo, M. (2007). Automation-induced complacency for monitoring highly reliable systems: the role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science, 8*(4), 321-348.

Barg-Walkow, L. H., & Rogers, W. A. (2016). The effect of incorrect reliability information on expectations, perceptions, and use of automation. *Human factors, 58*(2), 242-260.

Barnard, Y., & Lai, F. (2010). *Spotting sheep in Yorkshire: Using eye-tracking for studying situation awareness in a driving simulator.* Paper presented at the HUMAN FACTORS: A SYSTEM VIEW OF HUMAN, TECHNOLOGY AND ORGANISATION. ANNUAL CONFERENCE OF THE EUROPE CHAPTER OF THE HUMAN FACTORS AND ERGONOMICS SOCIETY 2009.

Berkovsky, S., Taib, R., Koprinska, I., Wang, E., Zeng, Y., Li, J., & Kleitman, S. (2019). *Detecting Personality Traits Using Eye-Tracking Data.* Paper presented at the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Bisantz, A. M., & Seong, Y. (2001). Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics, 28*(2), 85-97.

Breton, R., & Bossé, É. (2003). *The cognitive costs and benefits of automation*. Retrieved from

Brown, C. M., & Noy, Y. I. (2004). Behavioural adaptation to in-vehicle safety measures: Past ideas and future directions. *Traffic and transport psychology: Theory and application*, 25-46.

Buettner, R., Sauer, S., Maier, C., & Eckhardt, A. (2018). Real-time Prediction of User Performance based on Pupillary Assessment via Eye Tracking. *AIS Transactions on Human-Computer Interaction, 10*(1), 26-56.

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. (2017). Trust and the Compliance–Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence. *Human factors, 59*(3), 333-345.

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied ergonomics, 52*, 333-342.

Choo, S., Sanders, N., Kim, N., Kim, W., Nam, C. S., & Fitts, E. P. (2019). *Detecting Human Trust Calibration in Automation: A Deep Learning Approach.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Clare, A. S., Cummings, M. L., & Repenning, N. P. (2015). Influencing Trust for Human–Automation Collaborative Scheduling of Multiple Unmanned Vehicles. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(7), 1208-1218.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). *Trust in decision aids: A model and its training implications.* Paper presented at the in Proc. Command and Control Research and Technology Symp.

Cullen, R. H., Rogers, W. A., & Fisk, A. D. (2013). Human performance in a multiple-task environment: Effects of automation reliability on visual attention allocation. *Applied ergonomics, 44*(6), 962-968.

de Visser, E. J., Beatty, P. J., Estepp, J. R., Kohn, S., Abubshait, A., Fedota, J. R., & McDonald, C. G. (2018). Learning from the slips of others: Neural correlates of trust in automated agents. *Frontiers in human neuroscience, 12*.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). *Impact of robot failures and feedback on real-time trust.* Paper presented at the Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., . . . Yanco, H. (2012). *Effects of changing reliability on trust of robot systems.* Paper presented at the Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on.

Diez, M., Boehm-Davis, D. A., Holt, R. W., Pinney, M. E., Hansberger, J. T., & Schoppek, W. (2001). *Tracking pilot interactions with flight management systems through eye movements.* Paper presented at the Proceedings of the 11th International Symposium on Aviation Psychology.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human factors, 49*(4), 564-572.

Duchowski, A. T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers, 34*(4), 455-470.

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.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human factors, 44*(1), 79-94.

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.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2007). *Reliance on automation as a function of expectation of reliability, cost of verification, and age.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2008). Age-related differences in reliance behavior attributable to costs within a human-decision aid system. *Human factors, 50*(6), 853-863.

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). *Psycho-physiological measures for assessing cognitive load.* Paper presented at the Proceedings of the 12th ACM international conference on Ubiquitous computing.

Hebden, J. N. (2015). Slow adoption of automated infection prevention surveillance: Are human factors contributing? *American journal of infection control, 43*(6), 559-562.

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human factors, 58*(3), 509-519.

Ho, G., Wheatley, D., & Scialfa, C. T. (2005). Age differences in trust and reliance of a medication management system. *Interacting with Computers, 17*(6), 690-710.

Hoff, K. A., & Bashir, M. (2015). Trust in automation integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(3), 407-434.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & psychophysics, 57*(6), 787-795.

Holliday, D., Wilson, S., & Stumpf, S. (2016). *User trust in intelligent systems: A journey over time.* Paper presented at the Proceedings of the 21st International Conference on Intelligent User Interfaces.

Hu, W.-L., Akash, K., Jain, N., & Reid, T. (2016). Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine, 49*(32), 48-53.

Huerta, E., Glandon, T., & Petrides, Y. (2012). Framing, decision-aid systems, and culture: Exploring influences on fraud investigations. *International Journal of Accounting Information Systems, 13*(4), 316-333.

Jacob, R. J. (1995). Eye tracking in advanced interface design. *Virtual environments and advanced interface design*, 258-288.

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye* (pp. 573-605): Elsevier.

Ji, Q., Zhu, Z., & Lan, P. (2004). Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE transactions on vehicular technology, 53*(4), 1052-1068.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53-71.

Jung, E.-S., Dong, S.-Y., & Lee, S.-Y. (2019). Neural Correlates of Variations in Human Trust in Human-like Machines during Non-reciprocal Interactions. *Scientific reports, 9*(1), 1-10.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological review, 87*(4), 329.

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2017). Calibrating trust to increase the use of automated systems in a vehicle. In *Advances in Human Aspects of Transportation* (pp. 535-546): Springer.

Khastgir, S., Birrell, S., Dhadyalla, G., & Jennings, P. (2018). Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation research part C: emerging technologies, 96*, 290-303.

Kircher, K., Ahlstrom, C., & Kircher, A. (2009). Comparison of two eye-gaze based real-time driver distraction detection algorithms in a small-scale field operational test.

Klein, R. M., & Pontefract, A. (1994). 13 Does Oculomotor Readiness Mediate Cognitive Control of Visual Attention? Revisited! *Attention and performance XV: Conscious and nonconscious information processing*, 333.

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied ergonomics, 66*, 18-31.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2019). The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human factors*, 0018720819853686.

Lacson, F. C., Wiegmann, D. A., & Madhavan, P. (2005). *Effects of attribute and goal framing on automation reliance and compliance.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153-184.

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

Lichtenstein, A., Oehme, A., Kupschick, S., & Jürgensohn, T. (2008). Comparing two emotion models for deriving affective states from physiological data. In *Affect and emotion in human-computer interaction* (pp. 35-50): Springer.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277-301.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human factors, 48*(2), 241-256.

Madsen, M., & Gregor, S. (2000). *Measuring human-computer trust.* Paper presented at the 11th australasian conference on information systems.

Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based human–computer interaction. In *Advances in physiological computing* (pp. 39-65): Springer.

Mandell, A. (2018). *An Investigation of the Individual Differences and Causal Attributions That Make or Break Dynamic Trust in Automation.* George Mason University,

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 1555343411433844.

Marshall, S. P. (2002). *The index of cognitive activity: Measuring cognitive workload.* Paper presented at the Proceedings of the IEEE 7th conference on Human Factors and Power Plants.

Masalonis, A. J. (2003). *Effects of training operators on situation-specific automation reliability.* Paper presented at the SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483).

Mayer, A. K. (2008). *The manipulation of user expectancies: Effects on reliance, compliance, and trust using an automated system.* Georgia Institute of Technology,

Mayer, A. K., Sanchez, J., Fisk, A. D., & Rogers, W. A. (2006). *Don't let me down: The role of operator expectations in human-automation interaction.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don’t know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors, 55*(3), 520-534.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(2), 194-210.

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human factors, 57*(1), 34-47.

Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human factors, 47*(1), 35-49.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human factors, 43*(4), 563-572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human factors, 46*(2), 196-204.

Meyer, J., & Lee, J. D. (2013). Trust, reliance, and compliance.

Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human factors, 38*(2), 311-322.

Moore, K., & Gugerty, L. (2010). *Development of a novel measure of situation awareness: The case for eye movement analysis.* Paper presented at the Proceedings of the human factors and ergonomics society annual meeting.

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control, 21*(4-5), 203-211.

Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science, 1*(4), 354-365.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27*(5-6), 527-539.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429-460.

Niezgoda, M., Tarnowski, A., Kruszewski, M., & Kamiński, T. (2015). Towards testing auditory–vocal interfaces and detecting distraction while driving: A comparison of eye-movement measures in the assessment of cognitive workload. *Transportation research part F: traffic psychology and behaviour, 32*, 23-34.

Olsen, A. (2012). The Tobii I-VT fixation filter. *Tobii Technology*.

Onnasch, L., Ruff, S., & Manzey, D. (2014). Operators′ adaptation to imperfect automation–Impact of miss-prone alarm systems on attention allocation and performance. *International Journal of Human-Computer Studies, 72*(10-11), 772-782.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors, 52*(3), 381-410.

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

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 39*(2), 230-253.

Peißl, S., Wickens, C. D., & Baruah, R. (2018). Eye-tracking measures in aviation: a selective literature review. *The International Journal of Aerospace Psychology, 28*(3-4), 98-112.

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual differences in the calibration of trust in automation. *Human factors, 57*(4), 545-556.

Posner, M. I. (1980). Orienting of attention. *Quarterly journal of experimental psychology, 32*(1), 3-25.

Prinet, J. (2016). *Attentional Narrowing: Triggering, Detecting and Overcoming a Threat to Safety.*

Rayner, K. (1995). Eye movements and cognitive processes in reading, visual search, and scene perception. In *Studies in visual information processing* (Vol. 6, pp. 3-22): Elsevier.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin, 124*(3), 372.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology, 62*(8), 1457-1506.

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

Robinette, P., Howard, A. M., & Wagner, A. R. (2017). Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems, 47*(4), 425-436.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human factors, 49*(1), 76-87.

Salvucci, D. D. (1999). *Mapping eye movements to cognitive processes*: Carnegie Mellon University Pittsburgh, PA.

Salvucci, D. D., & Goldberg, J. H. (2000). *Identifying fixations and saccades in eye-tracking protocols.* Paper presented at the Proceedings of the 2000 symposium on Eye tracking research & applications.

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science, 15*(2), 134-160.

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data. *Human factors, 49*(3), 347-357.

Satterfield, K., Baldwin, C., de Visser, E., & Shaw, T. (2017). *The Influence of Risky Conditions in Trust in Autonomous Systems.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Sauer, J., Chavaillaz, A., & Wastell, D. (2016). Experience of automation failures in training: effects on trust, automation bias, complacency and performance. *Ergonomics, 59*(6), 767-780.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors, 58*(3), 377-400.

Shepherd, M., Findlay, J. M., & Hockey, R. J. (1986). The relationship between eye movements and spatial attention. *The Quarterly Journal of Experimental Psychology, 38*(3), 475-491.

Shinar, D. (2008). Looks are (almost) everything: where drivers look to get information. *Human factors, 50*(3), 380-384.

Tenhundfeld, N. L., de Visser, E. J., Haring, K. S., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2019). Calibrating Trust in Automation Through Familiarity With the Autoparking Feature of a Tesla Model X. *Journal of Cognitive Engineering and Decision Making, 13*(4), 279-294.

Tsai, Y.-F., Viirre, E., Strychacz, C., Chase, B., & Jung, T.-P. (2007). Task performance and eye activity: predicting behavior relating to cognitive workload. *Aviation, space, and environmental medicine, 78*(5), B176-B185.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247*(4940), 301-306.

Vachon, F., & Tremblay, S. (2014). What eye tracking can reveal about dynamic decision-making. *Advances in cognitive engineering and neuroergonomics, 11*, 157.

Van De Merwe, K., Van Dijk, H., & Zon, R. (2012). Eye movements as an indicator of situation awareness in a flight simulator experiment. *The International Journal of Aviation Psychology, 22*(1), 78-95.

Walker, G. H., Stanton, N. A., & Salmon, P. (2016). Trust in vehicle technology. *International journal of vehicle design, 70*(2), 157-182.

Walliser, J. C., de Visser, E. J., & Shaw, T. H. (2016). *Application of a System-Wide Trust Strategy when Supervising Multiple Autonomous Agents.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

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

Wickens, C., Dixon, S., Goh, J., & Hammer, B. (2005). *Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis*. Retrieved from

Wickens, C. D. (1995). Designing for situation awareness and trust in automation. *IFAC Proceedings Volumes, 28*(23), 365-370.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science, 2*(4), 352-367.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). *Evaluating Effects of User Experience and System Transparency on Trust in Automation.* Paper presented at the Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction.

Yang, X. J., Wickens, C. D., & Hölttä-Otto, K. (2016). *How users adjust trust in automation: Contrast effect and hindsight bias.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). *User trust dynamics: An investigation driven by differences in system performance.* Paper presented at the Proceedings of the 22nd International Conference on Intelligent User Interfaces.

Yuviler-Gavish, N., & Gopher, D. (2011). Effect of descriptive information and experience on automation reliance. *Human factors, 53*(3), 230-244.

Chapter 2

Eye Tracking: A Process-oriented Method for Inferring Trust in Automation as A Function of Priming and System Reliability

The ultimate goals of this research are (1) to better understand the process of trust development in automated systems and (2) to create a technique for detecting and addressing trust miscalibration in order to avoid the misuse and disuse of modern technology. One prerequisite for achieving these goals is the availability of an unobtrusive technique for measuring or inferring trust in real time. There is increasing evidence that eye tracking is well suited for this purpose (e.g., Petersen, Robert, Yang, & Tilbury, 2019; Strauch et al., 2019; F. Walker, Verwey, & Martens, 2018. Operators' gaze behavior while monitoring an automated system has been shown to be highly associated with subjective trust ratings (Hergeth et al., 2016). However, studies that documented this association involve a number of limitations. First, they explored the diagnosticity of very few eye tracking metrics, mainly the number of fixations (count metrics) and fixation duration (temporal metrics). Other eye tracking metrics, such as mean saccade amplitude or backtrack rate, which are related to different (in this case, spatial) search and monitoring features, are worth investigating as low trust levels might lead to less efficient and organized search behavior when collaborating with an automated system. Also, studies suggesting that eye tracking is a valid technique for inferring trust were conducted almost exclusively in the context of automated driving. Research is needed to establish the effectiveness of eye tracking in other application domains as the selection and calculation of eye tracking metrics are usually display-dependent and can vary based on tasks and settings. To address the above shortcomings, the experiment reported in this chapter compared eight different eye tracking metrics to determine how well each captures levels and changes in trust resulting from variations in system reliability and priming. The application domain for this study was military intelligence gathering with the help of multiple unmanned aerial vehicles (UAVs). Multi-UAV control imposes considerable attentional demands on operators. It involves high levels of time pressure and uncertainty and thus is an appropriate setting for studying human trust in automation.

Specifically, eye tracking and other trust-related measures were used in this study to assess the long-term effects of priming on trust, and how priming may interact with, and possibly be overridden by observations of actual system performance during human interaction with automation. Participants experienced both high and low reliability automation. Half of the participants were informed about the automation reliability in advance of each trial (priming condition) while the other half were not provided with this information (no priming). Dependent measures included eye tracking data, subjective trust ratings, and performance on the target detection task (including response times and error rates).

Method

Participants

Thirty-five University of Michigan undergraduate and graduate students participated in the experiment. Data from three participants were excluded due to malfunctions of the eye tracker or incomplete data. The 32 participants whose data were included in the data analysis were between the ages of 18 to 35 years (Mean=24.50, SD=3.86). 18 participants were males. None of the participants had any experience with Unmanned Aerial Vehicle (UAV) control before this study. Because the eye tracker used in this experiment cannot be worn with additional eyewear (such as prescription glasses), all participants were required to have normal or corrected-to-normal vision (contact lenses were allowed). This study was approved by the University of Michigan Institutional Review Board (IRB Reference ID: HUM00126745).

Apparatus and tasks

The application domain for this study was military reconnaissance and intelligence gathering with the assistance of unmanned aerial vehicles (UAVs). A UAV simulation replicating military target identification tasks was developed in our laboratory, as shown in Figure 2.1.

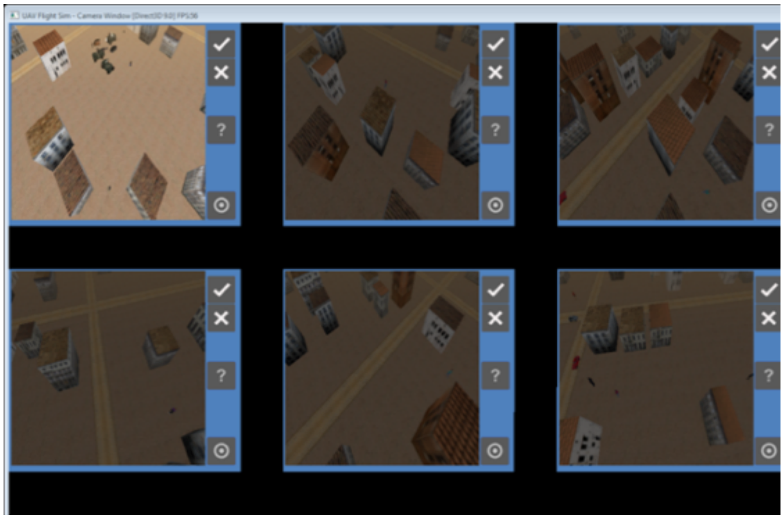


Figure 2.1. UAV simulation (video feed highlighted if automation detects a target)

During the 30-minute scenario, automation onboard six UAVs scanned pre-defined regions to help with the detection of a military target (a truck carrying a gun; see Figure 2.2). The simulated video feeds from the six UAVs were displayed in a 2*3 grid on a single 27” monitor display. They included both actual targets and similar looking non-targets, as shown in Figure 2.3.

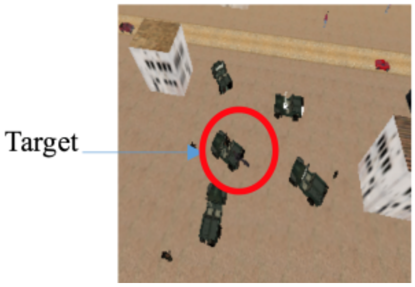


Figure 2.2. Target example

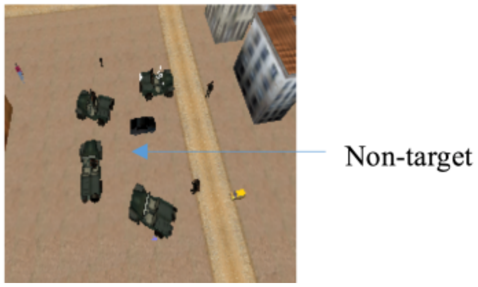


Figure 2.3. Non-Target example

When a UAV identified a possible target, its video feed was highlighted (see Figure 2.1). The participant then reviewed the scene and pressed one of two buttons to either confirm (√) or reject (X) the presence of a target. In addition, participants were asked to scan the various UAV video feeds on a continuing basis to make sure no targets were missed. If a target was missed by a UAV but detected by a participant, s/he pressed a third ‘target’ (⊙) button to record the event. Participants were instructed to press the ‘?’ button if they were uncertain about the presence of a target, independent of whether the screen was highlighted or not.

Experiment design

The experiment employed a 2 (reliability: high, low) * 2 (priming: reliability information, no reliability information) full factorial design. Automation reliability was a within-subject factor. Half of the UAVs were highly reliable (95% correct) while the other three UAVs were only 50% reliable. The low level was selected based on past trust research which set low reliability at 50-70% (e.g., Chavaillaz, Wastell & Sauer, 2016; Chugh & Caird, 1999; Wiegmann, Rich & Zhang, 2001). The lowest value in this range - 50% - was chosen in this experiment to create a large enough difference between low and high reliability and thus ensure that participants would reliably notice low system reliability and adjust their trust accordingly. For half of the participants, the upper three UAVs were the highly reliable ones whereas for the other half of the participants, the three highly reliable UAVs were shown at the bottom of the screen. Detailed information about the performance of the high and low reliability UAVs is shown in Table 2.1. The total number of hits and false alarms was the same for the two levels of automation reliability to avoid biasing participants' attention allocation. 95% reliable automation only made false alarms to make sure that participants could detect these errors and understand that the automation was not perfectly reliable.

Table 2.1 Overall UAV reliability and corresponding numbers of hits, correct rejections, false alarms and misses

(Hit: UAV successfully detects a target and the screen is highlighted. Correct rejection: UAV correctly determines that an object is not a target and the screen is not highlighted. False alarm: UAV incorrectly identifies an object as a target and the screen is highlighted. Miss: UAV fails to detect a target and the screen is not highlighted)

Reliability = 95%		Reliability = 50%	
Hits	35	Hits	30
Correct rejections	51	Correct rejections	15
False alarms	4	False alarms	9
Misses	0	Misses	36

Priming was a between-subject factor. Half of the participants were informed about the overall reliability of the two groups of UAVs in advance of the experiment while the other half did not receive any reliability information. The latter group of participants assessed system reliability solely based on their observations of UAV performance throughout the experiment.

The dependent measures in this experiment included eye tracking data, performance on the target detection task (including response times and error rates), and subjective trust ratings. Throughout the experiment, participants were prompted every two minutes to rate their trust in the upper three UAVs and the lower three UAVs on a scale of 0 ("I do not trust these UAVs at all") to 9 ("I completely trust these UAVs") using the keyboard. Once participants completed the two ratings, the target detection task automatically resumed.

Eye movement data were collected using Tobii Pro Glasses 2 and the Tobii Pro Lab software. The sampling rate of the eye tracking glasses was 50 Hz. The raw eye tracking data were used to calculate the eight metrics listed in Table 2.2. These metrics fall into three commonly used categories (Goldberg & Kotval, 1999; Lai et al., 2013) : (1) temporal metrics (2) spatial metrics, and (3) count metrics. The two temporal metrics in this study were total and average fixation duration. A fixation is defined as "a relatively stable eye-in-head position within some threshold of dispersion (typically ~2°) over some minimum duration (typically 100-200 ms), and with a velocity below some threshold (typically 15-100 degrees per second)" (Jacob & Karn, 2003). The fixation filter in the Tobii Pro Lab Analyzer is based on the velocity of the directional shifts of the eye. The default threshold is 100 degrees/second. The 4 spatial metrics used in this study – mean saccade amplitude, backtrack rate, rate of transitions, and scanpath length per second – relate to the efficiency and randomness of the search and scanpath (Moacdieh & Sarter, 2017). Efficiency refers to how quickly and easily participants can locate and detect a target and has traditionally been measured by response time (e.g., Beck, Lohrenz & Trafton, 2010). Finally, the two count metrics capture the number/frequency of fixations and transitions between areas of interest (AOI). An area of interest (AOI) is defined by the experimenter as the specific area for which eye movement data are being analyzed. In this experiment, two sets of AOIs were used: (1) for the purpose of studying the effects of automation reliability and the relationship between eye tracking and subjective trust ratings, the three upper and the three lower UAV windows, respectively, were considered separate AOIs, and (2) to examine participants' scan patterns in more detail, each of the six UAV windows was defined as a separate AOI. The definition of each eye tracking metric and predictions on how the metrics are associated with different trust levels can be found in Table 2.2.

Table 2.2. Eye tracking metrics calculated in this study

Metric	Definition	Prediction
Temporal metrics		
Total fixation duration(s)	The total time each participant fixated each AOI	Low system reliability results in low trust which, in turn, is expected to lead to longer fixation duration for affected UAVs (both average and total). Priming will, at least initially, result in a larger difference in fixation duration between low and high reliability automation.
Average fixation duration (s)	The average duration of the fixations within each AOI	
Spatial metrics		
Mean saccade amplitude (pixel)	The average amplitude of all saccades	Low trust levels resulting from low system reliability will lead to less efficient and organized search behavior. The no-priming group needs time to build up appropriate trust levels. Therefore, their search behavior will initially be more random and less efficient (longer mean saccade amplitude, larger backtrack rate, larger transition rate between AOIs and longer scanpath length per second).
Backtrack rate (/s)	The number of saccade angles larger than 90 degrees, divided by the total time	
Rate of transitions (/s)	The number of transitions between AOIs, divided by the total time	
Scanpath length per second (pixel/s)	The total length of the scanpath, divided by the total time	
Count metrics		
Total fixation count	The number of fixations within each AOI	Participants are likely to monitor low reliability UAVs more often; therefore, the total fixation count is expected to be larger for those vehicles. Participants are also likely to switch more often between the three low-reliability UAV windows to detect misses, resulting in a higher transition count. Priming information will further differentiate participants' trust in low and high reliability UAVs, leading to more efficient search behavior (smaller total fixation counts and transition counts).
Transition count	The number of transitions between AOIs	

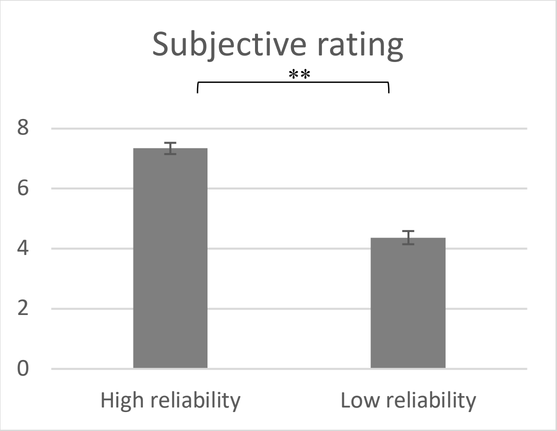
Experiment procedure

Each experiment session started with participants being informed that the goal of the experiment was to study trust in human-automation interaction. Participants then read and signed the consent form. The eye tracker was calibrated, and participants were trained on the target identification task for 10 minutes. At the end of the training, participants were expected to correctly identify at least 90% of all targets. If they did not meet this criterion, they received additional training until their performance was acceptable. Participants in the priming group were informed about the overall reliability of the two groups of UAVs (95% vs 50%) in advance of the experiment. However, they were not told about the distribution of particular types of errors (false alarms versus misses), nor were they informed that there would be no misses in the high reliability condition. Participants in the no priming group did not receive any information about overall system reliability. Following the 30-minute experiment session, a debriefing was conducted to ask participants for feedback about various aspects of the experiment, such as their overall monitoring strategy and the effectiveness of the automatic highlighting of UAV windows.

Results

Subjective trust ratings

Subjective trust ratings were analyzed using a 2 (Reliability: high vs. low) *2 (Priming: reliability information provided vs. not) linear mixed model. The significance level was set at 0.05. Participants rated the highly reliable UAVs (Mean=7.34, SD=1.06) as significantly more trustworthy than the low reliability UAVs (Mean=4.37, SD=1.24, F (1,30) =112.71, p <0.001), as shown in Figure 2.4. There was no significant effect of priming, nor was there an interaction between reliability and priming (see Figure 2.5).



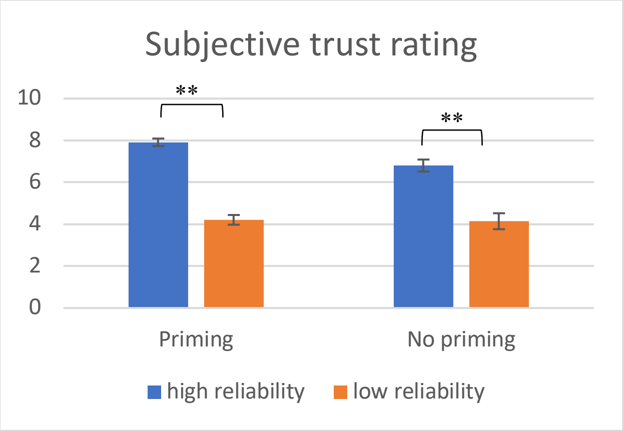


Figure 2.4. Subjective trust ratings as a function of reliability

(**: *p*<0.01)

Figure 2.5. Subjective trust ratings as a function of priming and reliability

(*: *p*<0.05)

Eye Tracking Metrics

Eye tracking metrics were analyzed with a 2 (Reliability: high vs. low) *2 (Priming: reliability information provided vs. not) linear mixed model. The participant number was entered as a random effect. The significance level was set at 0.05.

The analysis revealed a significant main effect of automation reliability on one of the two temporal eye tracking metrics (see Table 2.3). Longer total fixation durations were observed for low reliability UAVs (Mean = 817.75, SD = 168.65), compared to high reliability UAVs (Mean = 650.06, SD=160.82; *F* (1, 30) =18.12, *p*<0.001). Average fixation duration remained unchanged. Except for mean saccade amplitude, all spatial metrics were significantly affected by system reliability. Participants' backtrack rate was significantly higher in the low reliability condition (Mean=0.09, SD=0.07), compared to the high reliability condition (Mean=0.05, SD=0.04, F(1,30)=12.16, *p* =0.001). The rate of transitions for highly reliable UAVs (Mean=0.37, SD=0.16) was significantly smaller than for low reliability UAVs (Mean=0.46, SD=0.17, F(1,30)=5.06, *p* =0.028). And participants' scanpath length per second was significantly shorter for highly reliable UAVs, compared to low reliability UAVs (F(1,30)=5.24, *p* =0.029). Finally, both count metrics changed significantly as a function of reliability: fixation counts were significantly higher for low reliability UAVs (Mean=2277.63, SD=669.72), compared to high reliability UAVs (Mean=2679.75, SD=469.49; *F* (1,30) = 8.78, *p*=0.004), and transition counts were significantly larger for low reliability UAVs (Mean=975.12, SD=374.36) than for highly reliable vehicles (Mean=781.91, SD=330.17, F(1,30)=5.21, *p*=0.03).

Table 2.3. Effects of system reliability on eye tracking metrics

Metric	High reliability	Low reliability	Main effects of reliability
Temporal metrics			
Total fixation duration(s)	650.06(160.82)	817.65(168.71)	F(1,30)=18.12, p<0.001
Average fixation duration (s)	0.30(0.06)	0.31(0.06)	Not significant
Spatial metrics			
Mean saccade amplitude (pixel)	31.92(9.67)	32.09(11.20)	Not significant
Backtrack rate (/s)	0.05(0.04)	0.09(0.07)	F(1,30)=12.16, p=0.001
Rate of transitions (/s)	0.37(0.16)	0.46(0.17)	F(1,30)=5.06, p=0.028
Scanpath length per second (pixel/s)	522.07(221.79)	646.20(248.03)	F(1,30)=5.24, p=0.029
Count metrics			
Total fixation count	2277.63(669.72)	2679.75(469.49)	F(1,30)=8.78, p=0.004
Transition count	781.91(330.17)	975.12(374.36)	F(1,30)=5.21, p=0.030

A significant main effect of priming was found for one of the two temporal metrics (see Table 2.4). The average fixation duration was longer for the priming group (Mean=0.33, SD=0.08) than for the no priming group (Mean=0.27, SD=0.03, F (1, 30) = 8.45, *p* = 0.007). A significant effect of priming was observed also for the three spatial metrics. In the priming group, mean saccade amplitude was significantly shorter (Mean=27.89, SD=10.25), compared to the no priming condition (Mean=36.12, SD=8.88, F(1,30)=6.56, *p* =0.016); the rate of transitions (Mean=0.36, SD=0.16) was significantly smaller, compared to participants who had no information about UAV reliability (Mean=0.47, SD=0.16, F(1,30)=7.24, *p* =0.009); and the scanpath length per second was significantly longer (Mean=518.36, SD=244.38, F(1,30)=4.89, *p* =0.035). As for count metrics, the priming group showed a smaller fixation count (Mean=2293.41, SD=639.752), compared with the no priming group (Mean=2663.97, SD=521.78, F (1,30) = 7.46, *p*=0.008); however, the second count metric – transition counts – did not change as a function of priming.

Table 2.4. Effects of priming on eye tracking metrics

Metric	Priming	No priming	Main effects of priming
Temporal metrics			
Total fixation duration(s)	744.32(225.41)	723.49(133.36)	Not significant
Average fixation duration (s)	0.33(0.08)	0.27(0.04)	F(1,30)=8.45, p=0.007
Spatial metrics			
Mean saccade amplitude (pixel)	27.89(10.25)	36.12(8.88)	F(1,30)=6.56, p=0.016
Backtrack rate (/s)	0.07(0.05)	0.07(0.05)	Not significant
Rate of transitions (/s)	0.36(0.16)	0.47(0.16)	F(1,30)=7.24, p=0.009
Scanpath length per second (pixel/s)	518.36(244.38)	649.91(223.63)	F(1,30)=4.89, p=0.035
Count metrics			
Total fixation count	2293.41(639.75)	2663.97(521.78)	F(1,30)=7.46, p=0.008
Transition count	799.44(383.83)	957.59(328.91)	Not significant

There was also a significant interaction between automation reliability and priming for total fixation duration (F (1,60) = 7.50, *p*=0.008). A simple effect analysis showed that total fixation duration (F(1,60)=24.47), *p* <0.001) differed significantly as a function of automation reliability only for the priming group where it was longer for low reliability vehicles (see Figure 2.6).

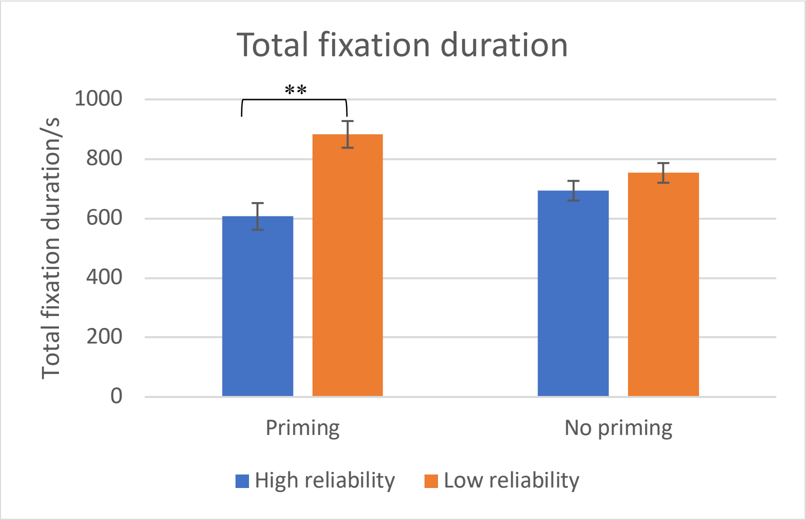


Figure 2.6. Interaction effect between priming and reliability for fixation duration (**: *p*<0.01)

Relationship between eye tracking metrics and subjective trust ratings

To validate the eye tracking metrics, we calculated their correlations with the subjective trust ratings, both for all participants combined and separately for the priming and no priming groups (see Table 2.5). The eye tracking metrics showed a significant negative correlation with subjective trust ratings, with two exceptions. There was no significant correlation between mean saccade amplitude and subjective trust ratings, and average fixation duration was not correlated with subjective ratings in the no priming group.

Table 2.5. Correlations between eye tracking metrics and subjective ratings

	Total fixation duration (s)	Average fixation duration (s)	Mean saccade amplitude (pixel)	Backtrack rate (/s)	Rate of transitions (/s)	Scanpath per second (pixel/s)	Total fixation count	Transition count
Subjective trust ratings (Overall)	-0.918**	-0.433*	0.058	-0.858**	-0.821**	-0.793**	-0.796**	-0.783**
Subjective trust ratings (Priming)	-0.928**	-0.487**	0.012	-0.839**	-0.852**	-0.837**	-0.849**	-0.829**
Subjective trust ratings(No priming)	-0.799**	-0.188	0.035	-0.687**	-0.565**	-0.621**	-0.556**	-0.491**

*: *p* <0.05, **: *p* <0.01

Changes in eye tracking metrics as a function of short-term variations in actual system reliability

The overall reliability for the two groups of UAVs was 95% (group 1, high reliability) versus 50% (group 2, low reliability), respectively. It varied slightly for each vehicle throughout the experiment (every 2 minutes; range: 0-100%) (see Table 2.6). A correlation analysis was conducted on the percentage differences for the various eye tracking metrics between the high- and low-reliability UAVs to determine whether these reliability variations were reflected in short-term changes in attention allocation. Only total fixation count was found to be correlated with reliability changes over time (r=0.685, *p*=0.005), as shown in Figure 2.7.

Table 2.6. Actual reliability settings for high (Group 1) and low (Group 2) reliability automation

	Interval 1	Interval 2	Interval 3	Interval 4	Interval 5
Group 1	1	1	1	1	1
Group 2	0.5	0.4	0.375	0.5	0.625
	Interval 6	Interval 7	Interval 8	Interval 9	Interval 10
Group 1	0.88	0.67	1	1	0.8
Group 2	0.67	0.4	0.75	0.286	0.6
	Interval 11	Interval 12	Interval 13	Interval 14	Interval 15
Group 1	1	1	0.857	1	1
Group 2	0.57	0.625	0	0.4	0.56

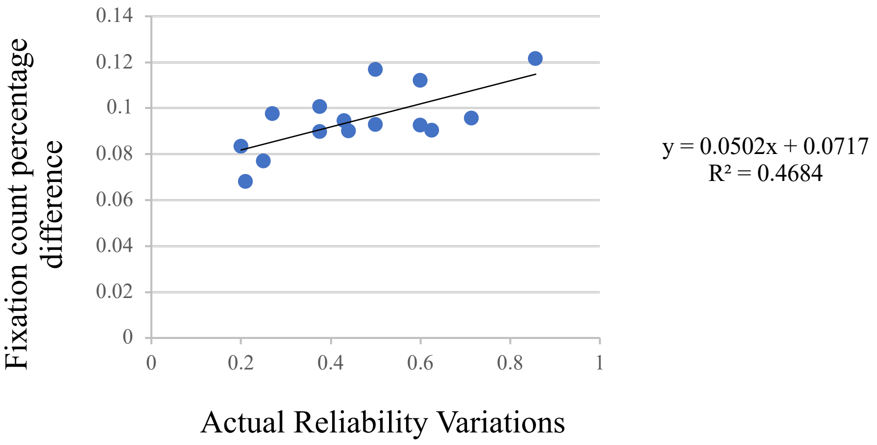


Figure 2.7. Correlation between fixation count and actual reliability variations

Performance on target detection task

Reaction time

Response time was defined as the time between the first appearance of the target/non-target on the screen and the participant's button press (confirm, reject, uncertain, miss). A 2 (Reliability: high vs. low) *2 (Priming: reliability information provided vs. not) mixed linear model showed a significant main effect for UAV reliability. Participants' response time for low reliability UAVs (Mean=2.14s, SD=0.17) was significantly slower than for high reliability UAVs (Mean=1.88s, SD=0.18; F(1,30)=205.07, *p*<0.001). Response time did not differ significantly between the priming and the no priming groups, and no interaction was observed between priming and reliability.

Overall error rate

The overall error rate on the target detection task was defined as the total number of errors, divided by the total number of trials in that condition. A 2 (Reliability: high vs. low) *2 (Priming: reliability information provided vs. not) mixed linear model revealed a significant main effect of reliability on error rate. Participants made more errors when interacting with the low reliability UAVs (Mean=9%, SD=0.005), compared to the highly reliable vehicles (Mean=2.7%, SD=0.005; F(1,30)=73.60, *p*<0.001). No significant effect of priming and no interaction effect were found.

Error types

There were five possible types of errors that participants could make during the experiment, as listed in Table 2.7. A 2 (Reliability)*2 (Priming) mixed linear model was conducted on participants' error rates for each of these categories. Results indicated a significant main effect of reliability on error types 3 (missing a target that was missed by the UAV), 4 (false alarms) and 5 (failure to respond when the UAV window was highlighted). Participants showed significantly higher type 3 and 5 error rates (Mean=12.5%, SD=0.072 and Mean=5.6%, SD=0.0126, respectively) for low reliability UAVs, compared to high reliability UAVs where no such errors were observed. The type 4 error rate was also significantly higher (Mean=13.5%, SD=0.055, F(1,30)=147.71, *p*<0.001) when participants interacted with low reliability UAVs, compared to the highly reliable UAVs (Mean=1.5%, SD=0.033). No significant effect of priming and no interaction effect between priming and reliability were found.

Table 2.7. Error definition and effect of reliability

Error type	Definition	Error rate - high reliability	Error rate - low reliability	Significance

Error 1	When the sub-screen was highlighted and there was a target, the participant clicked “reject” button.	M=0.031
		SD=0.027
			M=0.032
			SD=0.036
				Not significant

Error 2	When the sub-screen was highlighted and there was not a target, the participant clicked “confirm” button.	M=0.094
		SD=0.177
			M=0.070
			SD=0.100
				Not significant

Error 3	When the sub-screen was not highlighted and there was a target, the participant didn’t click “miss” button.	M=0
		SD=0
			M=0.125
			SD=0.072
				F(1,30)=94.92, *p*<0.001

Error 4	When the sub-screen was not highlighted and there was not a target, the participant clicked “miss” button.	M=0.015
		SD=0.033
			M=0.135
			SD=0.055
				F(1,30)=147.71, *p*<0.001

Error 5	When the sub-screen was highlighted, the participant didn’t click any button.	M=0.0056
		SD=0.0126
			M=0
			SD=0
				F(1,30)=6.39, *p*=0.017

Discussion

The purpose of this study was to develop and validate an eye tracking-based method for inferring and tracing, in real time, changes in operator trust levels as a function of automation reliability and priming. A total of eight different eye tracking metrics were calculated from raw eye movement data. They fall into three categories: temporal, spatial, and count metrics. To validate these eye tracking metrics, they were correlated with subjective trust ratings, the traditional means of assessing trust levels. Participants' performance on a UAV control task was recorded to determine whether and how it was affected by changes in attention allocation resulting from different levels of trust.

Effects of changes in system reliability on eye movements and monitoring

The analysis of the eye tracking data revealed that low automation reliability was associated with longer total fixation durations, higher backtrack and transition rates, an increased scanpath length per second, and higher total fixation and transition counts.

Longer fixation durations were expected for low system reliability as participants would trust the automation less and therefore examine potential targets more carefully. This effect was observed for total, but not for average fixation duration. One possible explanation for this finding is that, while participants visited low reliability UAV windows more often (resulting in longer total fixation durations), they were able to examine potential targets equally quickly for both high and low reliability UAVs (translating into comparable average fixation durations) since the appearance of the targets was identical.

As predicted, most of the spatial metrics were significantly affected by system reliability, except for mean saccade amplitude. Backtrack rate, transition rate and scanpath length per second can be considered indicators of the efficiency of visual search and scanning which suffered as a result of low automation reliability (Goldberg & Kotval, 1999). This may be explained by the high attentional load imposed and by a high level of uncertainty of where to look, leading to less systematic monitoring behavior. The fact that mean saccade length was not affected by reliability may be explained by earlier findings showing that this metric is particularly sensitive to mental effort (Chen, Epps, Ruiz, & Chen, 2011). Mental effort has been defined as "the amount of attentional demand that participants allocated to the task (how hard they were trying) (Vicente, Thornton & Moray, 1987)". All targets in this experiment for both high and low reliability UAVs were identical, and thus likely required the same amount of effort. Other metrics such as heart-rate variability (Aasman, Mulder & Mulder, 1987) or pupillary dilation (Kahneman, 1973) could be used to validate whether participants' mental effort was indeed the same when detecting targets with different groups of UAVs.

Finally, system reliability significantly affected both count metrics. This result confirms our expectations and is consistent with earlier research findings (Bagheri & Jamieson, 2004; Moray & Inagaki, 2000; F. Walker et al., 2018). When people trust automation to perform a task reliably, they will monitor the system less frequently (as expressed by fixation counts). The number of transitions between the low reliability UAV windows was also higher, most likely because participants felt they needed to scan these windows more frequently to make sure the UAVs did not miss any targets.

The above results indicate that it is useful to consider and combine multiple eye tracking metrics when trying to infer trust as various aspects (temporal, spatial and count) of eye movements respond differently to variations in system reliability.

Effects of priming on eye movements and monitoring

Even though priming did not affect subjective trust ratings, it was associated with changes in participants' monitoring behavior. Specifically, priming seemed to contribute to smaller mean saccade amplitudes, transition rates, scanpath length per second and total fixation counts, as well as longer average fixation durations.

Average fixation duration was the only temporal metric affected by priming. Receiving reliability information in advance likely resulted in participants allocating more effort and attention to the low reliability UAVs, in a top-down fashion. In contrast, total fixation duration did not differ between the priming and the no priming group. One explanation could be that participants' attention allocation across the entire screen was decided mainly by the total task duration instead of other factors. Priming significantly affected all spatial metrics, except the backtrack rate. This confirms that participants' visual search and scanning were indeed less efficient and systematic if participants were not informed about system reliability. It is not clear why the backtrack rate was not affected by priming when it showed the expected change due to differences in UAV reliability.

Among the count metrics, significantly fewer fixation counts, but not fewer transition counts, were observed in the priming condition. It is possible that any difference in transition counts was masked as it was calculated for the entire screen (including both high and low reliability UAVs).

Relationship between subjective trust ratings and eye tracking metrics

For the most part, the observed differences in monitoring behavior for the high- and low-reliability UAVs aligned with participants' subjective trust ratings. The only exceptions were average fixation duration and mean saccade amplitude. These metrics were not affected by system reliability but did change as a function of priming, while subjective trust ratings were affected only by system reliability but not priming. This explains why there was no correlation between subjective trust ratings and the two eye tracking metrics. The result also suggests that, overall, system reliability has a more pronounced effect on trust, compared to priming. This finding is similar to previous research finding (Liao & MacDonald, 2019) showing that affective primes did not affect trust propensity while product performance was critical in shaping trust.

Another important difference between the eye tracking metrics and subjective trust ratings was that subjective ratings differed significantly between high and low reliability UAVs, independent of whether participants were informed about system reliability at the beginning of the experiment. In contrast, one eye tracking metric, total fixation duration, differed between the two reliability levels only in the priming condition. Participants in the no priming group explained during the debrief that, even though they had noticed differences in automation reliability during the experiment, they still monitored all UAVs to the same extent because they were not sure whether observed reliability levels would remain constant.

A third difference between subjective ratings and the eye tracking metrics was the higher temporal resolution of total fixation counts which closely mirrored the changes in UAV reliability every two minutes. In other words, monitoring behavior changed even though participants' attitude towards the vehicle appeared unaffected. This finding is similar to the results from an earlier experiment on autonomous driving (Miller et al., 2016), where drivers intervened even when they expressed verbally that they expected the automation to be able to handle the emergency situation successfully.

Effects of system reliability and priming on performance

In terms of performance, participants' overall error rate was quite low. Their reaction time was longer, and their overall error rate was higher for low reliability UAVs. The high rate of type 3 errors (failure to notice a missed target when the window was not highlighted) for low reliability UAVs was likely due to the lack of attention guidance in the form of highlighting of the video feed. The high type 4 error rate (false alarms when the window was not highlighted) for low reliability UAVs may be due to participants' response bias. They expected these vehicles to miss more targets and were therefore more willing to call an ambiguous object a target. And the high rate of type 5 errors (failure to respond to highlighting of window) for highly reliable UAVs may be the result of participants focusing so much on the low reliability vehicles that they did not react in time or totally missed in their peripheral vision the highlighting of the high reliability UAV windows. Priming did not affect participants' performance significantly which may be explained by a ceiling effect (above 90% accuracy rate).

In conclusion, the findings from this experiment suggest that eye tracking is indeed an effective technique for inferring changes in operator trust levels in real time. Compared to other psychophysiological measures, eye tracking has the benefits of easier implementation, less intrusion and a more fine-grained analysis of monitoring behavior. Given the dissociation between some trust-related measures in this study, eye tracking is ideally combined with other techniques, such as subjective ratings or behavioral data, to assess and study various facets of trust.

It is important to note two limitations of the present study. First, false alarms and misses were evenly distributed among UAVs of equal reliability and over each short time interval. Past studies have shown that these two failure types can have different effects on one's trust in automation (C. Wickens et al., 2005). Automation that mostly triggers false alarms may more strongly affect eye tracking metrics that relate to information processing; in contrast, automation that is prone to misses may lead to changes in eye tracking metrics that reflect search behavior. Second, in this experiment, high and low automation reliability were coupled, which means that, if a participant looked more at one group of UAVs, his/her monitoring of the other group was necessarily reduced, independent of whether or not his/her trust in those vehicles was different. The experiments reported in the following chapters addressed the above shortcomings.

References

Aasman, J., Mulder, G., & Mulder, L. J. (1987). Operator effort and the measurement of heart-rate variability. *Human factors*, *29*(2), 161-170.

Bagheri, N., & Jamieson, G. A. (2004). Considering subjective trust and monitoring behavior in assessing automation-induced "complacency.". *Human performance, situation awareness, and automation: Current research and trends*, 54-59.

Beck, M. R., Lohrenz, M. C., & Trafton, J. G. (2010). Measuring search efficiency in complex visual search tasks: Global and local clutter. *Journal of experimental psychology: applied*, *16*(3), 238.

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied Ergonomics*, *52*, 333-342.

Chen, S., Epps, J., Ruiz, N., & Chen, F. (2011). *Eye activity as a measure of human mental effort in HCI.* Paper presented at the Proceedings of the 16th international conference on Intelligent user interfaces.

Chugh, J. S., & Caird, J. K. (1999, September). In-vehicle train warnings (ITW): The effect of reliability and failure type on driver perception response time and trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 43, No. 18, pp. 1012-1016). Sage CA: Los Angeles, CA: SAGE Publications.

Goldberg, J. H., & Kotval, X. P. (1999). Computer interface evaluation using eye movements: methods and constructs. *International Journal of Industrial Ergonomics, 24*(6), 631-645.

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human factors, 58*(3), 509-519.

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye* (pp. 573-605): Elsevier.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Englewood Cliffs, NJ: Prentice-Hall.

Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S. W.-Y., . . . Tsai, C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational research review, 10*, 90-115.

Liao, T., & MacDonald, E. *Manipulating Trust of Autonomous Products With Affective Priming.* Paper presented at the ASME 2019 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). *Behavioral Measurement of Trust in Automation: The Trust Fall.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Moacdieh, N. M., & Sarter, N. (2017). The effects of data density, display organization, and stress on search performance: An eye tracking study of clutter. *IEEE Transactions on Human-Machine Systems*, *47*(6), 886-895.

Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science, 1*(4), 354-365.

Petersen, L., Robert, L., Yang, J., & Tilbury, D. (2019). Situational awareness, driver's trust in automated driving systems and secondary task performance. *SAE International Journal of Connected and Autonomous Vehicles, Forthcoming*.

Strauch, C., Mühl, K., Patro, K., Grabmaier, C., Reithinger, S., Baumann, M., & Huckauf, A. (2019). Real autonomous driving from a passenger's perspective: Two experimental investigations using gaze behaviour and trust ratings in field and simulator. *Transportation research part F: traffic psychology and behaviour, 66*, 15-28.

Vicente, K. J., Thornton, D. C., & Moray, N. (1987). Spectral analysis of sinus arrhythmia: A measure of mental effort. *Human factors*, *29*(2), 171-182.

Walker, F., Verwey, W., & Martens, M. (2018). *Gaze behaviour as a measure of trust in automated vehicles.* Paper presented at the Proceedings of the the 6th Humanist Conference (June 2018).

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201-212.

Wickens, C., Dixon, S., Goh, J., & Hammer, B. (2005). *Pilot dependence on imperfect diagnostic automation in simulated UAV flights: An attentional visual scanning analysis*. ILLINOIS UNIV AT URBANA SAVOY

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*(4), 352-367.

Chapter 3

Effects of Magnitude and Duration of Changes in System Reliability on Trust Calibration and Recovery

As mentioned in the literature review (Chapter 1), and as shown in Experiment 1, system reliability is one of the most critical factors shaping operators' trust in automated systems (Lee & Moray, 1994; Lee & See, 2004; Schaefer et al., 2016). While the effects of system reliability on trust have been studied quite extensively, research in this area involves limitations related to the type of changes in system reliability and the aspects of trust that are being examined. Most experiments vary system reliability between but not within trials (e.g., Chavaillaz et al., 2016; Merritt et al., 2013; Yu et al., 2017). Also, trust is usually measured only one time - at the end of the experiment (e.g., Dzindolet et al., 2001; Rossi, Dautenhahn, Koay, & Walters, 2017). This means that very few studies provide any insight into how changes in system reliability affect trust calibration and evolution over time. One such study was conducted by Desai and colleagues (Desai et al., 2013; Desai et al., 2012) who investigated how the timing of a breakdown in system performance influenced trust development. Trust changes were captured in real time by asking participants for trust ratings every 20 seconds. However, data from only one scenario, where system reliability dropped halfway through the trial, could be compared with the baseline scenario where the automation performed highly reliably throughout; other scenarios were eliminated from the analysis because they did not last long enough to allow for trust development or trust recovery. Also, the study did not examine the effects of changes in system reliability on trust calibration. The experiment presented in this chapter will fill these gaps.

In terms of the nature of system reliability changes, past studies have focused primarily on the timing (i.e., at what point during a scenario reliability increased or decreased) (e.g., Rossi et al., 2017) and the type of breakdowns in system performance (i.e., false alarm and miss)(e.g., Guznov, Lyons, Nelson, & Woolley, 2016). Very few studies have examined how the severity of a drop in system reliability (Lewis, Sycara, & Walker, 2018), in particular its magnitude and duration, affect operators' trust and reliance on automation – the focus of the present experiment.

The main goals of this study then are: 1) to establish the detectability of system reliability changes of different magnitudes and durations, 2) to assess how trust calibration and recovery vary as a function of the magnitude and duration of variations in system reliability, and 3) to examine whether the magnitude and duration of system reliability changes interact with respect to how they affect trust calibration and recovery. To this end, participants performed a tracking task and a target detection task similar to the one used in the first experiment, with the aid of automation on board eight UAVs. The dependent measures were subjective trust ratings, perceived automation reliability, eye movements and performance on both tasks. The main expectations were that (1) participants would likely miss small and short drops but would reliably notice large or long decreases in system reliability, (2) perceived reliability of the automation would decline when a drop in system reliability was observed and trust ratings would change accordingly, (3) in case of large reliability drops, participants' trust in the automation would change immediately and significantly, (4) during long periods of low system reliability, participants' perceived reliability and trust would continue to decrease and approach actual system reliability, (5) participants' trust calibration during the first three minutes after system reliability returned to its original high level would suffer the most (as reflected by a large perceived reliability difference and inappropriate subjective trust levels) following a combined large and long drop in reliability, (6) trust recovery would take longer after a large or a long reduction in system reliability, and (7) the most severe trust miscalibration would lead to the worst task performance due to inappropriate reliance on the automation.

Method

Participants

A total of 20 students from the University of Michigan participated in the experiment. Their average age was 22.2 years (SD=3.5). 12 participants were males. Participants were required to have normal or corrected-to-normal vision. None of the participants had any experience with Unmanned Aerial Vehicle (UAV) control or the MATB tracking task before the experiment. This study was approved by the UM Health Sciences and Behavioral Sciences Institutional Review Board (HSBS-IRB; ID: HUM00143537).

Apparatus and tasks

The experiment was set in a computer-based Unmanned Aerial Vehicle (UAV) control simulation built in the THInC lab. Participants were asked to complete a target detection task and a tracking task at the same time (see Figure 3.1). The two tasks were displayed on a 27'' monitor.

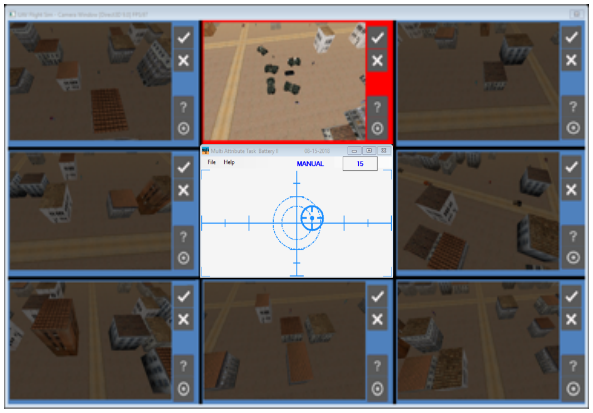


Figure 3.1. UAV simulation screenshot

(tracking task in the center of the screen)

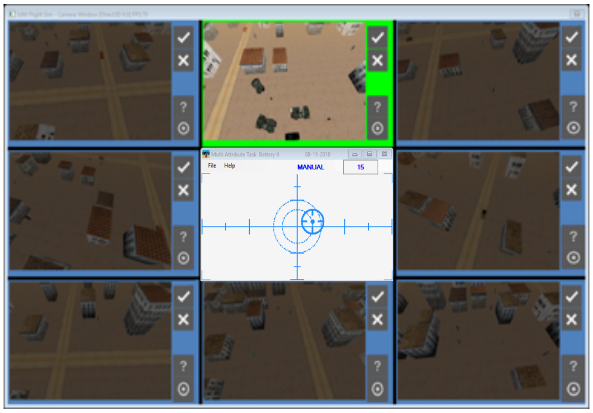


Figure 3.2. Simulation screenshot when

reaching a non-target area

For the target detection task, participants were asked to monitor the video feeds from 8 unmanned aerial vehicles (UAVs) to detect the presence of military targets. They were informed that automation onboard the UAVs would scan pre-defined regions to help with the detection of a target. Whenever a UAV reached one of the target regions, its video feed was highlighted. The surround of the video feed changed to red when it identified a possible target (as shown in Figure 3.1). The surround appeared green when the automation reached a region but did not detect a target (as shown in Figure 3.2). The target was defined as a black truck carrying a gun (See Figure 3.3). The video feeds also included a similar-looking object, a black truck without a gun, as a distractor, as shown in Figure 3.4.

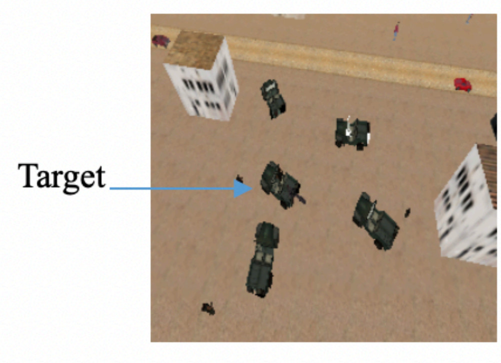


Figure 3.3. Target example

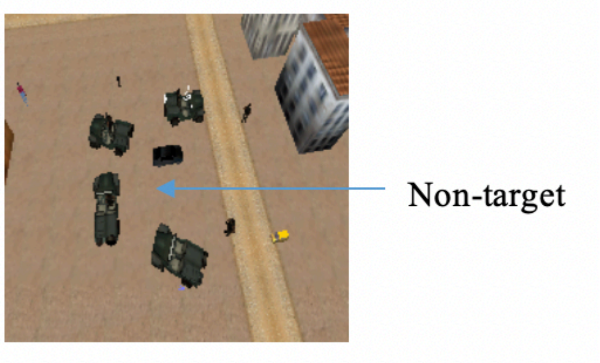


Figure 3.4. Non-Target example

It was up to the participant to decide whether or not to review the scene once a video feed was highlighted. If they did, they could then press one of two buttons (√ or ×) to either agree or disagree with the automation assessment regarding the presence of a target. If the participant did not review the scene, the automation would make a decision for the participant. Independent of whether the participant reviewed the scene and responded to the automation assessment, they would receive auditory feedback regarding the accuracy of the automation (Lu & Sarter, 2019b). If the automation was correct, a 'pleasant' high pitch chime sounded; when the automation assessment was wrong, a low pitch buzzing sound was played. Participants were presented with both sounds during the training session to make sure they were reliably distinguished.

In parallel with the target detection task, participants had to perform a tracking task (adopted from MATB-II (NASA)) in the center of the screen (see Figure 3.1). Using a joystick, they had to keep the smaller target circle inside the two larger circles around the origin of the coordinate frame. They were told that, if the target circle moved outside the larger circles, this could result in the UAV crashing. Participants were instructed to treat the target detection task and the tracking task as equally important and to try to achieve the best possible performance on both tasks.

Experiment design

The experiment employed a 2 (magnitude of reliability change: large vs. small) *2 (duration of reliability change: long vs. short) full factorial design. Both the magnitude and the duration of variations in system reliability were within-subject factors, leading to a total of four reliability scenarios, as shown in Figures 3.5-3.8. In Scenario 1, system reliability dropped from 95% to 50% and stayed at 50% for 9 minutes before recovering to 95%. Scenario 2 was the same as Scenario 1, except that system reliability dropped to 80% (instead of 50%). In Scenario 3, system reliability changed from 95% to 50% (like in scenario 1) but returned to 95% after only 3 minutes. Finally, Scenario 4 was different from Scenario 3 in that the reliability dropped to 80%, as opposed to 50%. The sequence in which each participant experienced the four scenarios was counterbalanced between subjects. Half of the participants experienced the four scenarios in the order of Scenario 1-2-3-4. The other half of the participants experienced the four scenarios in reverse order. In the data analysis, this "sequence" factor was entered as an independent variable to check if the manipulation was successful.

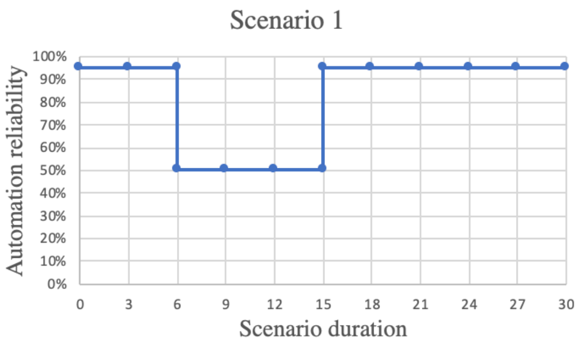


Figure 3.5. Large & long drop

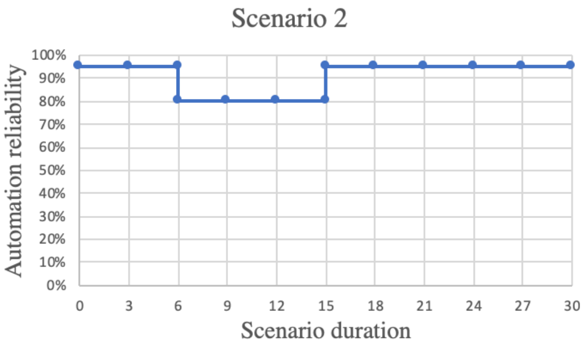


Figure 3.6. Small & long drop

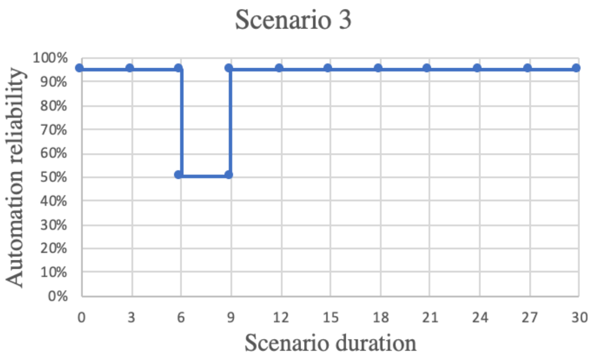


Figure 3.7. Large & short drop

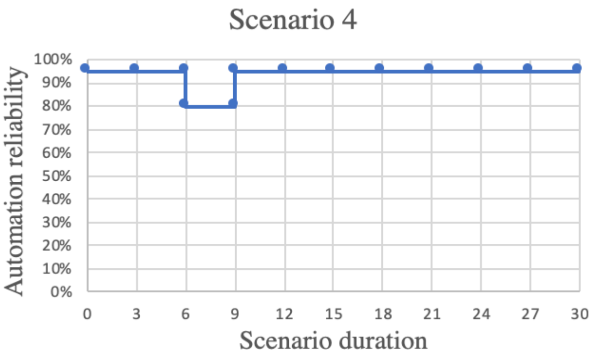


Figure 3.8. Small & short drop

The dependent measures in this study were subjective trust ratings, perceived automation reliability (as a measure of trust miscalibration (Merritt et al., 2015)), eye movements, as well as the frequency to review automation assessments and performance on the target identification and tracking tasks. Trust ratings and perceived reliability estimates were collected every three minutes, throughout each trial, by asking participants to answer two questions verbally: (1) “How much do you trust the automation?” (on a scale from 0-10, with 0 being the lowest possible trust) and (2) “What is your perceived reliability of the system” (on a scale of 0-100%). The 3-minute time interval was selected because it was shown in pilot testing to be long enough to trace trust changes and not as disruptive as the shorter time intervals used in earlier research. The eye tracking metrics used in this study were total fixation duration percentage, fixation count percentage and transition count. Total fixation duration percentage is the percentage of fixation time that a participant focuses on the display area showing the tracking task (see AOI2 in Figure 3.9). Fixation count percentage is the percentage of fixation count on the tracking task. Percentages, rather than absolute values, were used in this study to account for large individual differences in eye tracking metrics. Transition count is number of transitions between AOI 1 (Area of Interest 1: the target detection area) and AOI 2 (the tracking task area). These eye tracking metrics have been confirmed in Experiment 1 as indicators of different trust levels in response to system reliability variance. Metrics in the spatial dimension (related to different search features) were not used in this study because in Experiment 2, participants did not need to actively search for missing targets. The performance data consisted of participants' tracking task performance and target detection performance. Tracking task performance refers to the Mean Square Distance (MSD) of the small circle target from the center (as shown in Figure 3.1), calculated over five-second windows. Target detection task performance includes overall error rate for target identification, and response time which was defined as the time between highlighting of video feed and button press.

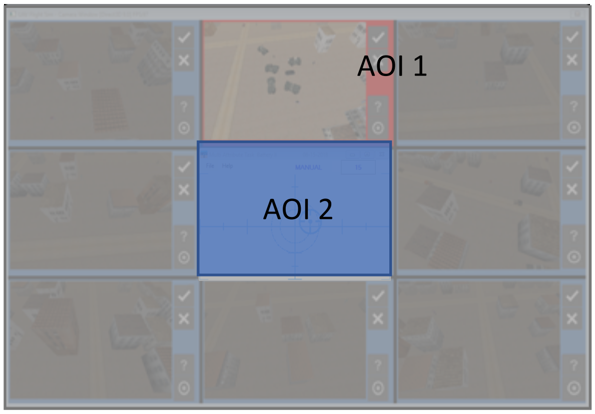


Figure 3.9. Area of Interest (AOI) definition

Experiment procedure

Upon arrival at the laboratory, participants first read and signed the consent form. They were then asked to fill out a background questionnaire asking for basic information, such as their age, gender, and nationality. Next, participants were instructed on the two tasks and completed a 12-minute training session before the actual experiment started. Each participant completed the 4 automation reliability scenarios described above, with a 5-minute break in between. Each scenario ended once a steady state of trust in the automation was reached (i.e., when the participant provided 2 similar subjective trust ratings that were close to their initial trust ratings). After completing the four scenarios, baseline eye tracking behavior was collected for three 6-minute periods that differed with respect to system reliability (95%, 80% or 50%). Participants were informed about the respective system reliability in advance of each trial during which they performed the same tasks as in the previous four scenarios. These baseline data were collected for later modeling purposes that will be described in Chapter 4. This baseline data collection was conducted after the four formal trials to avoid priming effects on participants. Finally, participants were asked to fill out a debriefing questionnaire. It took participants around 2 hours to complete the entire experiment, and they were compensated $30 for their participation.

Results

The Results section consists of three main parts. The first part reports on data analyses related to the detection of system reliability drops of varying duration and magnitude (expectation 1). Specifically, polynomial contrast analyses were performed to assess how well actual changes in system reliability were reflected in participants' perceived reliability, subjective trust ratings and eye movement data. The second part of the Results section describes findings related to expectations 2-6 on page 55. Linear mixed models were applied to establish the effects of magnitude and duration of system reliability changes on participants' perceived reliability, subjective trust ratings, eye tracking metrics, their review of automation assessments and various performance outcome measures (see expectation 7). The analyses examined the following four time periods (illustrated in Figure 3.10): 1) the first three minutes after system reliability dropped (the red rectangle), 2) the nine-minute period of low system reliability (the yellow rectangle), 3) the three-minute period after system reliability returned to its original level (the green rectangle), and 4) the time until participants' trust recovered after experiencing the system performance breakdown (the blue rectangle). For statistical analysis, the significance was set at $p < 0.05$. Error bars on the figures indicate the standard error of the mean.

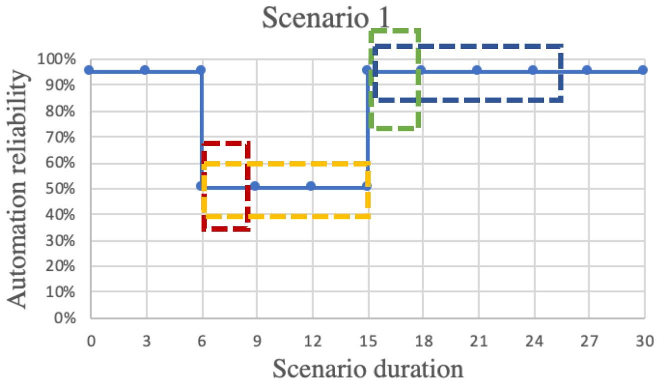


Figure 3.10. Four time periods for which data were analyzed

Part 1: Detection of system reliability changes of different magnitude and duration

Polynomial contrast analyses show that participants' perceived reliability and trust ratings reflect actual changes in system reliability in all four scenarios (see Table 3.1). A polynomial contrast analysis was conducted also on the eye tracking metrics to determine if participants adjusted their attention allocation in response to changes in system reliability. In Scenarios 1-3, all three eye tracking metrics changed when system reliability dropped. However, in Scenario 4 (the small and short reliability change), only total fixation duration showed a significant quadratic trend. The results from the above tests are summarized in Table 3.1.

Table 3.1. Polynomial contrast analyses for perceived reliability ratings, subjective trust ratings and eye tracking metrics

Scenario	Perceived reliability	Subjective trust ratings	Eye tracking		
			Total fixation duration PC	Fixation count PC	Transition count
1-large and long reliability drop	t(1,95)=11.175, $p < 0.001$	t(1,95)=8.688, $p < 0.001$	t(1,85) = 7.126, $p < 0.001$	t(1,85) = 5.331, $p < 0.001$	t(1,85) = -6.321, $p < 0.001$
2-small and long reliability drop	t(1,92)=8.022, $p < 0.001$	t(1,92)=7.002, $p < 0.001$	t(1,87) = 3.059, $p = 0.003$	t(1,87) = 3.677, $p < 0.001$	t(1,87) = -3.748, $p < 0.001$
3-large and short reliability drop	t(1,38)=8.106, $p < 0.001$	t(1,38)=5.539, $p < 0.001$	t(1,34) = 3.233, $p = 0.003$	t(1,34) = 3.807, $p = 0.001$	t(1,34) = -2.739, $p = 0.01$
4-small and short reliability drop	t(1,38)=3.117, $p = 0.003$	t(1,38)=2.392, $p = 0.022$	t(1,38) = 2.306, $p = 0.027$	t(1,38) = 0.019, $p = 0.985$	t(1,38) = -1.243, $p = 0.222$

*Significant findings ($p < 0.05$) are highlighted in grey

Part 2: The effects of magnitude and duration of system reliability changes on perceived reliability, subjective trust ratings, eye movements, review of automation assessments and performance

Period 1: the first three minutes after a drop in system reliability

This period was examined to determine whether abrupt changes in system reliability resulted in a temporary miscalibration of trust, as reflected by the difference between the actual system reliability and participants' perceived reliability and by subjective trust ratings, eye movement data, participant behavior and performance outcomes. Note that this period has a fixed duration, and this factor was therefore not included in the analyses.

Perceived reliability calibration

To calculate perceived reliability calibration, both perceived reliability and actual reliability were first re-scaled to a range of ‘0-10’ and then compared. A 2(magnitude)*2(sequence) linear mixed model analysis was performed on the difference between actual and perceived reliability. It reveals a significant main effect of magnitude (F(1,58) = 10.720, *p* = 0.002), as shown in Figure 3.11. Large reliability drops (Mean = -1.063, SD = 1.542) resulted in a significantly larger difference between actual and perceived reliability, compared to small reliability drops (Mean = -0.243, SD = 0.905).

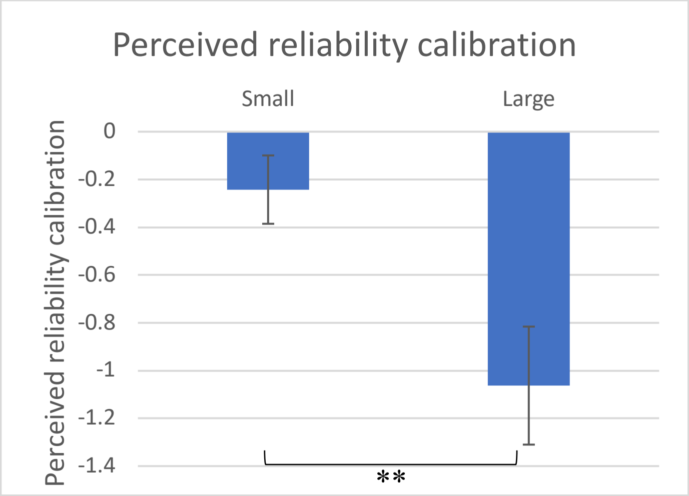


Figure 3.11. Effects of magnitude of system reliability changes on perceived reliability calibration

Subjective trust ratings

A 2 (magnitude)*2(sequence) linear mixed model analysis was conducted on the subjective trust ratings. Participant ID was entered as a random effect, and the subjective trust ratings immediately preceding the system reliability change were entered as a covariant. The result indicates that participants’ trust ratings were significantly lower (Mean = 6.308, SD =

2.173, F (1,58) = 42.281, $p < 0.001$) when they experienced a large reliability drop compared to their ratings (Mean = 8.189, SD = 1.214) following a small reliability drop (see Figure 3.12).

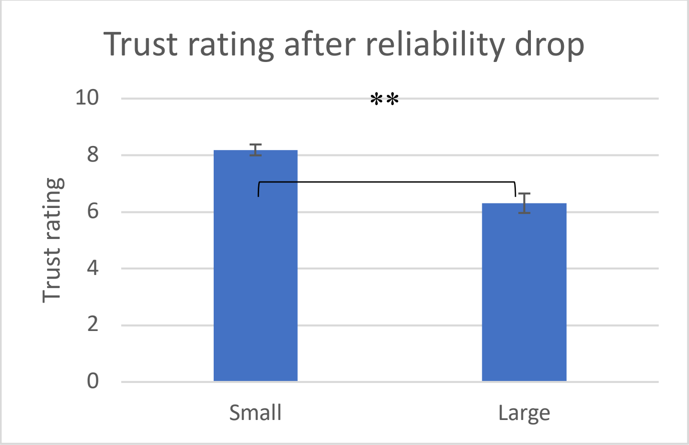


Figure 3.12. Effects of magnitude of system reliability changes on trust ratings

Eye tracking metrics

Three 2(magnitude)*2(sequence) linear mixed model analyses were conducted separately on the eye tracking metrics: total fixation duration percentage, total fixation count percentage, and transition count. Participant ID was entered as a random effect. There was a significant effect of magnitude on all three metrics (see Table 3.2). The average percentage of total fixation duration on AOI 2 (the tracking task area) was significantly lower (85.9%) following a large reduction in system reliability, compared to a small reduction in reliability (91.6%). Similarly, the fixation count percentage on AOI 2 significantly decreased to 68.7% following a large drop in reliability, as compared to 78.5% following a smaller drop. And the transition counts between the two AOIs following a large versus as small change in system reliability were 37.78 and 25.68, respectively. There was no significant effect of sequence and no interaction effect between magnitude and sequence.

Table 3.2. Changes in three eye tracking metrics as a function of magnitude of reliability change

Eye movement metrics	Small magnitude mean	Large magnitude mean	Main effects of magnitude
Total fixation duration percentage	0.916 (0.068)	0.859 (0.066)	F(1,54.14) = 20.99, *p* <0.001
Total fixation count percentage	0.785 (0.142)	0.687 (0.118)	F(1,53.48) = 17.389, *p* <0.001
Transition count	25.68 (17.82)	37.78 (16.41)	F(1,54.029) = 14.269, *p* <0.001

Review of automation assessments

Finally, a 2(magnitude)*2(sequence) linear mixed model was applied to the number of times participants reviewed the automation assessments in the three-minute period following the system reliability drop. Participant ID was entered as a random effect, and the number of times immediately preceding the system reliability drop was entered as a covariant. A large drop in system reliability resulted in a significantly larger number of times (Mean = 11.58, SD = 6.19), compared to a small drop in system reliability (Mean = 6.78, SD = 6.53, F(1,58) = 32.884, *p* <0.001), as shown in Figure 3.13. There was no significant effect of sequence and no interaction effect between magnitude and sequence.

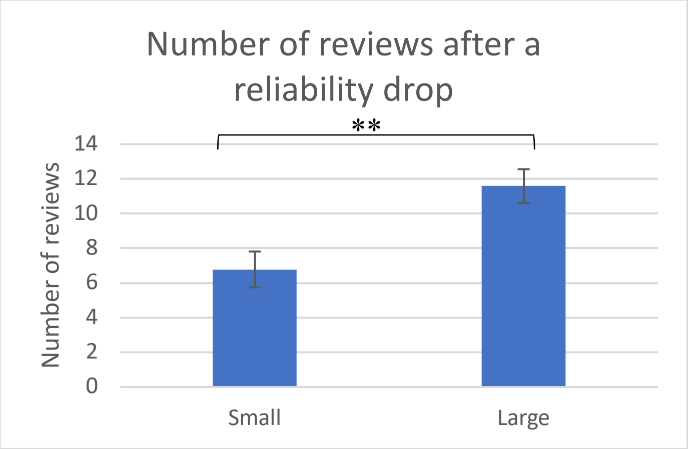


Figure 3.13. Effects of magnitude on the number of reviews after a reliability drop

Tracking task performance

A 2(magnitude)*2(sequence) linear mixed model, with MSD before the reliability change entered as a covariant, did not show any main or interaction effects for MSD following the drop in system reliability.

Target detection performance

Overall error rate

The overall error rate on the target detection task during the three-minute period after system reliability suddenly dropped was examined using a 2(magnitude)*2(sequence) linear mixed model. As shown in Figure 3.14, a significant effect of magnitude on overall error rate was observed (F(1,58) = 59.25, *p* < 0.001). Participants' error rate was significantly smaller (13.88%; SD = 6.25%) when reliability dropped by only 15%, compared to 29.25% (SD=15.09%) when reliability dropped by 45%. No significant effect of sequence and no interaction effect were observed.

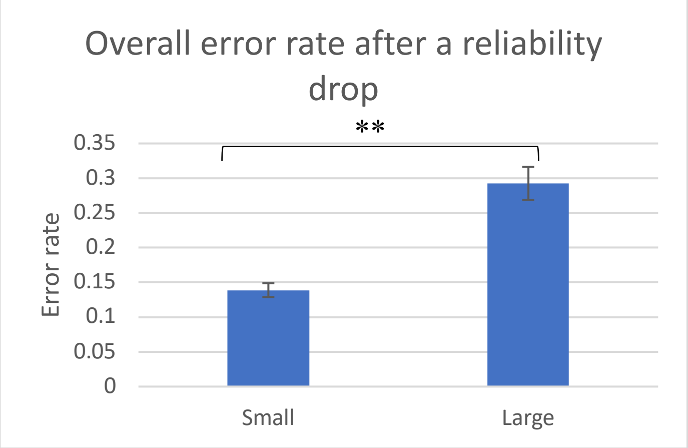


Figure 3.14. Effects of magnitude on overall error rate after a reliability drop

Response time

The analysis of response times for button pushes did not yield any significant main or interaction effects.

Period 2: the nine-minute period of low system reliability

The nine-minute period of low system reliability was examined to determine whether trust calibration would improve over time, i.e., whether participants' perceived reliability estimation would get closer to actual system reliability, and to assess how subjective trust ratings, attention allocation and performance would change as a function of magnitude and duration (3 three-minute intervals) of the reliability change.

Perceived reliability calibration

A 2 (magnitude) *3 (time interval) * 2(sequence) linear mixed model analysis was conducted on perceived reliability calibration. Participant ID was entered as a random effect. Similar to the results in period 1, there was a significant effect of magnitude on perceived reliability (F(1,90) = 19.074, $p < 0.001$). A large reliability drop (Mean = -0.813, SD = 1.674) resulted in a significantly larger difference between actual and perceived reliability, compared to a small reliability drop (Mean = -0.042, SD = 0.824), as shown in Figure 3.15. No significant effects of time interval and sequence were observed.

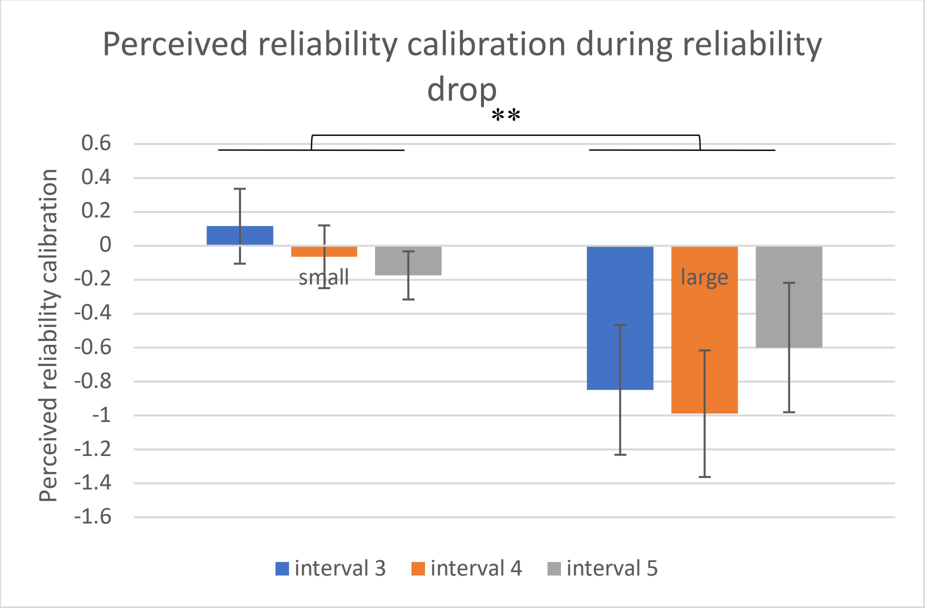


Figure 3.15. Effects of magnitude on perceived reliability calibration during the 9-minute low reliability period

Subjective trust ratings

A 2 (magnitude) *3 (time interval) * 2(sequence) linear mixed model analysis was conducted also on subjective trust ratings. Consistent with the findings on perceived reliability differences, only magnitude was found to significantly affect participants' trust ratings for the 9-minute period. Participants' trust ratings were significantly lower (Mean = 5.905, SD = 2.352, F(1,90) = 93.071, *p* <0.001) but did not change throughout the large reliability drop, compared to their ratings (Mean = 7.838, SD = 1.427) during the small drop in reliability, as shown in Figure 3.16.

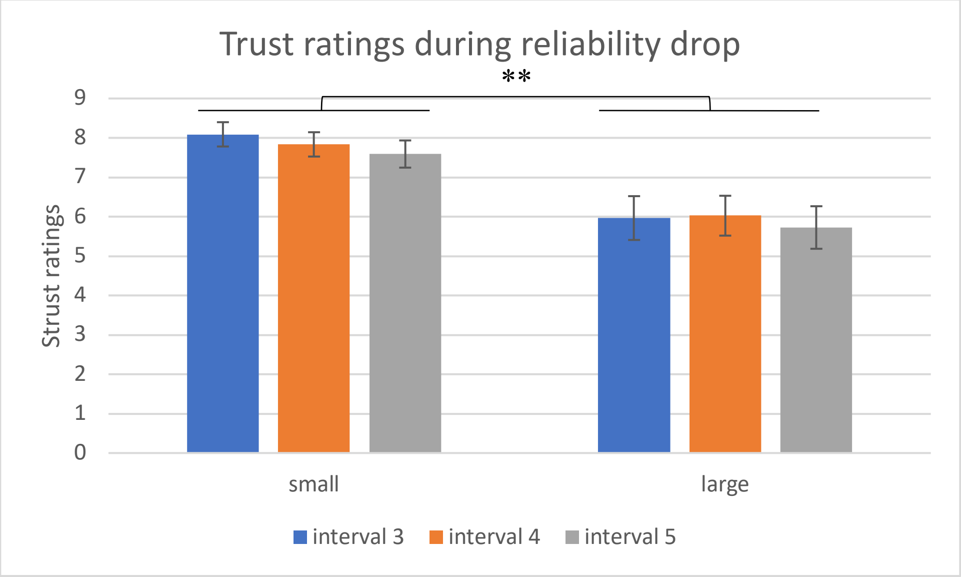


Figure 3.16. Effects of magnitude on subjective trust ratings during the 9-minute low reliability period

Eye tracking metrics

A 2 (magnitude) *3 (time interval) * 2(sequence) linear mixed model analysis on the eye tracking data revealed a significant effect of magnitude for all three eye tracking metrics (see Table 3.3). The average percentage of total fixation duration on AOI 2 (the tracking task area) was significantly lower (84.5%) during a large reduction in system reliability, compared to a small reduction in reliability (90.2%). Similarly, the fixation count percentage on AOI 2 decreased to 66.1% during a large reduction in reliability, as compared to 75.2% during a smaller reduction. And the transition counts between the two AOIs during a large versus as small magnitude change in reliability were 45.81 and 31.09, respectively.

The analysis also shows a significant effect of time interval on two of the eye tracking metrics. A post-hoc analysis with Bonferroni corrections indicates that the average percentage of total fixation durations on AOI 2 (the tracking area) during the first three minutes after the reliability drop (interval 3) was significantly higher (89.5%) than during interval 4 (86.5%) and interval 5 (85.9%). For transition counts between the two AOIs, the post-hoc analysis revealed a significantly smaller transition count during interval 3 (32.97), compared to interval 5 (42.19).

Table 3.3. Changes in three eye tracking metrics as a function of magnitude of reliability change and time intervals

Eye movement metrics	Magnitude effect		Interval effect		
	Small	Large	Interval 3	Interval 4	Interval 5
Total fixation duration percentage	0.902 (0.074)	0.842 (0.077)	0.895 (0.067)	0.865 (0.083)	0.859 (0.088)
	F(1,90.9) = 30.661, $p < 0.001$		F(1,88.396) = 5.952, p = 0.004 Post-hoc: Interval 3 significantly larger than interval 4 (p = 0.028) and 5 (p = 0.005)		
Total fixation count percentage	0.752 (0.145)	0.661 (0.095)	0.734 (0.135)	0.697 (0.124)	0.693 (0.134)
	F(1,91.75) = 20.592, $p < 0.001$		Not significant		
Transition count	31.09 (17.73)	45.81 (23.17)	32.97 (17.79)	39.59 (21.53)	42.19 (24.90)
	F(1,91.45) = 20.46, $p < 0.001$		F(1,88.4) = 3.869, p = 0.025 Post-hoc: Interval 3 significantly smaller than interval 5 (p = 0.025)		

Review of automation assessments

A 2 (magnitude)*3 (time interval)* 2(sequence) linear mixed model analysis on the number of times participants reviewed the automation assessments over the nine-minute period shows that a large system reliability drop led to significantly (F(1,90) = 71.634, $p < 0.001$) more reviews (Mean = 13.22, SD = 6.00) than a small drop in reliability (Mean = 8.23, SD = 6.63). The number of reviews also changed across the three 3-minute time intervals (F(1,90)=8.271, p = 0.001). A post-hoc analysis with Bonferroni corrections shows that the number of reviews in interval 3 (Mean = 9.1, SD = 6.95) was significantly smaller than in intervals 4 (Mean = 11.13, SD = 6.25, *p* = 0.044) and 5 (Mean = 11.95, SD = 6.95, *p* = 0.002; see Figure 3.17).

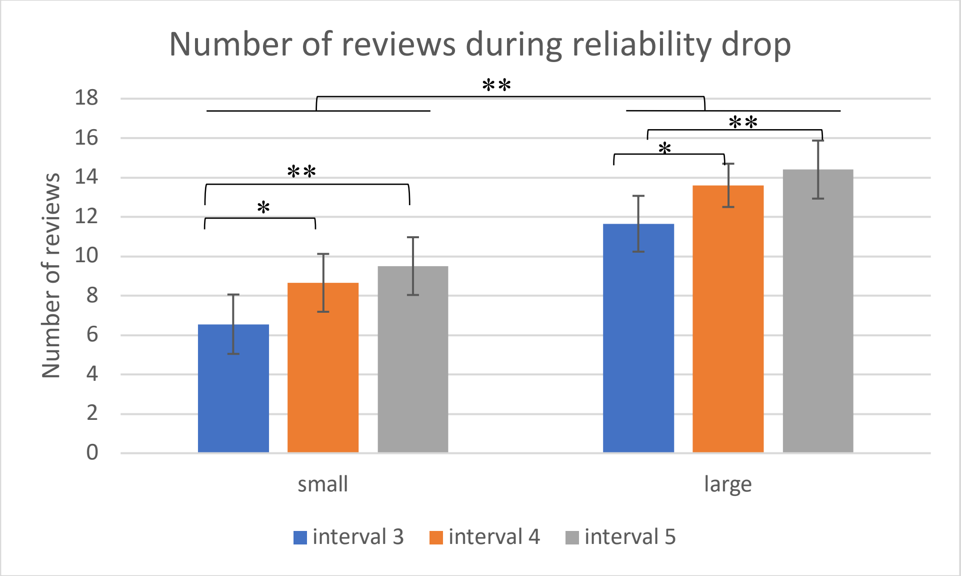


Figure 3.17. Effects of magnitude and time interval on number of reviews during the 9-minute low reliability period

Tracking task performance

A 2(magnitude) *3(time interval) * 2(sequence) linear mixed model, with MSD before the reliability change entered as a covariant, showed a significant effect of magnitude on MSD during the 9-minute period of low system reliability (F(1,90)=6.008, *p* =0.016). Participants who experienced a large reliability drop (from 95% to 50%) produced a significantly larger MSD (Mean = 42.058, SD = 5.467) compared to participants who experienced a small reliability drop (from 95% to 80%) (Mean = 40.727, SD = 5.441) (See Figure 3.18).

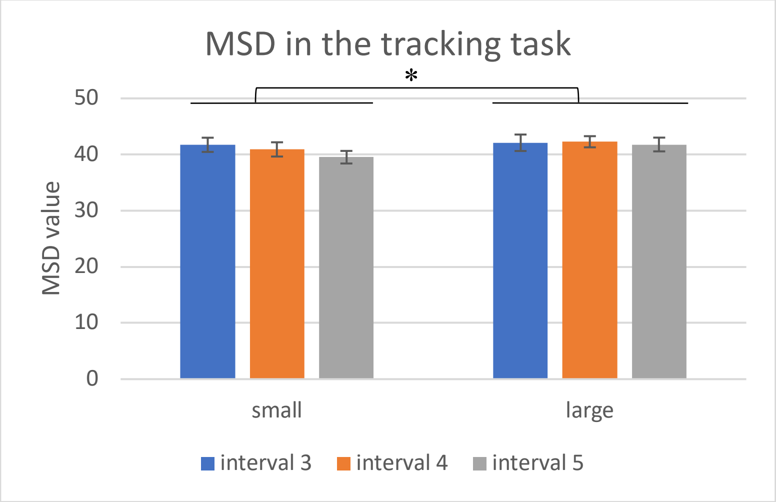


Figure 3.18. Effects of magnitude on tracking performance during the 9-minute reliability drop

Target detection performance

Overall error rate

A 2(magnitude) *3(time interval) *2(sequence) linear mixed model was performed for the 9-minute time period after system reliability dropped. The results revealed a significant effect of magnitude (F(1,90) = 24.015, $p < 0.001$). Participants' overall error rate was significantly higher following the large reliability drop (Mean = 0.235, SD = 0.160) compared to the overall error rate following the small reliability drop (Mean = 0.154, SD = 0.074), as shown in Figure 3.19. There was also a significant effect of time interval on the overall error rate. A post-hoc analysis with Bonferroni correction indicated that the error rate 4-6 minutes (Interval 4) (Mean = 0.186, SD = 0.122) and 7-9 minutes (Interval 5) after the reliability drop (Mean = 0.163, SD = 0.135) was significantly lower than during the first three minutes (Interval 3) after the decline in reliability (Mean = 0.235, SD = 0.127). However, a significant interaction effect between time interval and magnitude indicates that this was true only with a large reliability drop.

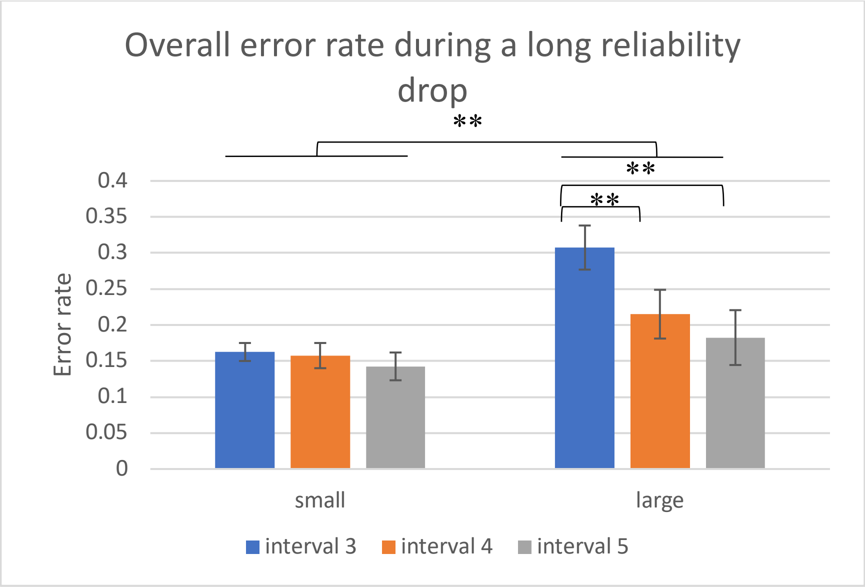


Figure 3.19. Effects of magnitude and time intervals on overall error rate during a long reliability drop

Response time

The analysis of response times for button pushes did not yield any significant main or interaction effects.

Period 3: the three-minute period after system reliability recovered to its initial high level

Perceived reliability calibration

A 2(magnitude)*2(duration)*2(sequence) linear mixed model analysis was performed for the three-minute period following system reliability recovery. There was a main effect of magnitude (F(1,54) = 12.539, $p = 0.001$) on the difference between actual and perceived reliability, as shown in Figure 3.20. A large reliability drop resulted in a significantly larger miscalibration in perceived reliability (Mean = 1.29, SD = 1.18), compared to a small reliability drop (Mean = 0.668, SD = 0.848). No significant effect of duration and no interaction effect were observed.

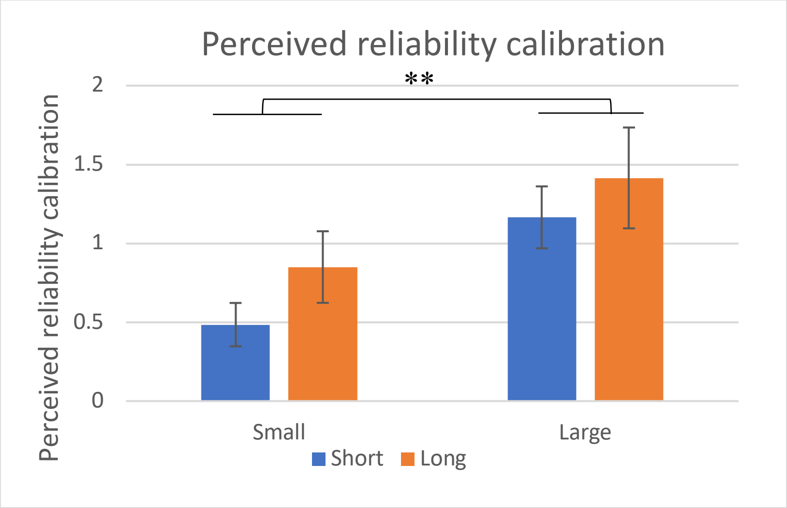


Figure 3.20. Effects of magnitude and duration on perceived reliability calibration

Subjective trust ratings

A 2 (magnitude) *2 (duration) * 2(sequence) linear mixed model analysis was conducted also on subjective trust ratings right after system reliability recovered to its initial high level. Participant ID was entered as a random effect, and the subjective trust ratings right before the change in system reliability were entered as a covariant. The result indicates that both magnitude and duration of the change significantly affected trust ratings (see Figure 3.21). Participants' subjective trust ratings after reliability had returned to its initial value were significantly lower (Mean = 7.633, SD = 1.261, F(1,54) = 15.88, $p < 0.001$) when system reliability had dropped by 45%, compared to when it had dropped by only 15% (Mean = 8.463, SD = 1.263). Also, a longer duration of low reliability led to a more significant decrease in subjective trust ratings, from an overall mean rating of 8.288 (SD = 1.169) following a short reliability drop to a rating of 7.808 (SD = 1.432) (F(1,54) = 5.311, $p = 0.025$) following a longer drop. No significant effect of sequence and no interaction effect were observed.

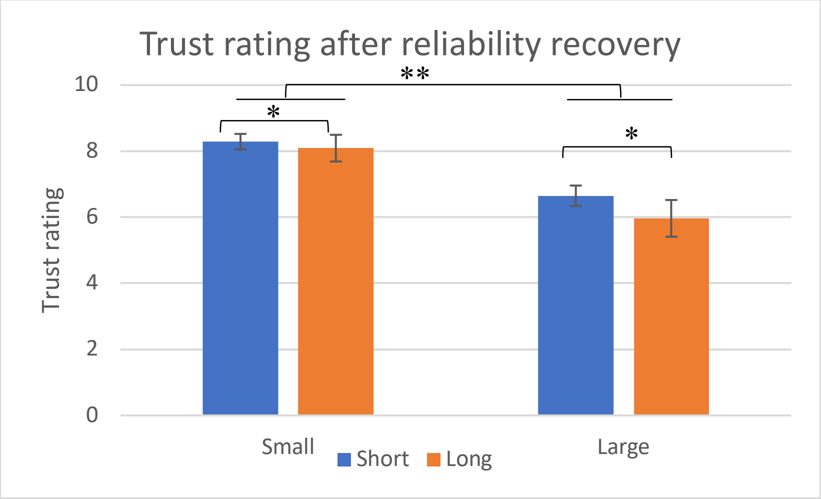


Figure 3.21. Effects of magnitude and duration on trust ratings after a reliability recovery

Eye tracking metrics

A 2 (magnitude) *2 (duration) * 2(sequence) linear mixed model analysis was conducted on total fixation duration percentage, total fixation count percentage and transition count after system reliability recovered. There was a marginally significant effect of magnitude on total fixation duration (F(1,50.47) = 3.749, *p* =0.058). No significant effect of duration was observed. There was, however, a significant interaction between magnitude and duration such that large versus small changes in reliability differed significantly (*p* = 0.006) in their effect on total fixation duration only for short duration changes in reliability.

The total fixation count percentage on AOI 2 was significantly smaller (Mean = 71.2%, SD = 0.144) when system reliability dropped from 95% to 50%, compared to when it dropped by 15% only (Mean = 77.3%, SD = 0.157, F(1,48.78) = 6.47, *p* = 0.014). No significant effect of duration and no interaction effect were observed.

Similar results were found for the transition count data which were affected significantly only by the magnitude of the change in system reliability. Transition counts between the two

AOIs were significantly larger (Mean =35.46, SD = 23.17) when system reliability dropped from 95% to 50%, compared to when it dropped from 95% to 80% (Mean =25.63, SD = 18.80, F(1,50.28) = 6.15, *p* =0.017). The results of the statistical analysis are summarized in Table 3.4.

Table 3.4. Effects of magnitude and duration on three eye tracking metrics after a reliability recovery

	Main effects		Interaction effects
Eye movement metrics	Magnitude	Duration	Magnitude & Duration interaction
Total fixation duration percentage	Not significant	Not significant	F(1,49.79) = 4.3, *p* = 0.043
Total fixation count percentage	Large: 0.712 (0.144) Small: 0.773(0.157) F(1,48.78) = 6.47, *p* = 0.014	Not significant	Not significant
Transition count	Large: 35.46(23.17) Small: 25.63(18.80) F(1,50.28) = 6.15, *p* = 0.017	Not significant	Not significant

Review of automation assessments

A 2(magnitude)*2(duration)*2(sequence) linear mixed model was applied to the number of reviews right after system reliability recovered. Participant ID was entered as random effect, and the number of reviews before system reliability dropped was entered as covariant. Again, the magnitude of a system reliability drop was the only factor that significantly affected the number of reviews (F(1,54) = 7.54, *p* = 0.008) which increased from 8 (SD = 7.23) in the small reliability change scenario to 10.9 (SD = 6.88) following a large reliability drop (Figure 3.22). No other main or interaction effects were observed.

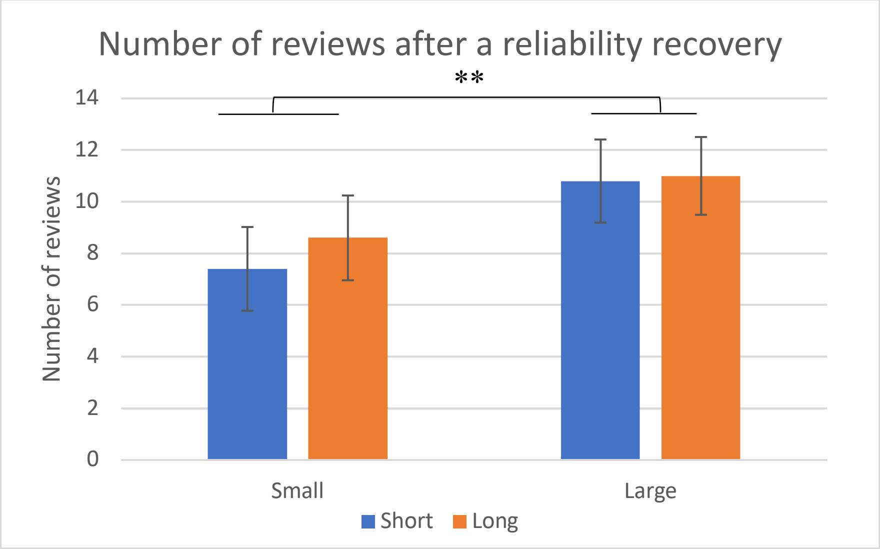


Figure 3.22. Effects of magnitude on the number of reviews after a reliability recovery

Tracking task performance

A 2(magnitude)*2(duration)*2(sequence) linear mixed model, with MSD before the reliability change entered as a covariant, did not yield main effects of magnitude or duration following system reliability recovery. However, a significant interaction between magnitude and duration (F(1,54) = 17.78, p <0.001) was observed, as shown in Figure 3.23. MSD during a small and short drop in reliability was larger than during a small and long drop (p = 0.003), while MSD during a large and short drop was smaller than during a large and long drop (p = 0.007).

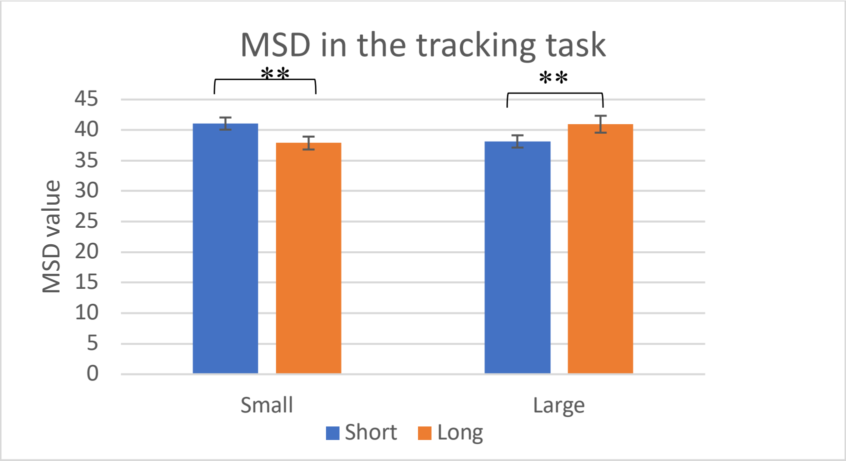


Figure 3.23. Interaction effect of magnitude and duration on tracking performance

Target detection performance

Overall error rate

Finally, the overall error rate during the three-minute period after system reliability had recovered to its original level was examined using a 2(magnitude)*2(duration)*2(sequence) linear mixed model. Only a marginally significant interaction between magnitude and duration (F(1,54) = 3.717, *p* = 0.059)(See Figure 3.24) was observed, such that the error rate was higher following recovery from a longer reliability drop only in case of a large drop (45%).

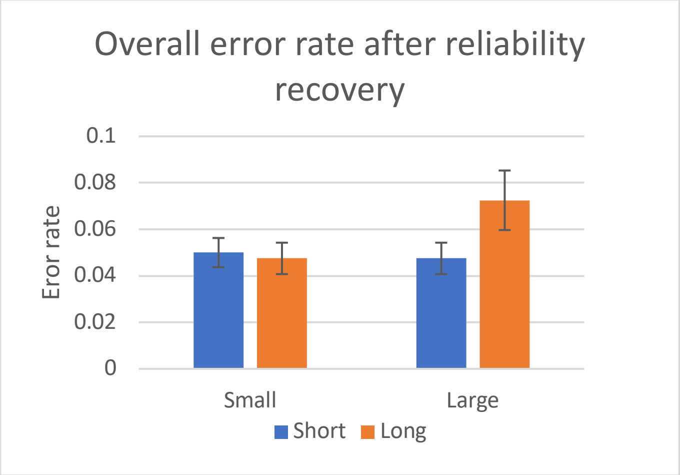


Figure 3.24. Effects of magnitude and duration on overall error rate after a reliability drop

Response time

The analysis of response times for button pushes did not yield any significant main or interaction effects.

Period 4: Time until trust recovery

'Time until trust recovery' is defined as the number of 3-minute periods that it took for participants' trust ratings to return to a steady (though not necessarily the initial) level following the recovery of system reliability to its initial high level. A 2 (magnitude) *2 (duration) * 2 (sequence) linear mixed model analysis was performed on the number of 3-minute periods required. Large reliability changes resulted in a significant increase in the time taken to recover trust (F(1,54) = 27.323, $p < 0.001$, Mean = 3.700, SD = 1.114), compared to trust recovery after small reliability changes (Mean = 2.725, SD = 0.987). There was no significant effect of the duration but there was a significant interaction between the magnitude of the reliability change and the sequence in which scenarios were experienced (F(1,18) = 4.042, $p = 0.049$). For both sequences, trust recovery took longer with large reliability drops; however, this effect was more pronounced when participants experienced the scenario sequence 4(small and short drop)->3(large and short drop)->2(small and long drop)->1(large and long drop) (Sequence 2) compared to the sequence 1->2->3->4 (Sequence 1), as shown in Figure 3.25.

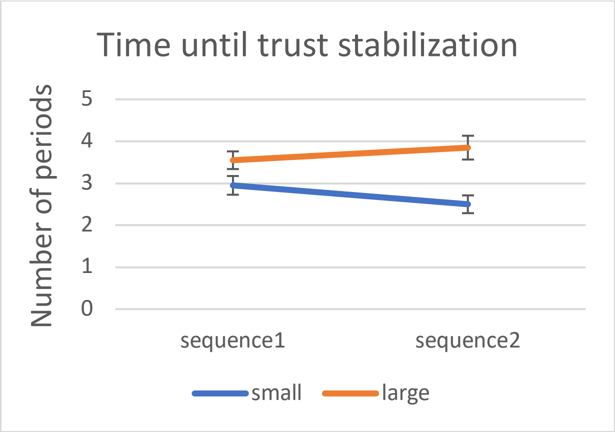


Figure 3.25. Effects of magnitude and sequence on recovery time

Results Summary

The following three tables provide a summary of the findings that were reported in the Results section.

Table 3.5. Detection of system reliability change

	Subjective ratings		**Eye tracking**		
	Perceived reliability	Subjective trust ratings	Total fixation duration PC	Fixation count PC	Transition count
Detection of system reliability change	✓ in all scenarios	✓ in all scenarios	✓ in all scenarios	✓ in all scenarios except when system reliability dropped for a short duration and small magnitude	✓ in all scenarios except when system reliability dropped for a short duration and small magnitude

Table 3.6. Effects of magnitude and duration of system reliability change on perceived reliability, subjective trust ratings, eye movements and behavior (NS: not significant)

Effects of magnitude and duration		**Subjective ratings**		**Eye tracking**			**Behavior**
		Perceived reliability calibration	Subjective trust ratings	Total fixation duration PC	Fixation count PC	Transition count	Number of reviews
3 min after the reliability drop	Magnitude ↑	↑	↓	↓	↓	↑	↑
9 min low reliability period	Magnitude ↑	↑	↓	↓	↓	↑	↑
	Time interval ↑	NS	NS	↓	NS	↑	↑
3 min after the reliability recovery	Magnitude ↑	↑	↓	NS	↓	↑	↑
	Duration ↑	NS	↓	NS	NS	NS	NS
	Interaction	NS	NS	Significant effect of magnitude when duration is short	NS	NS	NS

Table 3.7. Performance outcomes

Effects of magnitude and duration		**Tracking task (Mean square distance)**	**Target detection performance**	
			Response time	Error rate
3 min after the reliability drop	Magnitude ↑	NS	NS	↑
9 min low reliability period	Magnitude ↑	↑	NS	↑
	Time interval ↑	NS	NS	↓
	Interaction	NS	NS	Significant effect of time interval when magnitude was large Significant effect of magnitude in time interval 3 and 4
3 min after the reliability recovery	Magnitude ↑	NS	NS	NS
	Duration ↑	NS	NS	NS
	Interaction	After a short reliability drop, significant higher MSD in the small reliability drop. Reversed results after the long reliability drop	NS	Marginal significant: error rate was higher following recovery from a longer reliability drop only in case of a large drop

Discussion

The main goal of this experiment was to examine how the magnitude and duration of variations in system performance affect participants' trust calibration and recovery. The following sections summarize and discuss the main findings for the various trust-related dependent variables in the study, including subjective trust ratings, perceived reliability, eye movement metrics, behavioral data and performance outcomes.

Detection of system reliability changes as evidenced by changes in participants' perceived reliability, subjective trust ratings and eye movements

The analysis of the perceived reliability and trust ratings shows that participants were sensitive to all changes in system reliability, even to small and short drops in performance. However, this did not necessarily translate into changes to their visual attention allocation. Small and short drops in reliability affected fixation duration only (but not fixation or transition count). In those cases, participants did not immediately change their monitoring strategy in terms of how often they checked the target detection task. However, when they did check, they reviewed the scene more carefully as reflected in the longer fixation duration. One possible explanation for this behavior may be a fairly large switching cost (both in terms of performance and response time (Crandall & Cummings, 2007; Rogers & Monsell, 1995) associated with alternating between two different tasks - the target detection and the tracking task – and associated control settings. Participants may have been willing to incur this cost only in case of more severe (larger and/or longer) breakdowns in system performance.

Effects of magnitude and duration of reliability changes on perceived reliability, subjective trust ratings, eye movements, review of automation assessments and corresponding performance outcomes

Period 1: the first three minutes after a drop in system reliability

During the first three minutes of low system reliability, a large drop resulted in the worst perceived reliability calibration. In particular, participants tended to overestimate the automation reliability after it had dropped to its lowest level (50%). This result differs from Wiegmann (2001) who found that, when system reliability was low (in that study, 60%), participants still underestimated it. One possible explanation for this discrepancy in findings is that, in our experiment, participants always experienced high reliability first, which may have had a priming effect leading to a fairly high and robust baseline trust; in contrast, in Wiegmann's study, trials started at the low level of reliability of 60%.

In terms of participants' attention allocation and monitoring behavior, a large drop in system reliability resulted in more and longer fixations on the target detection task, more transitions between the tracking and target detection tasks, and a more frequent review of automation assessments. These results closely mirror the findings for perceived reliability and subjective trust ratings: perception/attitude changes due to a large drop resulted in behavioral changes in the sense that participants monitored and checked the automation more closely. The error rate for the target detection task was still higher with a large drop. The main reason for this finding is that the overall error rate included the performance of the automation when participants did not review the scene on their own. The change in trust and automation monitoring during the first three minutes after a large reliability drop did not affect participants' tracking performance. This may be explained by masking due to large individual variances on this task.

Period 2: the nine-minute period of low system reliability

During the nine-minute period of low system reliability, the effect of magnitude of the reliability drop was similar for subjective and behavioral measures. However, the effect of duration on subjective and behavioral measures diverged. While participants monitored the automation more often and more closely throughout a long period of low reliability (as indicated by increases in total fixation duration and fixation count for the target detection task AOI and an increasing transition count between the tracking and target detection task), perceived reliability and their trust ratings did not continue to decrease. This finding is consistent with previous research findings (Barg-Walkow & Rogers, 2016; Ezer et al., 2008) showing that perception and attitudes towards technology change more quickly and easily compared to automation reliance, i.e., the actual use of the automation, as it takes time and is more costly to change behavior rather than attitudes.

An effect of magnitude on tracking performance was observed during the later stages of low reliability when participants who experienced a large reliability drop showed significantly worse performance compared to participants who experienced a more minor reduction in system reliability. As indicated by all three eye tracking metrics and the frequency of reviewing automation assessments, participants focused increasingly on the target detection task and their performance on the tracking task was eventually sacrificed over the long duration of low system reliability. Interestingly, participants' overall error rate on the target detection task continued to decrease in case of a large reliability drop but not with a small drop in system performance. This may be explained by participants investing more effort into timesharing both tasks (Christopher D Wickens, 1990; Christopher Dow Wickens et al., 2016) to compensate for very poor automation performance and maintain a reasonable level of joint system performance.

Period 3: the three-minute period after system reliability recovered to its initial high level

Perceived reliability during the first three minutes after system performance returned to its initial high level was affected only by magnitude but not by duration of the reliability change. This confirms earlier research findings suggesting that perceived reliability estimates are based almost exclusively on the actual level of system reliability (e.g., Dzindolet et al., 2002;

Madhavan & Phillips, 2010). When actual system reliability returned to its initial high level (95%; from both 50 and 80%), participants tended to underestimate the reliability of the automation. This finding is consistent with results from previous studies (e.g. Madhavan & Wiegmann, 2007a; Wiegmann & Cristina Jr, 2000; Wiegmann et al., 2001) showing that a breakdown in system performance leads to difficulty in trust recovery and calibration.

In contrast to perceived reliability, subjective trust ratings were affected by both magnitude and duration. Specifically, a large drop in system reliability resulted in a significant loss of trust (Lee & Moray, 1992) and a long period of low reliability led to a further drop in trust ratings during the first three minutes after system reliability recovered. This finding is in line with earlier work suggesting that perceived reliability is only one of many factors affecting people's trust in a system. Perceived reliability may simply reflect observed moment-to-moment changes in system performance (Madhavan & Wiegmann, 2007a) while trust is formed based on multiple factors and their interactions, including system performance, process and purpose (as summarized by Lee and See, 2004). The resulting attitude of trust appears to be less dynamic, i.e., lag behind observed changes in system reliability.

During the 3-minute period after reliability recovered to its original high level, magnitude and duration interacted in their effect on tracking performance. When system reliability dropped for a long time, participants' tracking performance was worse with a large (as compared to a small) drop. Unexpectedly, when system reliability dropped for a short period of time, participants' tracking performance was worse with a small (as compared to a large) drop. One possible explanation for this finding is the sequence in which participants experienced the various reliability scenarios. One of the two combinations above - scenario 1 (large and long drop) or scenario 4 (small and short drop) - were experienced first by participants in this study.

As their tracking performance improved when they interacted with the system for a longer time, this may explain why performance was so poor with a small and short drop even.

After system reliability recovered to its original level, participants more often incorrectly disagreed with the automation after a large and long reliability drop, compared to a small and long performance reduction, even though their perceived reliability did not differ for those two scenarios. This may have resulted from a significant loss of trust after experiencing highly unreliable automation performance and from trust being slower to recover than perceived reliability, affecting their interaction with and reliance on the automation (Dzindolet et al., 2003).

Period 4: Time until trust recovery

In terms of trust recovery, a large change in system reliability (but not a long duration of low reliability) resulted in a significantly longer time taken to recover trust, compared to a small reliability drop. This finding suggests that the effect of duration on trust is short-lived, again confirming that the magnitude of system reliability changes has a more pronounced impact on trust development.

The time it took for trust to recover differed as a function of the sequence in which participants experienced the reliability scenarios. Trust recovery took significantly longer with a large drop in reliability for sequence 2 (small and short drop-> large and short drop-> small and long drop->large and long drop) than when participants experienced reliability changes in reverse order. This might be explained by the contrast effect, a cognitive bias that either enhances or diminishes perception due to the previous exposure to something similar (Schwarz, Münkel, & Hippler, 1990). When participants experienced the most reliable automation first and finished with the most unreliable automation, it took longer to recover trust following a large reliability drop, because participants have been anchored with highly reliable automation and this contrast made them more difficult to recover trust in the automation.

Relationship between subjective trust ratings and eye tracking metrics

Both subjective ratings and eye tracking data were collected in this experiment as possible indicators of trust. The fact that variations in system reliability affected eye tracking metrics in a similar way to perceived reliability and trust ratings strongly suggests the feasibility of using eye tracking as a promising less intrusive technique for inferring trust in real time (Lu & Sarter, 2019a). However, similar to Experiment 1, there were some discrepancies and inconsistencies between these two types of measures. Perceived reliability, in particular, was highly sensitive to and quickly changed in response to variations in actual reliability. Trust ratings lagged behind to some extent and eye movements and behavioral changes were even slower to respond. Still, compared to the intrusive nature of subjective ratings, this lag in behavioral measures is still acceptable for real-time inference of trust levels.

In summary, this study examined the effects of magnitude and duration of system reliability changes on trust calibration and recovery. The results from this experiment show that, overall, participants were sensitive to all four types of system reliability change (small-short, small-long, large-short and large-long). Both magnitude and duration of the system reliability drop had some effect on participants' trust calibration and recovery, with magnitude being the more critical factor. The effect of duration was limited to significantly lower trust ratings during the first three minutes after system recovery from a large and long reliability drop. And not surprisingly, the large and long reliability drop in system performance had the most severe impact on trust and target detection performance. Therefore, in the next two chapters, the focus will be on developing and testing an algorithm and alerting measure for improving trust calibration in this scenario.

References

Barg-Walkow, L. H., & Rogers, W. A. (2016). The effect of incorrect reliability information on expectations, perceptions, and use of automation. *Human factors, 58*(2), 242-260.

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied ergonomics, 52*, 333-342.

Crandall, J., & Cummings, M. (2007). *Attention allocation efficiency in human-UV teams.* Paper presented at the AIAA Infotech@ Aerospace 2007 Conference and Exhibit.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). *Impact of robot failures and feedback on real-time trust.* Paper presented at the Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., . . . Yanco, H. (2012). *Effects of changing reliability on trust of robot systems.* Paper presented at the Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on.

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.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human factors, 44*(1), 79-94.

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.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2008). Age-related differences in reliance behavior attributable to costs within a human-decision aid system. *Human factors, 50*(6), 853-863.

Guznov, S., Lyons, J., Nelson, A., & Woolley, M. (2016). *The effects of automation error types on operators' trust and reliance.* Paper presented at the International Conference on Virtual, Augmented and Mixed Reality.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*(10), 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153-184.

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

Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In *Foundations of trusted autonomy* (pp. 135-159): Springer, Cham.

Lu, Y., & Sarter, N. (2019a). Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability. *IEEE Transactions on Human-Machine Systems, 49*(6), 560-568.

Lu, Y., & Sarter, N. (2019b). *Feedback on system or operator performance: Which is more useful for the timely detection of changes in reliability, trust calibration and appropriate automation usage?* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Madhavan, P., & Phillips, R. R. (2010). Effects of computer self-efficacy and system reliability on user interaction with decision support systems. *Computers in Human Behavior, 26*(2), 199-204.

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human factors, 49*(5), 773-785.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors, 55*(3), 520-534.

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human factors, 57*(1), 34-47.

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*(2), 207.

Rossi, A., Dautenhahn, K., Koay, K. L., & Walters, M. L. (2017). *How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario.* Paper presented at the International Conference on Social Robotics.

Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors, 58*(3), 377-400.

Schwarz, N., Münkel, T., & Hippler, H. J. (1990). What determines a 'perspective'? Contrast effects as a function of the dimension tapped by preceding questions. *European Journal of Social Psychology, 20*(4), 357-361.

Wickens, C. D. (1990). Resource management and time-sharing. In *Human performance models for computer-aided engineering* (pp. 180-202): Elsevier.

Wickens, C. D., Gutzwiller, R. S., Vieane, A., Clegg, B. A., Sebok, A., & Janes, J. (2016). Time sharing between robotics and process control: Validating a model of attention switching. *Human factors, 58*(2), 322-343.

Wiegmann, D. A., & Cristina Jr, F. J. (2000). Effects of feedback lag variability on the choice of an automated diagnostic aid: a preliminary predictive model. *Theoretical Issues in Ergonomics Science, 1*(2), 139-156.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science, 2*(4), 352-367.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). *User trust dynamics: An investigation driven by differences in system performance.* Paper presented at the Proceedings of the 22nd International Conference on Intelligent User Interfaces.

Chapter 4

Development of Eye-tracking Based Algorithms to Infer Trust in Real Time

In Chapter 2 (Experiment 1), several eye tracking metrics were shown to reflect trust changes that resulted from priming and variations in system reliability. However, eye tracking metrics in that study were calculated over fairly long-time windows (two or three minutes) and were analyzed post-hoc, following the experiment. In order to support the ultimate goal of this line of research, namely the development of an alert system that detects and responds to observed trust miscalibration and potential performance breakdowns in a timely manner, the next step serves to establish 1) whether eye tracking metrics calculated over a shorter time window (such as one minute) can serve as valid indicators of trust and 2) whether modeling techniques such as machine learning can be applied to eye tracking data to infer operators' trust levels in real time.

As reviewed in Chapter 1, eye tracking metrics extracted from raw eye movement data have been used successfully as input to various modeling methods. For example, pupil diameter has been used to predict performance on various information search and integration tasks, using a random forest method (Buettner et al., 2018). People's personality traits could be detected based on eye tracking using classifying methods such as Naïve Bayes (Berkovsky et al., 2019). In most studies, the datasets used as input for the modeling process were based on groups of participants because classifying algorithms tend to require large data samples to train the model and achieve relatively high and stable performance. This "generalized" modeling process is appropriate in many cases. However, prediction accuracy with this approach can suffer due to large inter-individual differences in a dataset (Krull & MacKinnon, 2001) which have been observed for eye movements (Castelhano & Henderson, 2008). Also, very few studies to date have used raw eye movement data as input to machine learning models in order to make predictions. One benefit of modeling raw eye tracking data, rather than processed eye tracking features, is that this approach is more generalizable to a wide range of tasks and settings as the selection and calculation of effective eye tracking metrics can be highly associated with certain tasks and settings.

To address these two limitations, the work reported in this chapter examines the performance of two different modeling processes to infer trust: 1) using eye tracking metrics from individuals for "personalized" modeling, and 2) using eye tracking metrics from a group of participants for "generalized" modeling. In the latter case, two types of input data were compared: 1) using defined eye tracking metrics, and 2) using raw eye movement data (gaze point coordinates).

There is no agreement on which classification modeling methods are preferable. The utility of each method depends heavily on the data to be modeled. In this study, supervised machine learning techniques including Logistic Regression (LR), k-Nearest Neighbor (kNN), Random Forest (RF), Multi-Layer Perceptron (MLP) and Convolutional Neural Networks (CNN) were compared to determine which method works best to infer trust levels in real time. LR uses an equation that is very similar to linear regression but predicts probabilities between 0 and 1. The coefficients of the equation were estimated based on training data using maximum likelihood estimation. LR is light-weight and efficient for real-time processing (Moacdieh, 2015). There is evidence that LR can achieve similar prediction performance as other more complex machine learning techniques (Jie, Collins, Steyerberg, Verbakel, & van Calster, 2019). LR was therefore used as the baseline against which all other techniques were compared. kNN returns classification labels by comparing the feature similarity between the new sample and labeled training data sample using Euclidean distance (Dudani, 1976). It has the advantage of working well with a relatively small number of features. RF is an ensemble learning method that classifies a new sample based on the vote result from many trained decision trees (Breiman, 2001), where each decision tree provides a classification prediction based on different split features/criteria. Bagging and feature randomness ensure the modeling performance of RF. Bagging, also known as bootstrap aggregation, is the process of randomly sampling the dataset for decision making of each individual tree. Feature randomness means that the features used in each decision tree to make classifications are randomly selected. These two characteristics enable uncorrelated predictions from individual decision trees and thus optimize the modeling performance. Both kNN and RF have been used successfully with eyetracking data in other application domains (e.g., Berkovsky et al., 2019; Buettner et al., 2018). They both require little data preprocessing and parameter handling; therefore, they are considered promising candidates for the modeling process described in this chapter.

The two deep learning approaches in our set of classification modeling methods, MLP and CNN, were only employed for modeling the raw eye tracking data. MLP is a type of feedforward artificial neural network consisting of at least three layers (input, hidden, output) (Collobert & Bengio, 2004). It finds the best approximation (through assigning weights and bias) to map the input x (raw gaze coordinate in this case) to an output classification y (high or low trust in this case). CNN has at least one convolution layer to replace the general matrix multiplication used in MLP (Goodfellow, Bengio, & Courville, 2016). Compared to MLP, it can capture the spatial and temporal dependencies in the data. These two methods were specifically applied to model raw eye tracking data because they are well suited for identifying the intricate structure of large data sets and have been shown to dramatically improve the modeling results in various application domains (LeCun, Bengio, & Hinton, 2015). Figure 4.1 provides an overview of the modeling process and classification methods used in this study.

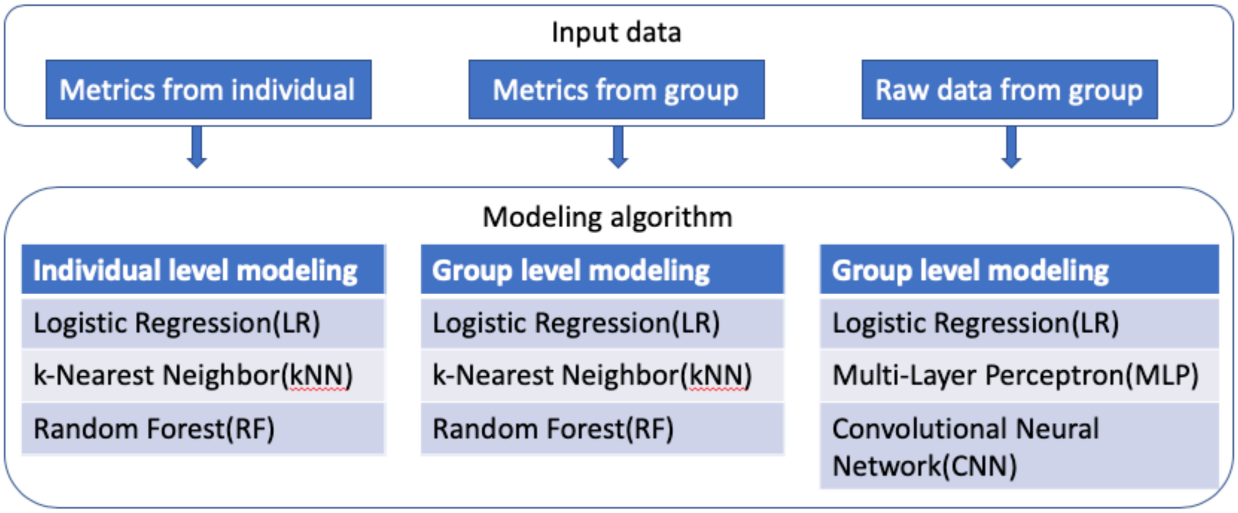


Figure 4.1. Modeling process overview

Data collection

Data collected in Experiment 2 (Chapter 3) were used as input to the modeling processes. Specifically, baseline eye tracking data were collected for two scenarios (high (95%) and low (50%) reliability) after participants completed all formal trials in Experiment 2. Each scenario lasted 6 minutes, and participants were informed about the respective system reliability in advance. The sequence in which the two reliability settings were experienced was randomized.

Data processing and modeling

Three types of data sets were tested as input for the modeling process, as shown in Figure 4.1: (1) eye tracking metrics for each individual participant, (2) eye tracking metrics for all participants combined, and (3) the raw eye movement data for all participants combined.

The eye tracking metrics used in this study were fixation duration percentage and transition count. Fixation duration percentage is the percentage of fixation time that a participant focuses on the display area showing the tracking task (see AOI2 in Figure 4.2). Transition count is number of transitions between AOI 1 (Area of Interest 1: the target detection area) and AOI 2 (the tracking task area) (see Figure 4.2). These metrics were selected for two reasons: 1) they were identified as promising indicators of trust levels in Experiments 1 and 2, and 2) they were found to be closely associated with target detection performance when trust calibration suffered in Experiment 2.

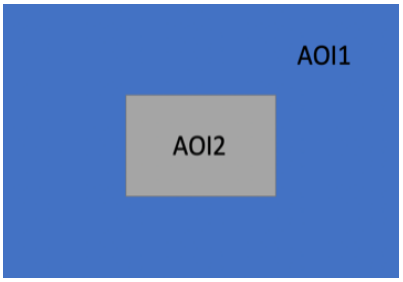


Figure 4.2. AOI definition

For the raw eye movement dataset, each data sample consisted of the coordinates of raw gaze points in the form of (x,y). These data were preprocessed, including dealing with missing data, down-sampling and standardization. Each missing data point was filled with the previous data point. The whole dataset was down-sampled from 50 Hz to 1 Hz using average pooling, i.e., calculating the average x, y value for every 50 raw eye movement points. Finally, input standardization (a process of being subtracted by the mean and then divided by the standard deviation) was performed on the data.

For the calculation of all model input data, a moving time window of one minute was used. This time window was shorter than those used in Experiments 1 and 2 in order to be able to detect trust miscalibration and trigger alerts in a more timely fashion. It was based on extensive piloting which showed that one minute of eye movement data is sufficient to avoid excessive variance in the data and reflect participants' trust levels in near real time. Successive time windows were overlapped by ten seconds to generate enough data samples and implement the trust level inference in real time.

Incomplete data from 4 participants were discarded, leaving data from 16 participants as input to the modeling process. For individual-level modeling, data from the first three minutes were used as the training set. The next 1.5 minutes were used as the validation set to select the most appropriate parameters in the model. The last 1.5 minutes were used as the test set. For group-level modeling, data from 4 of the 16 participants were randomly selected as test sets. For model selection, twelve-fold cross validation was applied, meaning that in each model training process, data from eleven participants were used as the training set and data from the twelfth was used as the validation set. This process was repeated twelve times so that data from each participant was used once as a validation set. The final validation performance was the average of the outcome of the twelve trials.

For the feature extraction data set, three classifying methods were explored: Logistic Regression (LR), k-Nearest Neighbors (kNN) and Random Forest (RF). These methods were applied using standard implementations from the Python library "Scikit-learn" (Pedregosa et al., 2011).

For the raw eye movement data set, LR, MLP and CNN were compared. MLP and CNN were implemented using TensorFlow (Abadi et al., 2016). Specifically, an MLP with two hidden layers was used. Each hidden layer had 64 units with Relu activations. Batch normalization was applied after each hidden layer. The output layer had two output units and Softmax activation. For CNN, one 1-D convolutional layer with 8 kernels of size 3 was used, followed by a fully connected layer with 32 units. Both the convolutional layer and the fully connected layer used Relu activations. As for MLP, batch normalization was applied after each hidden layer. The output layer had two output units and Softmax activation.

Modeling performance

In this section, results for each of the three modeling processes are presented. The performance of each classifying method was assessed by comparing the accuracy and efficiency of inferring trust levels in the test datasets.

Accuracy

Individual level modeling using eye tracking metrics

Figure 4.3 shows the accuracy of three classifiers when individual-based eye tracking metrics were used as input to the models. RF achieved the highest average accuracy rate at 93% (SD = 0.047). The mean accuracy for 16 participants when using kNN was 92.2% (SD = 0.053). And logistic regression resulted in a mean accuracy rate of 90.6% (SD = 0.059).

Group level modeling using eye tracking metrics

The modeling accuracy of the same three classification methods with group-level eye tracking metrics as input is shown in Figure 4.4. RF achieved the highest accuracy at 81.5% (SD = 0.077), followed by kNN at 80.7% (SD = 0.130), and logistic regression with an accuracy of 77.5% (SD = 0.055).

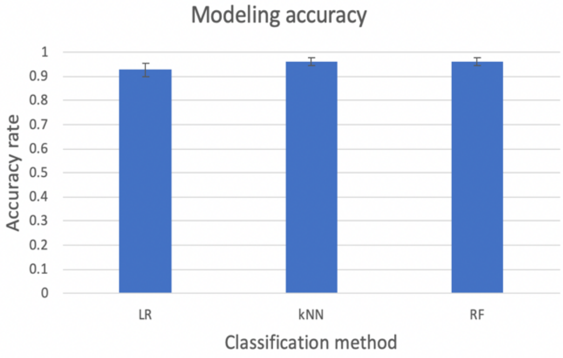


Figure 4.3. Personalized prediction

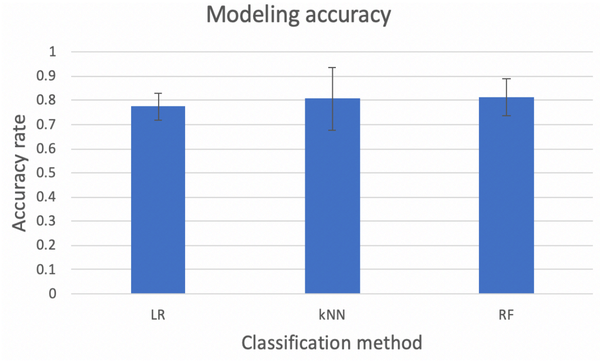


Figure 4.4. Generalized prediction

Group level modeling using raw eye movement data

The accuracy of the three classification methods when using raw eye movement data as input is shown in Figure 4.5. CNN outperformed the other methods and achieved an accuracy rate of 80.7% (SD = 0.096). MLP and LR did not perform well when using raw eye movement data as input. Their accuracy rate was substantially lower at 60.6% (SD = 0.031) and 44.1% (SD = 0.041), respectively.

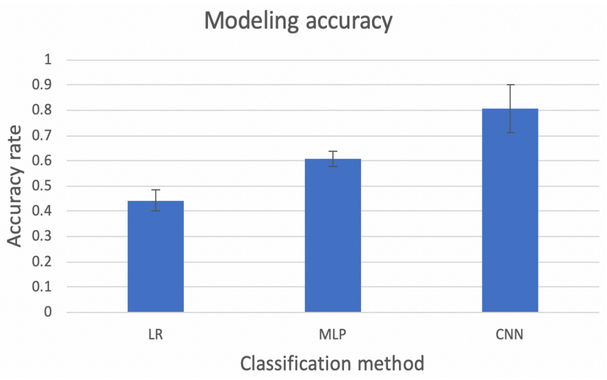


Figure 4.5. Generalized prediction using raw data

Efficiency

The modeling efficiency of the three procedures and for each algorithm is shown in Table 4.1. For individual-level modeling, LR was the most efficient (Mean = 0.48 ms, SD = 0.22). The modeling time using kNN was 1.3 ms (SD = 0.38), and the modeling time for RF was the slowest with a mean of 4 ms (SD = 0.61). Similar efficiency trends were observed for group-level modeling, with LR being the fastest and RF being the slowest approach. For group-level modeling using raw eye movement data, it took longer to infer trust than with the other two modeling procedures. Here, LR was the fastest technique with a mean modeling time of 2.4 ms (SD = 0.22). MLP and CNN were much slower with a mean modeling time of 72.5 ms (SD = 5.73) and 68.4 ms (SD = 2.06), respectively.

Table 4.1. Modeling efficiency

Algorithm	Time (ms)		Algorithm	Time (ms)
	Metrics from individual	Metrics from group		Raw data
LR	0.48 (0.22)	0.46 (0.17)	LR	2.43 (0.22)
kNN	1.25 (0.38)	5.09 (0.38)	MLP	72.45 (5.73)
RF	3.98 (0.61)	22.8 (0.56)	CNN	68.4 (2.06)

Discussion

Effective support for trust calibration requires an unobtrusive method for inferring trust in real time. To this end, eye tracking data were used as input to several modeling procedures and classification algorithms, and the accuracy and efficiency of these approaches were compared. The findings indicate that 1) eye tracking data extracted from a fairly short one-minute time window are sufficient to infer trust levels in real time, 2) modeling at an individual level outperformed modeling at a group level when using two eye tracking metrics as input, and 3) at a group level, both raw eye movement data and eye tracking metrics worked reliably (achieving an accuracy of around 80%) for inferring high and low trust levels when using machine/deep learning techniques.

The three modeling processes used in this research all achieved an accuracy of 80% or higher, demonstrating the feasibility and effectiveness of using eye tracking data to infer trust in real time. The much higher accuracy achieved with individual-level modeling, compared to generalized modeling, is likely due to significant interindividual differences due to dispositional factors affecting operators' trust in automation (Merritt et al., 2013; Merritt & Ilgen, 2008). In addition, this modeling process was much more efficient than group-level modeling in terms of the time taken to infer trust.

When modeling at an individual level, there was no significant difference between the accuracy of kNN and RF. However, other aspects of these techniques, such as their efficiency and interpretability (Wright, Chen, & Lakhmani, 2019) also need to be considered when selecting a method for use in real-world safety-critical environments. As shown in Table 1, kNN was about three times faster than RF, suggesting kNN may be the more appropriate method for settings that involve large data sets. Also, kNN appears to be more transparent, and its results are easier to interpret than those from RF. For example, as shown in Figure 4.6, it is easy to visualize how kNN works using a 2D figure. When a new data sample arrived, this algorithm found out the nearest k neighbors around the new data sample and determined the label of the new data sample based on the majority of labels in the circle. Thus, in summary, kNN may be considered the most appropriate algorithm for use in our subsequent work described in Chapter 5.

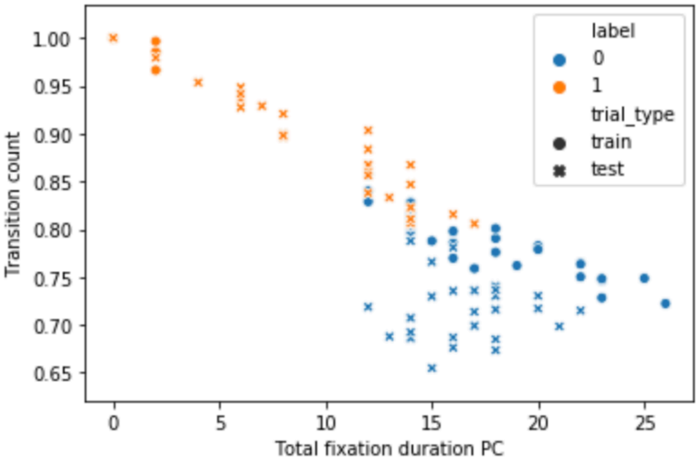


Figure 4.6. An example of kNN modeling

The accuracy observed for the generalized modeling approach suggests that data collected from one group of operators can be applied, to an extent, to other cohorts. While not as accurate as personalized modeling, generalized modeling was still rather effective at over 80% accuracy. This means that, in situations where training and collecting individual baseline data is not feasible, trust modeling and inference is still possible, to a more limited extent, using this method. Also, with more data samples being collected over time and being added to the modeling process, the accuracy of this approach may well increase further.

Based on our findings, it appears that using raw eye movement data as input to the modeling process can result in similar levels of accuracy as using eye tracking metrics. The increase in performance from LR to MLP demonstrates that non-linear transformation of raw data is required for classifying raw eye movement data. The increase in performance from MLP to CNN illustrates that incorporating the inductive bias of temporal invariance can further improve the performance of the classification model.

Our eye tracking-based approach outperforms the use of EEG and GSR in an earlier study on building trust classification models (Akash et al., 2018). In their study, their general trust sensor model (for all participants) achieved a mean accuracy of 71.22% and individual model based on customized features (for each participant) achieved a mean accuracy of 78.55%.

The fact that the performance of group-level modeling was similar when using either eye tracking metrics or raw eye movement data supports the robustness of deep learning algorithms. It suggests that the complex and time-consuming step of feature extraction may not be needed to achieve high performance. However, the notably longer time required to model trust using raw eye movement data suggests a tradeoff between the time spent on data preprocessing and modeling.

In summary, individual-level modeling using kNN can be considered the most effective approach for inferring trust levels in real time. It was shown to be most accurate, efficient and transparent technique of the ones compared in this study. In the next and final step of this line of research, kNN will be used for the development of an audio alert system to help facilitate trust calibration and prevent potential resulting performance breakdowns.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., . . . Isard, M. (2016). *Tensorflow: A system for large-scale machine learning.* Paper presented at the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16).

Akash, K., Hu, W.-L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using eeg and gsr. *ACM Transactions on Interactive Intelligent Systems (TiiS), 8*(4), 27.

Berkovsky, S., Taib, R., Koprinska, I., Wang, E., Zeng, Y., Li, J., & Kleitman, S. (2019). *Detecting Personality Traits Using Eye-Tracking Data.* Paper presented at the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Buettner, R., Sauer, S., Maier, C., & Eckhardt, A. (2018). Real-time Prediction of User Performance based on Pupillary Assessment via Eye Tracking. *AIS Transactions on Human-Computer Interaction, 10*(1), 26-56.

Castelhano, M. S., & Henderson, J. M. (2008). Stable individual differences across images in human saccadic eye movements. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale, 62*(1), 1.

Collobert, R., & Bengio, S. (2004). *Links between perceptrons, MLPs and SVMs.* Paper presented at the Proceedings of the twenty-first international conference on Machine learning.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*(4), 325-327.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*: MIT press.

Jie, M., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate behavioral research, 36*(2), 249-277.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors, 55*(3), 520-534.

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 50*(2), 194-210.

Moacdieh, N. M. (2015). *Eye Tracking: A Promising Means of Tracing, Explaining, and Preventing the Effects of Display Clutter in Real Time.*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research, 12*(Oct), 2825-2830.

Wright, J. L., Chen, J. Y., & Lakhmani, S. G. (2019). Agent transparency and reliability in human–robot interaction: the influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems*.

Chapter 5

Development and Evaluation of A Real-time Countermeasure to Trust Miscalibration, Inappropriate Reliance on Automation and Associated Performance Decrements

The experiment reported in Chapter 3 revealed that both the magnitude and the duration of a change in system reliability affected trust calibration and recovery, though to varying degrees. With a combined large and long drop in system reliability, participants' trust calibration suffered the most, and their performance was negatively affected due to inappropriate reliance on the automation. In particular, they underutilized the system when the automation was actually highly reliable. In an effort to try to correct trust miscalibration before it can lead to performance decrements, an eye-tracking based algorithm was developed that infers trust levels and variations in real time (Chapter 4). Personalized modeling using k-Nearest Neighbor was selected from among various candidates as the most accurate, transparent and fastest method. In the experiment reported in this chapter, this algorithm was used to trigger a candidate countermeasure to trust miscalibration in the form of an audio alert. Earlier studies have shown that discrete audio alerts are a good choice in domains such as Unmanned Aerial Vehicle (UAV) control as they do not interfere or compete with the visual demands imposed by the primary tasks. Audio alerts are reliably detected and particularly effective for supporting operator performance (Dixon, Wickens, & Chang, 2005; Graham & Cummings, 2007). In general, independent of a particular application domain, the auditory channel tends to be the preferred modality for simple warnings and for triggering immediate responses (Woods, 1995). This is true especially when the alert is implemented as a command display, i.e., a type of display that indicates what action to take (as opposed to status displays which describe the nature of a problem). Command displays have been shown to improve users' attention management and multitasking performance (e.g., McGuirl & Sarter, 2006; Prinet, 2016). In this study, the audio alert was used to notify participants when the algorithm described in chapter 4 detected that their trust level was not proportional to the trustworthiness of the automation. The alert instructed participants where to orient their visual attention, to the center (the tracking task) or the surround (the target detection task). Half of the participants were presented with the audio alert while the baseline condition completed the experiment without this feedback.

Based on the literature review and our findings from the previous two experiments, we expected that perceived reliability and trust in the audio alert group would more closely reflect actual levels and changes in system reliability and lead to behavioral adjustments in terms of attention allocation and automation reliance. More specially, the expectations were that:

(1) during the first 6 minutes of each trial when system reliability was high (red rectangle – see Figure 5.1 below), participants' perceived reliability and trust would increase over time,

(2) during that same time period (red rectangle), participants in the audio alert group would be better calibrated in terms of perceived reliability and trust, leading to more appropriate reliance on the automation and better overall task performance,

(3) during the three-minute period following a drop in actual system reliability (yellow rectangle), participants in the audio alert group would adjust their perceived reliability, trust and behavior more quickly and appropriately, while for participants in the control group, perceived

reliability and trust would initially be too high, leading to overreliance on the automation and worse performance,

(4) in the control group, perceived reliability and trust would continue to decrease and slowly and eventually approach actual system reliability during large and long reliability drops (blue rectangle); their behavior would change accordingly, ultimately leading to performance levels similar to those of participants in the audio alert group,

(5) during the first three minutes after system reliability returned to its original high level (black rectangle), participants in the audio alert group would show better trust calibration (as reflected by smaller perceived reliability calibration value and more appropriate trust levels) and thus more appropriate reliance on the automation; trust recovery (green bracket) would take longer in the control group leading to worse performance during those first three minutes (black rectangle).

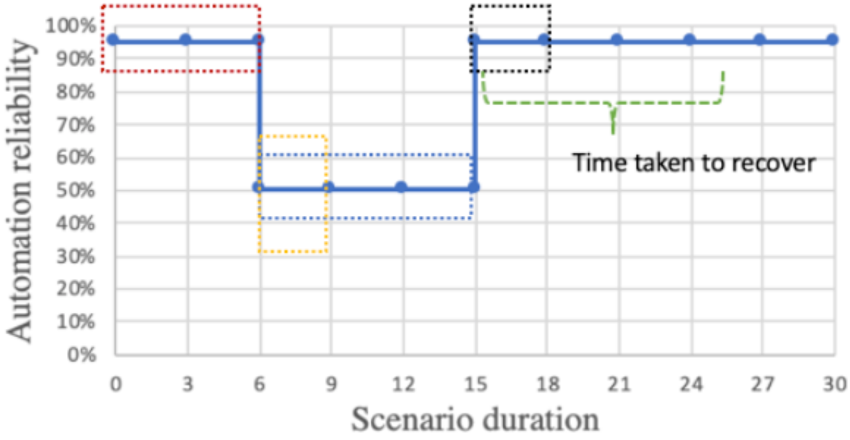


Figure 5.1. Scenario

Method

Participants

A total of 32 students from the University of Michigan participated in this experiment. Their average age was 22.5 years (SD=2.74). Half of the participants were males. Participants were required to have normal or corrected-to-normal vision. Due to the use of eye tracking glasses in the experiment, contact lenses were allowed but glasses were not. None of the participants had any experience with Unmanned Aerial Vehicle (UAV) control or the MATB tracking task before the experiment. Participants were randomly assigned to one of two groups: 1) baseline – no audio alert and 2) audio alert group. Independent samples t-tests showed that these two groups did not differ in terms of age, propensity to trust, experience with video games, problem-solving and decision-making capabilities (the assessment of these variables will be described in more detail in the Experiment Procedure section). This study was approved by the UM Health Sciences and Behavioral Sciences Institutional Review Board (exempt and not regulated; HSBS-IRB; ID: HUM00173710).

Apparatus and tasks

This study was set in a computer-based Unmanned Aerial Vehicle (UAV) control simulation built in the THInC lab. Participants were asked to complete a target detection task and a tracking task at the same time (see Figure 5.2). The two tasks were displayed on one 24'' monitor. A detailed description of the two tasks can be found in Chapter 3. The eye tracker used in this study was the same as in the earlier experiments – Tobii Glasses 2.

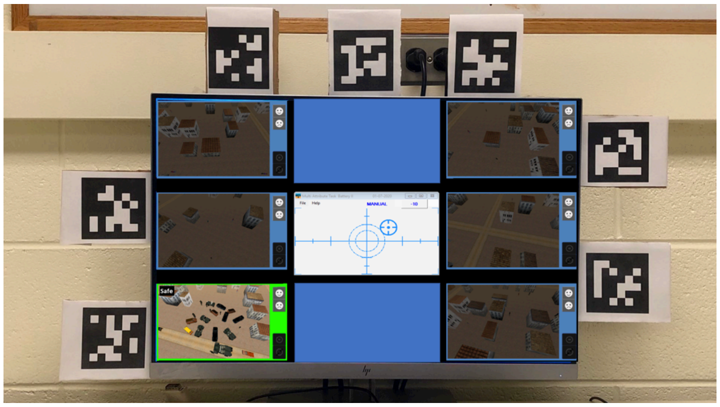


Figure 5.2. Simulation setup

For the most part, the setup in this experiment was the same as in Experiment 2 (Chapter 3); however, there were a few notable differences. First, for the target detection task, participants were asked to monitor the video feeds from 6 unmanned aerial vehicles (UAVs) (instead of 8 UAVs in Experiment 2) to detect the presence of military targets. As shown in Figure 5.2, QR code hardboards were attached to the display for real-time mapping of the 3-D eye tracking data points to the static 2-D image of the simulation setup (Figure 5.2). Pilot tests showed a large error when mapping data points in the vertical direction. Therefore, the number of video feeds was reduced to six only which allowed us to divide the screen into two Areas of Interest (AOIs), as shown in Figure 5.3.

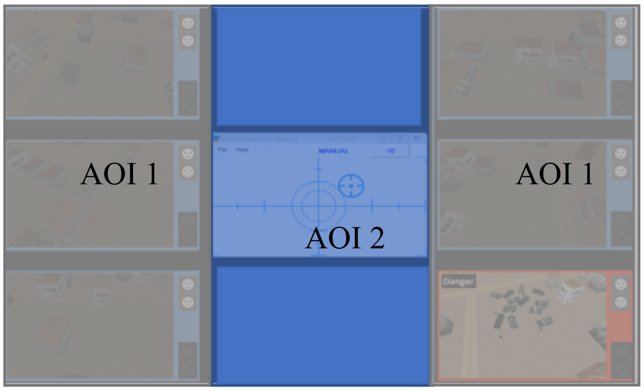


Figure 5.3. Area of Interest (AOI) definition

The second difference was that there were two types of targets which were defined as either a black truck carrying a gun (See Figure 5.4, inside the red circle) or a small tank (See Figure 5.5, inside the red circle) surrounded by other vehicles without weapons to increase the difficulty of the target search task. The video feeds also included distractors in the form of similar-looking vehicles that did not carry weapons, as shown in Figure 5.6. Figure 5.7 and Figure 5.8 show the targets and nontargets looked like in the simulation. The third difference was that, if participants decided to review the scene, they had to press one of two buttons (☺ or ☹; rather than (√ or ×) which were used in the earlier experiments) to either agree or disagree with the automation assessment regarding the presence of a target. This change was made to increase the intuitiveness of the buttons. Lastly, to ensure that participants correctly interpreted the red and green border around the video feeds, as shown in Figure 5.7 and 5.8, the words “danger” or “safe” were added for redundancy.

Participants were instructed to treat the target detection task and the tracking task as equally important and to try to achieve the best possible performance on both tasks.

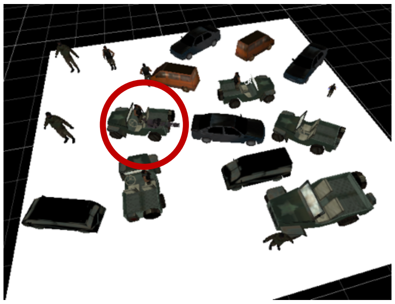


Figure 5.4. Target example one

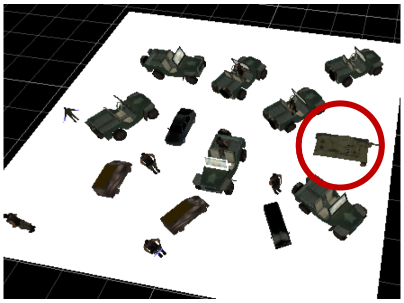


Figure 5.5. Target example two

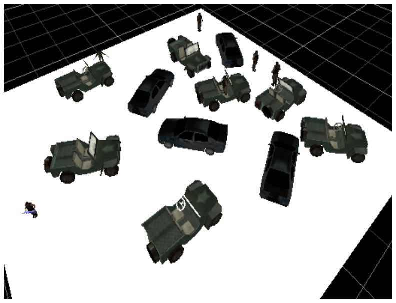


Figure 5.6. Non-Target example

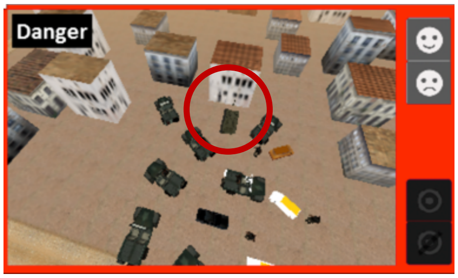


Figure 5.7. Target in simulation

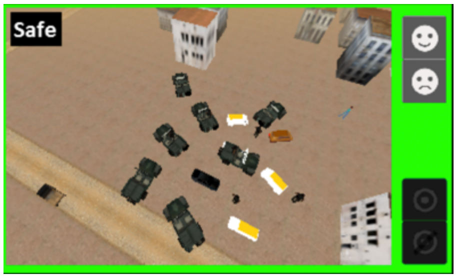


Figure 5.8. Non-target in simulation

Experiment design

The experiment employed a between-subject design. Participants were randomly assigned to one of two groups: half of the participants received an audio alert if trust miscalibration was detected by the algorithm, while the other half did not receive such feedback (baseline). The trust inference model was based on participants' eye tracking data over the last one-minute period. If the predicted value was 'high trust' but the system's trustworthiness was actually low, an audio alert "surround-surround" was triggered to encourage participants to re-orient their attention from the tracking to the target detection task. In contrast, if the predicted value was 'low trust' but the system was actually highly reliable, an audio alert "center-center" was played. The audio alert thus represented a command display in that it indicated to the participant where (on which task) to focus their attention.

The experiment employed the system reliability scenario shown in Figure 5.1. During the first six minutes, the automation was highly reliable (95% reliable), then system reliability dropped from 95% to 50% and stayed at 50% for a relatively long period of time (9 minutes) before recovering to the initial value of 95%. This scenario was chosen because it resulted in the worst trust miscalibration and poorest task performance in Experiment 2.

The dependent measures in this study were subjective trust ratings, perceived automation reliability, eye movement data, reviews of the automation assessments and overall task performance. Trust ratings and perceived reliability estimates were collected every three minutes, throughout each trial, by asking participants for a verbal response to two questions: (1) “How much do you trust the automation?” (on a scale from 0-10, with ‘0’ being the lowest possible trust) and (2) “What is your perceived reliability of the system” (on a scale of 0-100%).

The raw 3-D eye tracking data collected in real-time were first mapped to the fixed 2-D image as shown in Figure 5.2. The gaze points were then used to calculate fixations based on the Velocity-Threshold Identification (I-VT) fixation classification algorithm described by Tobii (Olsen, 2012a). This ensured that the outcomes were comparable to the outputs from Tobii Pro Lab used in our previous studies. The fixation classification algorithm was validated using eye tracking output from Experiment 2. The result showed that the fixation classification achieved an accuracy of 96%. The fixations were then used to calculate two eye tracking metrics: total fixation duration percentage (the percentage of fixation time that a participant focuses on the tracking task) and transition count (the number of transitions between the tracking task and target detection task).

The performance data consisted of participants’ scores on the tracking and monitoring tasks and their overall performance (cumulative combined score of monitoring and tracking). Tracking task performance was defined as the Mean Square Distance (MSD) of the small circle target from the central. The movement of the small circle was based on a series of random numbers and thus not predictable. If participants’ MSD over the past 3 seconds was smaller than 35, they would get ten points; if their MSD was between 35 and 55 over the past 3 seconds, their points remained the same; if their MSD was over 55, they would lose ten points. The monitoring task performance was calculated by combining the participants' accuracy and the performance of the automation. If participants decided to review the video feed to determine whether there was a target and if they made a correct decision, they would obtain ten points. If they made a wrong decision, they would lose ten points. If participants decided to rely on the automation and not review the video feed, they would get ten points if the automation was correct and lose ten points if the automation was incorrect.

Experiment procedure

Upon arriving at the laboratory, participants first read and signed the consent form. They were then asked to fill out a background questionnaire asking for basic information (e.g., age, gender, nationality). The questionnaire also assessed their propensity to trust (Merritt et al., 2013), experience with video games (Clare et al., 2015), and their problem-solving and decision-making capabilities (Wiegmann, 2002).

Next, participants were instructed on the two tasks and completed a 12-minute training session. Participants in the audio alert group were also informed about the meaning of the audio alerts. Then, all participants experienced a 6-minute scenario without any automation assistance on the monitoring/target identification task. Each time the video feed was highlighted (without any green/red border), participants had to determine on their own whether a target was present and then click the corresponding button to the side of the feed. At the end of this manual trial, participants were informed about their accuracy on the target detection task. This feedback served to help participants decide how much they should rely on the automation to achieve the best performance, as the choice to rely on automation has been shown to depend on both an estimate of automation reliability and one’s own proficiency/self-confidence (Lee & Moray, 1994).

Participants then completed four 6-minute trials to establish a baseline for their handling of the tracking and target detection tasks with 4 different levels of automation reliability (95%, 80%, 65% or 50%). Participants experienced these four trials in randomized order. They were informed about the system reliability in advance of each trial and then performed the same tasks as in the training session. These baseline data were collected for the purpose of developing the trust inference model and for triggering audio alerts in the formal trial. Even though these baseline data were not needed for participants in the control group as they would not receive audio alerts, the same procedure was followed to ensure that participants in both groups spent the same amount of time practicing the task and interacting with the automation. After each baseline trial, participants were asked for subjective ratings of their trust in the automation (on a scale from 0-10, with ‘0’ being the lowest possible trust). Based on their responses, the trial was marked as either high or low trust. If participants’ subjective trust ratings were lower than ‘7’ and they frequently checked the target detection task, the trial was considered low trust. If they reported trust ratings higher than or equal to ‘7’, and they rarely or occasionally checked the target detection task, it was considered high trust. If participants’ ratings were not consistent with their behavior, the trust level was based on their behavior (checking automation accuracy) rather than attitude (trust rating).

Following the baseline data collection, each participant completed the scenario shown in Figure 5.1. The scenario ended once a steady state of trust in the automation was reached after the return of system reliability to its initial high level. Participants were then asked to fill out a debriefing questionnaire. It took participants around 90 minutes to complete the entire experiment, and they were compensated $30 for participating in the experiment.

Results

This experiment examined how audio alerts in case of trust miscalibration affected both the processes of trust development/recovery and attention allocation and the resulting performance outcomes for the monitoring and tracking tasks. The first part of the Results section focuses on process. It is organized around five time periods within each trial, as shown in Figure 5.1: (1) the first 6 minutes when automation reliability was high (red rectangle), (2) the three-minute time period right after a drop in reliability (yellow rectangle), (3) the nine-minute time period during which system reliability was consistently low (blue rectangle), (4) the first three minutes after recovery of automation reliability back to its initial high level (black rectangle), and (5) the time taken to fully recover trust after the system breakdown (the green bracket). Unless otherwise specified, linear mixed models were applied to analyze all dependent variables. The fixed effects were audio alert and time interval (if there was more than one time interval). Participant ID was entered as a random effect. For all analyses, the significance level was set at p <0.05. Error bars on the figures indicate the standard error of the mean.

Period 1: The six-minute high-reliability period at the beginning of the trial

The six-minute period at the beginning of the formal trial was examined to determine whether perceived reliability and trust increased when automation performance was nearly perfect (expectation 1) and whether participants in the audio alert group were better calibrated in terms of perceived reliability, trust (including subjective trust ratings and eye tracking metrics) and automation reliance, leading to better overall performance (expectation 2).

Perceived reliability

To examine whether participants' perceived reliability was well calibrated, the difference between actual system reliability and participants' perceived reliability was calculated. This measure is considered one valid indicator of the trust calibration process (Merritt et al., 2015). The analysis shows a significant effect of time interval (F(1,30) = 5.922, $p = 0.021$). Participants' perceived reliability calibration was significantly smaller in the 2nd three-minute interval (Mean = 0.081, SD = 0.434) than in the 1st three-minute interval (Mean = 0.3, SD = 0.671) (Figure 5.9). No significant effect of audio alert and no interaction effects were observed.

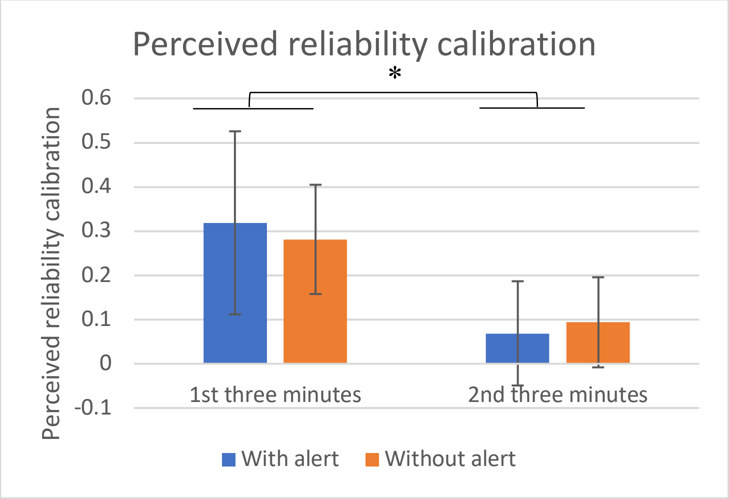


Figure 5.9. Perceived reliability calibration in the first 6 minutes

Subjective trust ratings

There was also a significant effect of time interval on subjective trust ratings. Participants' trust ratings were significantly higher (Mean = 8.609, SD = 1.210, F (1,30) = 8.576, *p* = 0.006) in the second three-minute period, compared to their ratings (Mean = 9.072, SD = 0.798) for the first three minutes. No significant effect of audio alert and no interaction effect was observed, as shown in Figure 5.10.

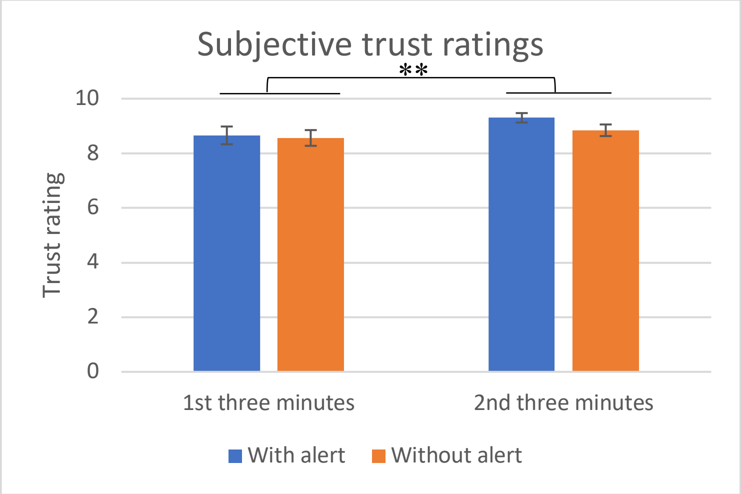


Figure 5.10. Subjective trust ratings in the first 6 minutes

Eye tracking metrics

The results show a significant effect of time interval on total fixation duration percentage (but not on transition count; see Table 5.1). The average percentage of total fixation duration on AOI2 (the tracking task window) was significantly lower (88.2%) during the first 3-minute period, compared to next three minutes (90.6%).

Table 5.1. Time interval effects on eye tracking metrics

Eye tracking metrics	First 3-min mean	Second 3-min mean	Main effect of time interval
Total fixation duration percentage	0.882 (0.078)	0.906 (0.073)	F(1,54.14) = 20.99, *p* < 0.001
Transition count	26.34 (18.24)	26 (16.21)	Not significant

Reviews of automation assessments

There was a significant effect of time interval on review frequency. In the second 3-minute period, participants reviewed the automation assessments less often (Mean = 5.97, SD=5.88), compared to the first three minutes (Mean = 8.38, SD=6.1, F(1,30)=7.993, *p* =0.008), as shown in Figure 5.11. There was also a significant effect of audio alerts on review frequency. Participants who received audio alerts reviewed the automation assessment less often (Mean = 4.69, SD = 4.86), compared to the baseline condition (Mean = 9.66, SD = 6.20, F(1,30) = 7.988, *p* = 0.008).

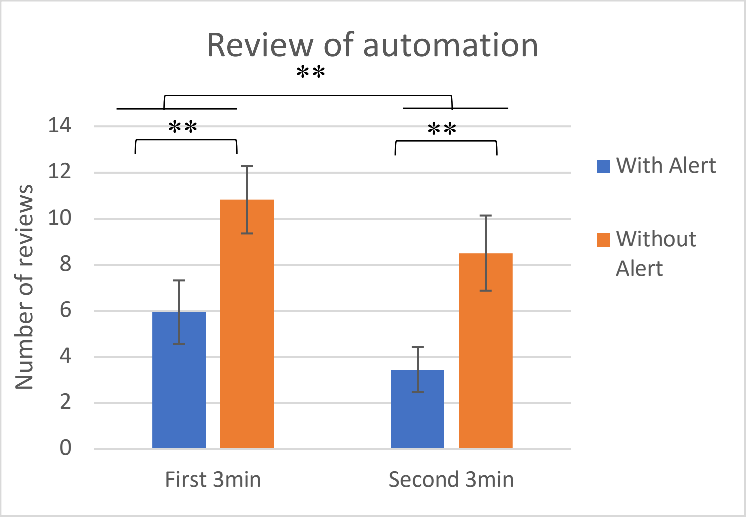


Figure 5.11. Number of reviews as a function of audio alerts and time interval

Tracking performance – MSD variance

The results show a significant effect of time interval for the initial 6-minute high reliability period of the trial (F(1,30) = 25.30, *p* <0.001). Participants' tracking performance in the second 3-minute period (Mean = 25.67, SD = 3.13) was significantly better than their performance during the first 3 minutes (Mean = 27.18, SD = 3.54), as shown in Figure 5.12. No other significant main effects or interaction effects were found.

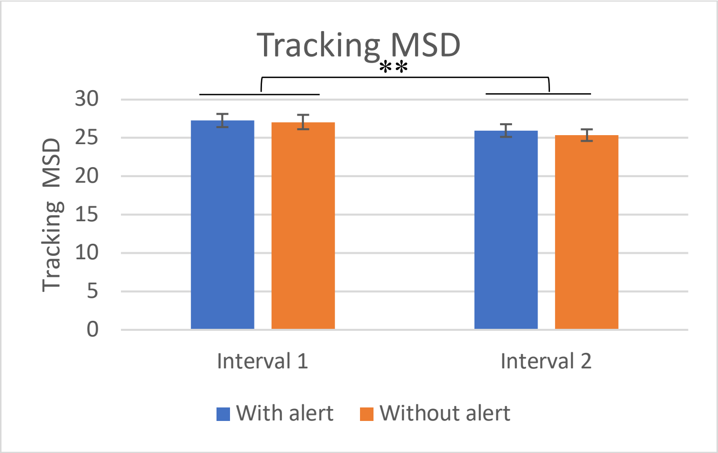


Figure 5.12. Tracking variance as a function of time intervals

Target detection performance

Accuracy and response time

No significant main effects or interaction effects were observed for accuracy and response time.

Overall performance

For the combined performance on the tracking and monitoring task, there was a significant effect of time interval (F(1,30) = 10.31, *p* = 0.003). Participants' overall performance during the second 3-minute period (Mean = 646.28, SD = 137.90) was significantly better than their performance during the first 3-minute period (Mean = 610.34, SD = 138.02), as shown in Figure 5.13. No other significant main or interaction effects were found.

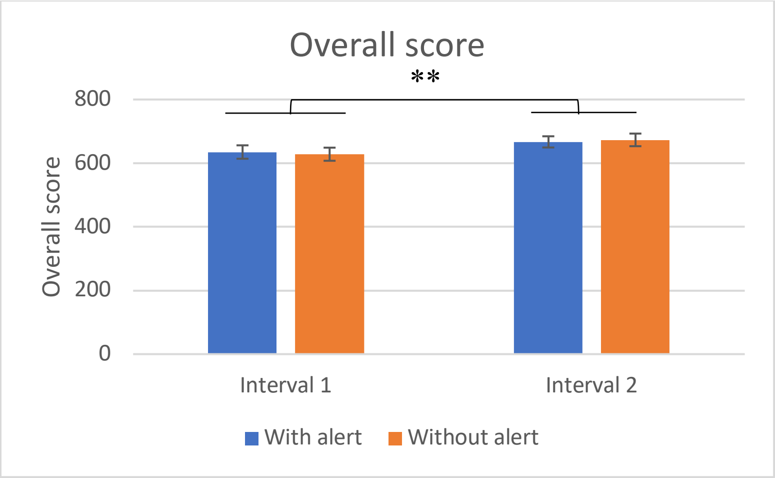


Figure 5.13. Overall score as a function of time intervals

Period 2: The three-minute period after the drop in system reliability

The three-minute period right after the reliability drop was examined to assess 1) whether the sudden and large change in system reliability resulted in a temporary miscalibration of trust, as reflected in the perceived reliability, subjective trust ratings, eye tracking metrics, and participants' reliance on the automation, and 2) whether participants in the audio alert group were able to adjust their perceived reliability, trust and behavior more quickly and appropriately, leading to better overall performance (expectation 3).

A linear mixed model analysis was conducted on each of the following dependent variables: perceived reliability difference, subjective trust ratings, two eye tracking metrics, reviews of automation assessments and performance on both tasks. Audio alert was entered as a fixed effect, and participant ID was entered as a random effect. There was no significant effect of audio alerts on any of these dependent variables.

Period 3: The nine-minute period of low system reliability

The nine-minute period (divided into 3 three-minute intervals) following the reliability drop was examined to determine if trust calibration improved over time for participants in the control group, leading to performance levels similar to those of participants in the audio alert group (expectation 4).

Perceived reliability

A 2(audio alert)*3(time interval) linear mixed model was performed on the perceived reliability calibration during the nine-minute low reliability period. Participant ID was entered as a random effect. The results reveal a significant effect of time interval (F(1,60) = 12.343, *p* < 0.001) (See Figure 5.14). Post hoc analyses with Bonferroni correction show that participants in both groups tended to overestimate the reliability of the automation early on, leading to a significantly smaller perceived reliability calibration (Mean =-0.772 , SD = 1.464) in interval 3 compared to both intervals 4 (Mean = 0.359 , SD = 1.541, *p* <0.001) and 5 (Mean =0.063 , SD = 1.038, *p* = 0.002). No effect of audio alerts and no interaction effects were observed.

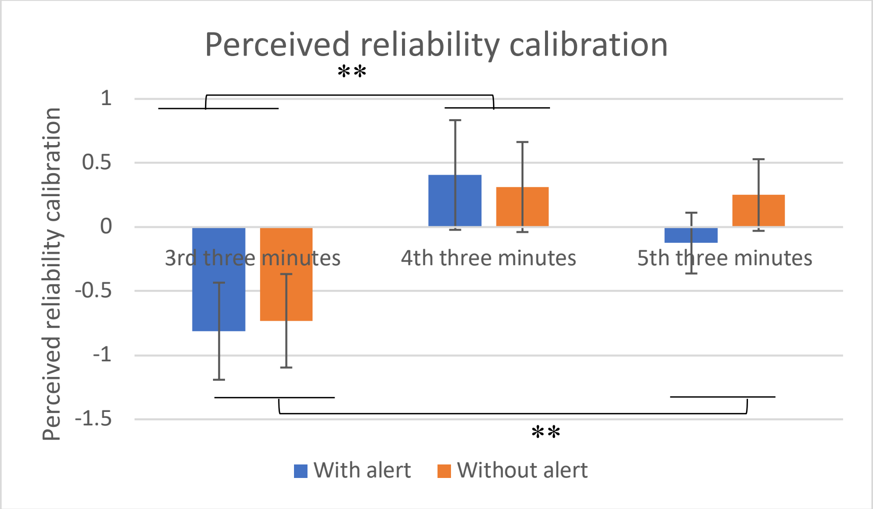


Figure 5.14. Time interval effects on perceived reliability calibration

Subjective trust ratings

Consistent with the findings for the perceived reliability, only time interval had a significant effect on participants' trust ratings over the 3 three-minute intervals. Specifically, participants' trust ratings were significantly lower for the 4th interval (Mean = 3.703, SD = 1.946, F(1,60) = 4.715, *p* =0.013), compared to their ratings (Mean = 4.672, SD = 2.010) in the 3rd interval, as shown in Figure 5.15. No significant effect of audio alerts or any interaction effects were found.

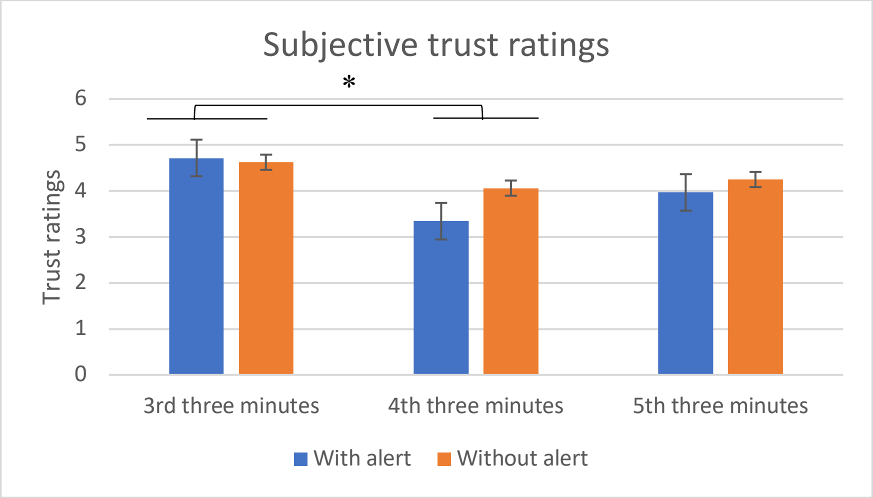


Figure 5.15. Time interval effects on subjective trust ratings

Eye tracking metrics

A 2 (audio alert) *3 (time interval) linear mixed model analysis was conducted on total fixation duration percentage and transition count. No significant main effects or interaction effects were observed.

Reviews of automation assessments

A 2 (audio alert) *3 (time interval) linear mixed model analysis was conducted on the reviews of automation assessments. No significant main effects or interaction effects were observed. Participants in both groups clicked quite frequently (Mean = 15.78, SD = 0.62) during the 9 minutes.

Tracking performance – MSD variance

The same analysis was conducted on tracking performance. No significant main effects or interaction effects were found.

Monitoring performance

Accuracy

There was a significant effect of time interval on accuracy for the 9-minute low-reliability period (F(1,33.58)=4.928, *p* =0.013)(see Figure 5.16). Participants' monitoring accuracy significantly improved over time, increasing from 72.81% (SD = 0.134) during the first 3 minutes after the reliability drop to 80% (SD=0.136) during the last three minutes of low reliability. No other significant main effects or interaction effects were observed.

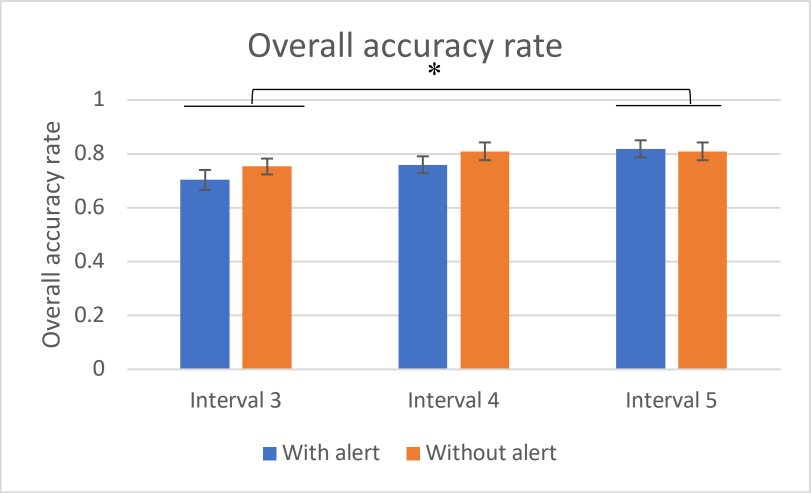


Figure 5.16. Time interval effects on overall accuracy rate

Response time and Overall performance

No significant main effects or interaction effects were observed for response time and overall performance.

Period 4: The three-minute period following system reliability recovery

The three-minute period right after system reliability had returned to its original high level was examined to determine if participants in the audio alert group showed faster and better trust calibration and more appropriate reliance on the automation than the control group. Performance was expected to be worse for the control group during this period due to a delay in their adjustment of trust and reliance (expectation 5).

Perceived reliability calibration

A linear mixed model analysis was performed on participants' perceived reliability calibration for the three-minute period following automation reliability recovery. No significant effect of audio alerts was observed.

Subjective trust ratings

The same analysis was conducted on subjective trust ratings. The analysis reveals a marginally significant effect of audio alerts on subjective trust ratings for the three-minute period following system reliability recovery (F (1,30) = 3.884, *p* = 0.058). Participants in the audio alert group expressed higher trust (Mean = 8.938, SD = 0.75) than participants in the control group (Mean = 8.219, SD = 1.251), as shown in Figure 5.17. No significant effect of audio alert was observed.

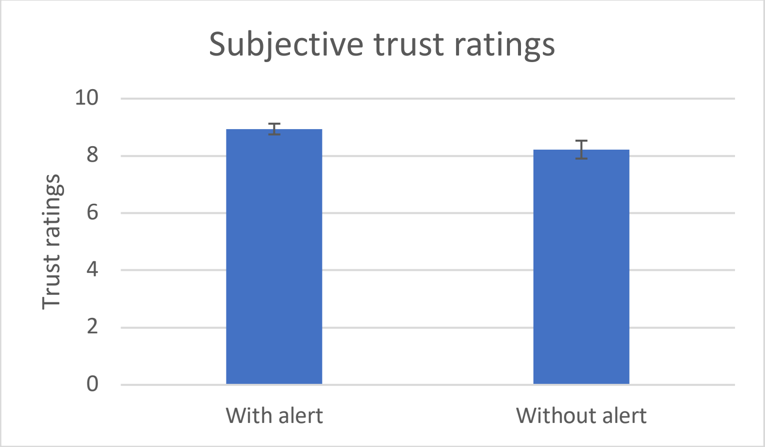


Figure 5.17. Marginal effects of audio alerts on subjective trust ratings

Eye tracking metrics

The same analysis was performed on the two eye tracking metrics. No significant difference was found between the control and the audio alert groups.

Reviews of automation assessments

A linear mixed model analysis was conducted with audio alert entered as a fixed effect and participant ID entered as a random effect. There was a marginally significant effect of audio alerts (F(1,30) = 3.927, *p* = 0.057). Participants in the control group reviewed the automation assessments more often (Mean = 7.25, SD=5.37) than participants in the audio alert group (Mean = 4.25, SD =2.79) (Figure 5.18).

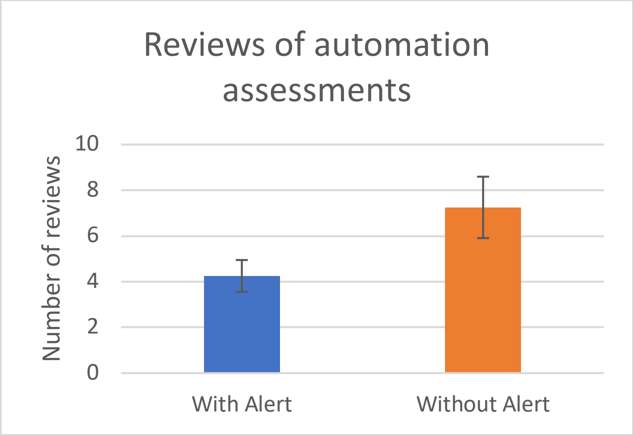


Figure 5.18. Marginal effects of audio alerts on number of reviews

Performance

A linear mixed model analysis was conducted on tracking variance, monitoring overall accuracy, monitoring response time and overall performance score. There was no significant effect of audio alerts on any of these performance outcomes.

Period 5: Time until trust recovery

'Time until trust recovery' is defined as the number of 3-minute periods it took for participants' trust to recover to a steady level after the system reliability had returned to its initial high level. A steady level is reached when 1) the trust rating returns to a steady (though not necessarily the initial) level and 2) the trust inference model indicates high trust for two consecutive periods. A linear mixed model was conducted on time taken to recover trust. The result shows that the audio alert significantly reduced the time it took for participants' trust to recover (F(1,30)=7.075, *p* =0.012) from 3 periods in the control group (SD = 1.60, without alert system) to 1.75 periods in the audio alert group (SD = 1), as shown in Figure 5.19.

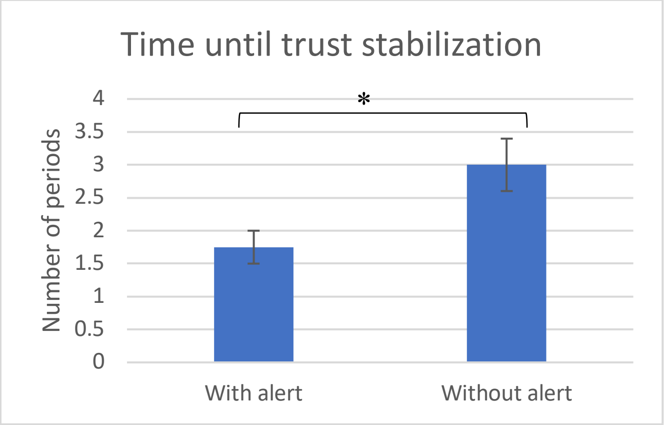


Figure 5.19. Effects of audio alerts on trust recovery

Additional analysis on model inference results and overall performance

To better understand why audio alerts did not appear to affect performance, an additional analysis was conducted on the model inference results and overall performance for the two groups of participants on a minute-by-minute basis. For participants in the control group, even though they did not receive any audio alerts, the trust inference model still worked in the background to infer whether a participant was in a high or low trust state during the past minute. A Chi-Square test was performed to compare the number of alerts actually presented to participants in the audio alert group and the number of alerts that would have been triggered for participants in the control group. The analysis revealed a significant difference in the number of alerts between the two groups for minute 5 (χ^2(1,N=32) = 6.788, *p* = 0.023), minute 7

(χ^2(1,N=32) = 7.575), *p* = 0.015) and minute 18 (χ^2(1,N=32) = 6, *p* = 0.037), as shown in Figure 20. The number of alerts would have been higher for the baseline condition for minutes 5 and 18 and higher for the audio alert group for the seventh minute of the trial.

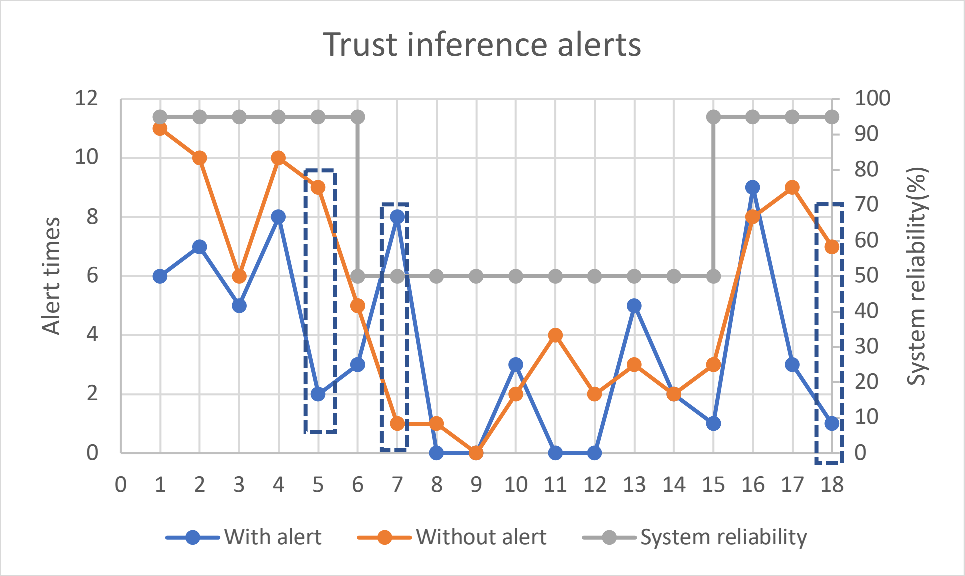


Figure 5.20. Alert number in each group for every 1 minute

Results summary

Table 5.2. Effects of audio alert and time intervals on perceived reliability, subjective trust ratings, eye movements, behavior and performance (NS: not significant)

Effects of audio alert and time intervals		**Subjective ratings**		**Eye tracking**		**Behavior**	**Performance**		
		Perceived reliability calibration	Subjective trust ratings	Total fixation duration PC	Transition count	Number of reviews	MSD	Accuracy	Overall score
First 6 min	Audio alert (Yes)	NS	NS	NS	NS	↓	NS	NS	NS
	Time interval ↑	↓	↑	↑	NS	↓	↓	NS	↑
3 min after the reliability drop	Audio alert (Yes)	NS	NS	NS	NS	NS	NS	NS	NS
9 min low reliability period	Audio alert (Yes)	NS	NS	NS	NS	NS	NS	NS	NS
	Time interval ↑	Worst trust mis-calibration in interval 3 compared to 4 and 5	Lowest trust rating in interval 4	NS	NS	NS	NS	Improved from interval 3 to 5	NS
3 min after the reliability recovery	Audio alert (Yes)	NS	Marginally higher trust in the audio alert group	NS	NS	↓	NS	NS	NS
Trust recovery	Audio alert (Yes)		Faster trust recovery						

Discussion

The final experiment in this line of research aimed to examine the effectiveness of using an audio alert for overcoming trust miscalibration and thus prevent inappropriate reliance on automation and resulting performance breakdowns. The results show that the audio alert did not affect participants' perceived reliability, and that trust ratings differed between the control and the alert group only during the trust recovery period when participants in the alert group were faster to recover trust in the automation. Participants in the alert group relied more heavily and appropriately on the highly reliable automation during the first 6 minutes of the trial, compared to the control group. The same difference in automation reliance between the two groups was observed during the first three minutes following the recovery of system reliability to its initial high level. The observed changes in trust ratings and automation reliance did not translate into significantly improved overall performance on the tracking and target detection tasks. The following sections will discuss these findings as well as the observed effects of time interval on the various dependent measures in more detail.

During the first 6 minutes of each trial, when system reliability was consistently high (95% reliable), perceived reliability and trust increased for participants in both the control and the audio alert group. This observation is consistent with previous research findings showing that participants' trust in the highly reliable systems increases gradually over time (Desai et al., 2013; Kraus et al., 2019; Muir & Moray, 1996). The audio alert did not lead to a more pronounced increase, likely due to a ceiling effect. Perceived reliability and trust in the automation were very high in both groups and could not increase by much. As trust increased, participants in both groups started to rely more on the automation to perform the target detection task for them, leading to significantly fewer reviews of automation assessments in the 2nd three-minute period.

This effect was more pronounced in the audio alert group. One likely explanation for this finding is that, even though all participants reported a high level of trust in the automation from the beginning, they initially still reviewed the UAV scene and automation assessments quite frequently. This discrepancy between participants' (high) subjective trust ratings and their (low) reliance on the system has been observed in earlier studies (Lu & Sarter, 2019a; Miller et al., 2016). In response to their disuse of the highly reliable automation, participants in the audio alert group (but not those in the control group) were presented with aural "center-center" commands instructing them to rely more heavily on the target detection system and focus on the tracking task instead. This promoted more appropriate trust in and reliance on the system.

Total fixation duration on the tracking task significantly increased over time for all participants as they relied more on the automation to perform the target detection task. Even though participants in the control group still continued to review automation assessments more often than those in the audio alert group, total fixation duration did not differ between the two groups, likely because the control group was aware of the high reliability of the automation and thus reviewed the target area only very briefly. Transition counts did not change significantly, probably due to large variance in the data.

The increased attention on the tracking task resulted in improved performance for both groups during the second 3-minute period. One possible reason why the audio group did not outperform the control group may be that the average response time to an automation assessment in this group was less than 3 seconds. Performance on the tracking task was calculated every 3 seconds. This means that, even if participants in the control group did review automation assessments more frequently, this did not necessarily hurt their tracking performance as long as they could keep the average MSD within the positive score range. The movement of the small circle in the tracking task was based on a series of random numbers, meaning that even if participants did not work on the tracking task, there was still a chance that the average MSD of the small circle was within the positive score range. Target detection performance in both groups remained unchanged, mainly due to a ceiling effect. Even when participants double-checked an automation assessment, they rarely disagreed with the system because of its high reliability (Hussein, Elsawah, & Abbass, 2019) and therefore the final target detection performance did not change.

During the three-minute period following a drop in system reliability, the control and audio alert groups did not differ with respect to any of the dependent measures. Participants in both groups were able to quickly detect the rather large change in system reliability and adjust their behavior. This may have been helped by the collection of baseline data for modeling purposes in advance of the experiment which did not take place in experiments 1 and 2. All participants experienced four baseline scenarios (50%, 65%, 80% and 95%) in a randomized order before they started the formal trial. Before each scenario, participants were informed of the actual capability of the automation to help them decide how much they should trust and rely on the automation. This process served as additional training and familiarization with the automation which has been shown to be an effective technique to help facilitate the trust calibration process (Masalonis, 2003; Tenhundfeld et al., 2019). This explanation is supported by the fact that the number of alerts that were actually triggered in the audio alert group was not significantly different from the number of alerts that would have been triggered in the control group after 2 minutes of low system reliability, meaning both groups were equally calibrated at that point. Interestingly, the number of alerts that would have been triggered during the first minute after the reliability drop was significantly lower in the control group, compared to the audio alert group. This unexpected result may be explained by the way the alert was designed. It was based on eye tracking data over the last one-minute period. Therefore, during the first minute after the reliability of the system dropped, some participants in the audio alert group may still have received audio alerts that carried over from the 6-minute high-reliability period, telling them to focus more on the tracking task.

Participants overestimated system reliability, and their subjective trust ratings were higher during the first 3 minutes of the 9-minute period of low system reliability. This confirms earlier findings (e.g., Lee and Moray, 1992) that trust adjustments lag behind changes in actual system reliability. Trust tends to reach its lowest point only after several performance breakdowns have been observed and then recovers gradually, even in the continued presence of system faults. The change in subjective ratings during the 9-minute period was not mirrored in participants' attention allocation and automation monitoring behavior. Both groups double-checked automation assessments nearly every time during the low reliability period. Audio alerts were rarely triggered and therefore the audio group did not benefit. While subjective ratings and behavior did not change, participants' performance on the automation monitoring task did improve (in terms of accuracy) over the course of the 9-minute low-reliability period. This is likely the result of their increased familiarity with the unreliable automation and increased scrutiny of automation decisions once they calibrated their trust in the system.

During the first 3 minutes after system reliability had recovered, perceived reliability did not differ significantly between the two groups; however, trust ratings for participants in the audio alert group were marginally higher than for the control group, which translated into less frequent reviews of automation assessments. This difference in behavior was not reflected in the eye metrics which, as mentioned earlier, may relate to the fact that participants in the control group had a tendency to agree very quickly with the automation when it was highly reliable.

With regard to trust recovery, the results show that audio alerts helped participants rebuild their trust in the automation in a more timely fashion. In this experiment, trust recovery was defined as 1) participants' trust ratings returned to a steady (though not necessarily the initial) level, and 2) the trust inference model indicated high trust twice in a row. While participants' trust ratings did not differ, the eye tracking based model output (criterion 2) for participants in the audio group showed higher level of trust and mirrored system reliability.

In conclusion, the results of this experiment indicate that the audio alert was partially successful in improving the trust calibration process, especially during trust recovery, and contributed to more appropriate reliance on automation, especially during the time period following a large system reliability drop. However, the alert did not improve participants' overall performance as expected, mainly because of two factors. First, all participants received extra exposure to various system reliability scenarios during the baseline data collection, which served as training and improved trust calibration in both groups. Second, the cost of unnecessarily checking on the automation when it was highly reliable was minimal; as a result, performance in the control group did not suffer. These factors should be addressed in the future research to better investigate the effectiveness of the audio alert system.

References

Clare, A. S., Cummings, M. L., & Repenning, N. P. (2015). Influencing Trust for Human–Automation Collaborative Scheduling of Multiple Unmanned Vehicles. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(7), 1208-1218.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). *Impact of robot failures and feedback on real-time trust.* Paper presented at the Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction.

Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission control of multiple unmanned aerial vehicles: A workload analysis. *Human factors, 47*(3), 479-487.

Graham, H., & Cummings, M. (2007). *Assessing the impact of auditory peripheral displays for UAV operators*. Retrieved from

Hussein, A., Elsawah, S., & Abbass, H. A. (2019). Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions. *Human factors*, 0018720819879273.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2019). The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human factors*, 0018720819853686.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153-184.

Lu, Y., & Sarter, N. (2019). Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability. *IEEE Transactions on Human-Machine Systems, 49*(6), 560-568.

Masalonis, A. J. (2003). *Effects of training operators on situation-specific automation reliability.* Paper presented at the SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483).

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*(4), 656-665.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors, 55*(3), 520-534.

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human factors, 57*(1), 34-47.

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). *Behavioral Measurement of Trust in Automation: The Trust Fall.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429-460.

Olsen, A. (2012). The Tobii I-VT fixation filter. *Tobii Technology*, 1-21.

Prinet, J. (2016). *Attentional Narrowing: Triggering, Detecting and Overcoming a Threat to Safety.*

Schwarz, N., Münkel, T., & Hippler, H. J. (1990). What determines a 'perspective'? Contrast effects as a function of the dimension tapped by preceding questions. *European Journal of Social Psychology, 20*(4), 357-361.

Tenhundfeld, N. L., de Visser, E. J., Haring, K. S., Ries, A. J., Finomore, V. S., & Tossell, C. C. (2019). Calibrating Trust in Automation Through Familiarity With the Autoparking Feature of a Tesla Model X. *Journal of Cognitive Engineering and Decision Making, 13*(4), 279-294.

Wiegmann, D. A. (2002). Agreeing with automated diagnostic aids: A study of users' concurrence strategies. *Human factors, 44*(1), 44-50.

Woods, D. D. (1995). The alarm problem and directed attention in dynamic fault management. *Ergonomics, 38*(11), 2371-2393.

Chapter 6

Summary and Conclusion

Trust calibration remains a major challenge for safe and efficient human-machine interaction. Poor trust calibration, i.e., “a lack of correspondence between a person’s trust in a system and the system’s actual capabilities (Lee & Moray, 1994)”, can lead to either misuse (overreliance on) or disuse (slow adoption or complete rejection) of automated systems (Parasuraman & Riley, 1997). The consequences of inappropriate reliance on advanced technologies can be catastrophic, especially in high-risk, safety-critical domains such as military operations, aviation and process control (e.g., nuclear power plants). Addressing this challenge requires a better understanding of, and support for the process of trust development and calibration – the focus of the proposed research. Earlier research in this area suffers from a number of limitations. First, most studies rely heavily on subjective ratings to measure trust (Hoff & Bashir, 2015). Such ratings are susceptible to biases and are too disruptive for real-world applications. Also, discrete ratings fail to capture the dynamics of trust, i.e., they do not provide insight into the continuous evolution and temporal variability of trust in response to factors such as variations in system reliability. In the absence of a real-time continuous measure of trust it is also not possible to detect the need for, and trigger interventions to overcome inappropriate trust levels and resulting misuse or disuse of technology. This may explain why, to date, most research tried to improve trust calibration in a top-down fashion through training (Bahner et al., 2008; Masalonis, 2003) and priming (Pop et al., 2015) in advance of operations.

To address these gaps and shortcomings, the goals of this dissertation research were to:

1) Develop an eye-tracking based technique to infer trust levels in real time and identify the eye tracking metrics that are best suited for tracing trust variations.
2) Identify how the magnitude and duration of variations in system reliability affect trust evolution and calibration
3) Develop and evaluate the effectiveness of a real-time intervention (an audio alert) for supporting trust calibration and promoting proper use of automation

To achieve these goals, three experiments were conducted in the context of an Unmanned Arial Vehicle (UAV) control simulation. Participants were required to perform two tasks in parallel: a tracking task and a target detection task, the latter with the assistance of an imperfectly reliable automated system. Experiment 1 used subjective trust ratings and eye movement data to examine variations in trust as a function of system reliability and priming. Three types of eye tracking metrics were calculated from raw eye movement data: temporal metrics (e.g., total fixation duration), spatial metrics (e.g., backtrack rate), and count metrics (e.g., transition count). System reliability had a significant effect on both subjective trust ratings and eye tracking metrics whereas priming affected eye movements only. The observed differences in eye movements for high- and low-reliability UAVs were closely associated with participants' subjective trust ratings, suggesting that eye tracking is a promising less intrusive technique for measuring trust in real time. Some differences between subjective trust ratings and the eye tracking data also highlight, however, that the two measures should be employed in a complementary fashion. For example, subjective ratings of trust differed significantly between high- and low-reliability UAVs independent of the priming condition. In contrast, one of the eye tracking metrics, total fixation duration, was affected by reliability variations only in the priming condition. Also, one of the eye tracking metrics, total fixation count, was sensitive to variations in system reliability over time, while subjective ratings were not. These differences suggest that subjective trust ratings mainly reveal participants' explicit trust levels, while eye tracking metrics reflect more about implicit trust levels (Burns, Mearns, & McGeorge, 2006).

After Experiment 1 confirmed the feasibility of using eye tracking to infer trust in real time, this technique was employed in Experiment 2 to assess the effects of the magnitude (small and large) and duration (short and long) of system reliability changes on the process of trust development and calibration. Subjective trust ratings showed that participants detected changes in reliability even when they were small and short. However, awareness of these changes did not result in behavioral (monitoring) changes, as indicated by the eye movement data. Trust miscalibration was observed for all combinations of magnitude and duration of reliability changes: participants underestimated system reliability when it was actually quite high (95%) and overestimated it when it was low (50% & 80%). The magnitude of variations in system reliability had a more significant impact on trust calibration and trust recovery than duration. Trust miscalibration was most severe, and trust recovery took the longest in case of large and long drops in system reliability. In those cases, accuracy on a target detection task was also lowest due to inappropriate reliance on the automation.

After establishing that trust variations could be detected in real time, and that participants experienced trust miscalibration in various circumstances, we next developed a machine learning technique to support the development of a countermeasure to trust miscalibration in the form of an audio alert. Two different modeling approaches were compared: 1) using eye tracking metrics from individual participants for "personalized" modeling, and 2) using eye tracking metrics from a group of participants for “generalized” modeling. For the latter approach, two types of input data were used: 1) raw eye movement data (gaze point coordinates) and 2) eye tracking metrics calculated from those raw data. Linear Regression (LR), k-Nearest Neighbor (kNN), Random Forest (RF), Multi-Layer Perceptron (MLP) and Convolutional Neural Networks (CNN) were compared to determine which classification method was most sensitive and reliable and thus best suited for triggering an audio alert to notify participants that they needed to adjust their trust and reliance on the automation. Individual-level modeling using kNN or RF was found to achieve the best performance (accuracy rate around 92%). After considering factors including modeling efficiency, modeling variance, and interpretability, kNN was picked over RF for inferring trust levels in real time.

Finally, Experiment 3 was conducted to evaluate the effectiveness of the audio alert system for supporting trust calibration in the worst case scenario (long and large reliability drop) identified in Experiment 2. The audio alert resulted in more appropriate trust (reflected by both subjective trust ratings and eye tracking metrics) and more appropriate reliance on the automation (as seen in behavioral measures such as the number of times participants reviewed automation recommendation). Compared to participants in the baseline condition (no audio alerts), participants who experienced audio alerts recovered their trust in the automation much faster once the system performed well again. However, this improved trust calibration did not translate into improved performance on either the target detection or the tracking task. This result could be explained by the fact that the time required to double-check decisions of the automation was not long enough to negatively affect tracking performance. In addition, results indicated that participants had a tendency to always agree with the automation when the automation was highly reliable, resulting in no significant target detection performance difference between the audio alert group and the control group.

Overall, the findings from this line of research were quite consistent across the three experiments. All three studies showed that low trust levels were associated with longer fixation duration, more fixation counts on the target detection task and more transition counts between tasks, suggesting the feasibility of using eye tracking to infer trust in an unobtrusive way. It is important to note that the selection of specific eye tracking metrics as indicators of trust will likely have to be tailored to task demands. For example, if an operator performs only one task, transition counts will not be useful for capturing trust variations. And if the task does not require operators to search actively, spatial metrics may not be informative, as shown in Experiment 2. In addition, the shape, size and location of AOIs need to be carefully chosen by the experimenter to find the proper balance between selectivity and sensitivity for a given task set.

To facilitate trust calibration, this line of research examined two promising methods: 1) priming to assist with trust formation in advance of operations, in a top-down fashion (Experiment 1) and 2) real-time audio alerts to support timely and appropriate adjustments of trust throughout task performance, in a bottom-up fashion (in Experiment 3). The results indicate that priming affected overall trust levels (though not as strongly as system reliability). According to an integrative model of long-term trust development in human-robot teams (de Visser et al., 2019), priming promotes trust calibration because it allows operators to familiarize themselves with how an automated system works and thus anticipate system breakdowns which might otherwise represent violations of trust and lead to a “fall from grace” (de Visser et al., 2016; Madhavan, Wiegmann, & Lacson, 2006). Past research suggests, however, that this effect may not be long-lasting (Clare, Cummings, and Repenning, 2015). Real-time audio alerts, on the other hand, mainly promoted faster and appropriate trust recovery following a system breakdown. This confirms earlier work showing that real-time countermeasures, such as alerts, help repair trust after the automation has made a mistake or behaved unexpectedly (Baker, Phillips, Ullman and Keebler, 2018). Taken together, these findings suggest that both methods are valuable but should be employed to address different aspects of trust miscalibration.

Intellectual merit and broad impact

Findings from this dissertation carry both conceptual, methodological and practical value. First, they contribute to the knowledge base in trust and reliance on modern automation technologies. The results enhance our understanding of the effects of variations in system reliability on subjective trust ratings, behavioral markers and performance outcomes and of the connections between these three types of measures. A much-needed eye tracking based technique for tracing trust development in real time was developed which complements existing discrete and more intrusive trust measurement techniques, such as subjective ratings. Also, an eye tracking-based algorithm for detecting divergence between system trustworthiness and operator trust was developed and used to trigger and evaluate the effectiveness of a candidate countermeasure to trust miscalibration in the form of an audio alert. From an applied perspective, the findings from this line of research will lead to more appropriate attitudes towards, and adoption of modern automation technologies in a variety of application domains such as aviation, driving, space operations and medicine. This will, in turn, help avoid unexpected performance breakdowns and the rejection of useful technologies, thus contributing to more efficient and safer operations.

Future work

Through a series of empirical studies, this dissertation improved our understanding of and addressed challenges for studying trust in automation. However, as with all research, the reported work involves some notable limitations.

First, the examination of the effects of duration of reliability changes was somewhat limited. Even the long duration changes lasted for nine minutes only. That may explain why this factor did not affect most of the trust-related variables in our experiments. Past research (Hoff & Bashir, 2015; Lee & See, 2004; Muir & Moray, 1996) has shown that operators' trust tends to be based largely on continuously observing and interacting with an automated system. Future work should therefore include longitudinal studies that allow operators to collaborate with the automation over more extended periods of time.

Second, the need for, and the effectiveness of the audio alerts should be investigated in the absence of training (in this case, during the baseline data collection) which may have enabled participants to develop a model of the automation that was sufficient for performance to reach a ceiling. Past research has shown that training is an effective means of promoting appropriate trust levels in automation (Tenhundfeld et al., 2019). It is therefore not clear to what extent the findings from Experiment 3 were due to bottom-up guidance (the audio alert) or top-down influences (the training/mental model) or a combination of both.

Third, as this was the first attempt to trace participants' trust evolution process in real time, the input for the trust inference model was quite simple (with only two eye tracking metrics). In order to build a more comprehensive trust inference model, other factors such as dispositional trust, situational trust and learned trust, should be carefully considered (Hoff & Bashir, 2015). For example, by integrating factors reflecting individual differences in dispositional trust (such as culture, age, gender), it is possible to develop a more robust and generalizable trust inference model, which can be specific to individuals. In terms of the techniques to measure trust, in addition to eye tracking metrics, other techniques such as physiological measures (Akash et al., 2018), subjective measures (Jian et al., 2000) and behavioral measures (Miller et al., 2016) can be used in combination to trace trust evolution over time in a more robust way.

Finally, a more comprehensive understanding of the relationships between trust calibration and performance outcomes is needed. To date, there is limited research on developing metrics that can quantify trust miscalibration and establish whether and how trust contributes to performance (Merritt et al., 2015). A better understanding of this relationship is needed as performance and safety are ultimately the main concerns for human factors practitioners working on human-automation collaboration.

References

Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an Understanding of Trust Repair in Human-Robot Interaction: Current Research and Future Directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *8*(4), 1-30.

Clare, A. S., Cummings, M. L., & Repenning, N. P. (2015). Influencing Trust for Human–Automation Collaborative Scheduling of Multiple Unmanned Vehicles. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 57*(7), 1208-1218.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331.

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.

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013). *Impact of robot failures and feedback on real-time trust.* Paper presented at the Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction.

Dixon, S. R., Wickens, C. D., & Chang, D. (2005). Mission control of multiple unmanned aerial vehicles: A workload analysis. *Human factors, 47*(3), 479-487.

Graham, H., & Cummings, M. (2007). *Assessing the impact of auditory peripheral displays for UAV operators*. Retrieved from

Hussein, A., Elsawah, S., & Abbass, H. A. (2019). Trust Mediating Reliability–Reliance Relationship in Supervisory Control of Human–Swarm Interactions. *Human factors*, 0018720819879273.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2019). The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Human factors*, 0018720819853686.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153-184.

Lu, Y., & Sarter, N. (2019). Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability. *IEEE Transactions on Human-Machine Systems, 49*(6), 560-568.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human factors*, *48*(2), 241-256.

Masalonis, A. J. (2003). *Effects of training operators on situation-specific automation reliability.* Paper presented at the SMC'03 Conference Proceedings. 2003 IEEE International Conference on Systems, Man and Cybernetics. Conference Theme-System Security and Assurance (Cat. No. 03CH37483).

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 48*(4), 656-665.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors, 55*(3), 520-534.

Merritt, S. M., Lee, D., Unnerstall, J. L., & Huber, K. (2015). Are well-calibrated users effective users? Associations between calibration of trust and performance on an automation-aided task. *Human factors, 57*(1), 34-47.

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., & Ju, W. (2016). *Behavioral Measurement of Trust in Automation: The Trust Fall.* Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429-460.

Neisser, U. (1967). *Cognitive Psychology* (New York: Appleton-Century-Crofts)

Nikolic, M. I., Orr, J. M., & Sarter, N. B. (2004). Why pilots miss the green box: How display context undermines attention capture. *The International Journal of Aviation Psychology*, *14*(1), 39-52.

Olsen, A. (2012). The Tobii I-VT fixation filter. *Tobii Technology*, 1-21.

Prinet, J. (2016). *Attentional Narrowing: Triggering, Detecting and Overcoming a Threat to Safety.*

APPENDICES

APPENDIX A Debriefing Form for Experiment 1

Debriefing Questionnaire

Principal Investigator: Yidu Lu

Department of Industrial and Operations Engineering

University of Michigan

1. Did your trust differ among the 6 UAVs?
   o yes o no Please explain.

2. What was your general strategy for monitoring the UAVs? For example, did you monitor **certain** UAVs more than others? If so, why? Did you regularly check all video feeds for targets, independent of whether they were highlighted or not?

3. Do you think the automation improved your ability to detect targets? In other words, did the automatic highlighting of video feeds from UAVs reaching a target area help you notice and identify targets? Was the highlighting ever detrimental? Please explain.

4. Was it difficult to distinguish between targets & non-targets?

   o yes o no   if yes, please explain what made this task difficult

5. Please rate your fatigue level after finishing all the tasks:

```
|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|

0    1    2    3    4    5    6    7    8    9         10
```

6. Did the information we provided regarding the different degrees of reliability of the 6 UAVs affect your monitoring behavior? If so, was this true especially early on, or did this effect persist throughout the experiment?

7. Please feel free to add any other comments about the automation assistance and the experiment in general.

APPENDIX B Debriefing Form for Experiment 2

Debriefing Questionnaire

Principal Investigator: Yidu Lu

Department of Industrial and Operations Engineering

University of Michigan

1. How would you rate the difficulty of the monitoring task?

|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|

0 1 2 3 4 5 6 7 8 9 10
very easy very difficult

2. How would you rate the difficulty of the tracking task?

|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|

0 1 2 3 4 5 6 7 8 9 10
very easy very difficult

3. How would you rate the difficulty of timesharing the above two tasks?

|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|

0 1 2 3 4 5 6 7 8 9 10
very easy very difficult

4. What was your strategy for performing both tasks (monitoring the UAVs and doing the tracking task) at the same time? If your strategy differed for the four scenarios, please describe your strategy separately for each scenario and explain how and why it differed.

5. Do you think the automation improved your ability to detect targets? In other words, did the automatic highlighting of video feeds from UAVs reaching a target area help you notice the targets? Was the highlighting ever detrimental? Please explain.

6. Was it difficult to distinguish between targets & non-targets?
   o yes o no   if yes, please explain what made this task difficult

7. Did the sound feedback make you feel confused? In other words, did it give you a clear idea of your performance and the automation performance? Please write the meaning of the sound feedback.

8. Please rate your fatigue level after finishing all the tasks:

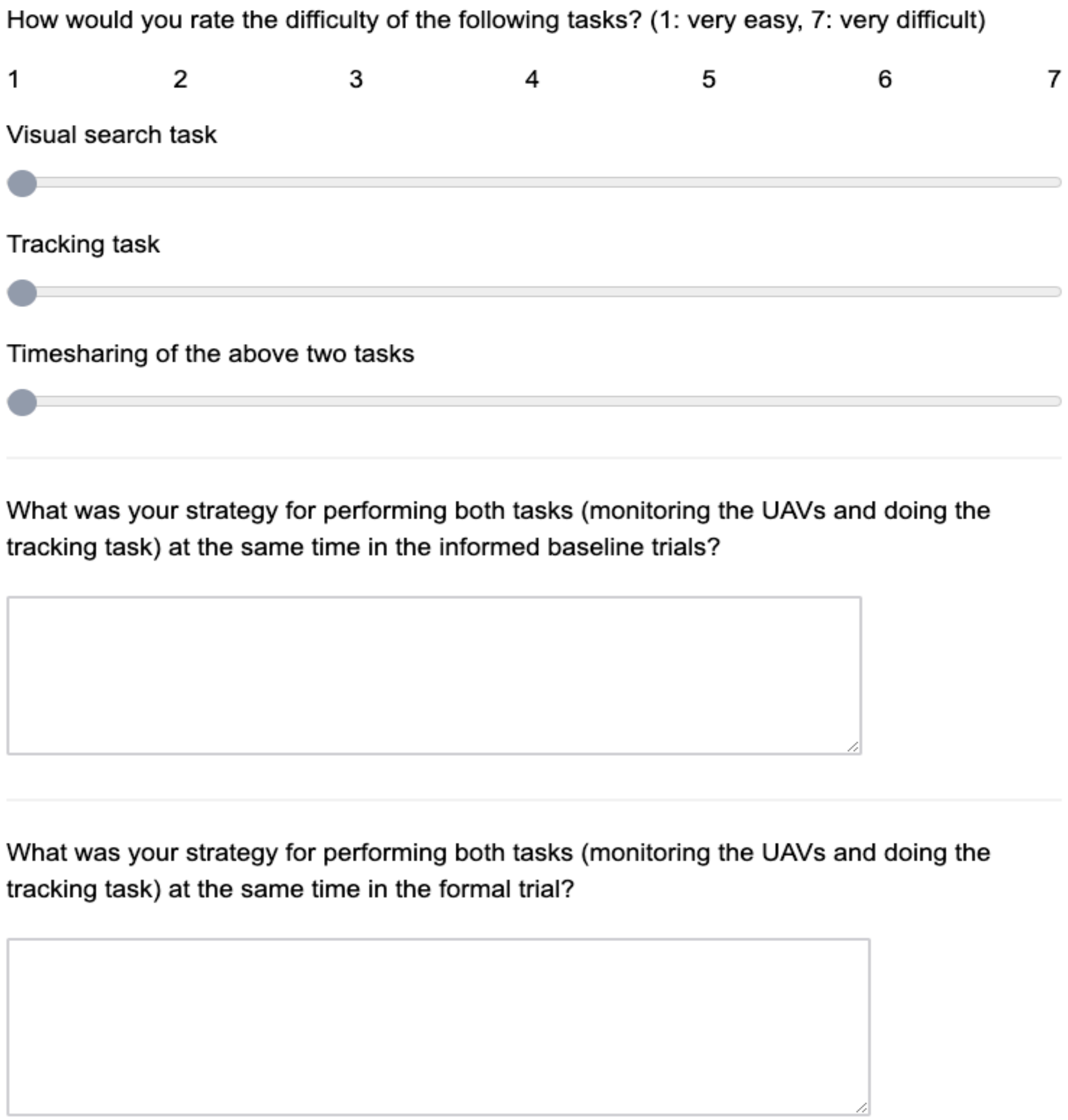
   |---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|

   0 1 2 3 4 5 6 7 8 9 10

   Not tired at all Extremely tired

9. Please feel free to add any other comments about the automation assistance and the experiment in general.

APPENDIX C Debriefing Form for Experiment 3



How would you rate the difficulty of the following tasks? (1: very easy, 7: very difficult)

1 2 3 4 5 6 7

Visual search task

●

Tracking task

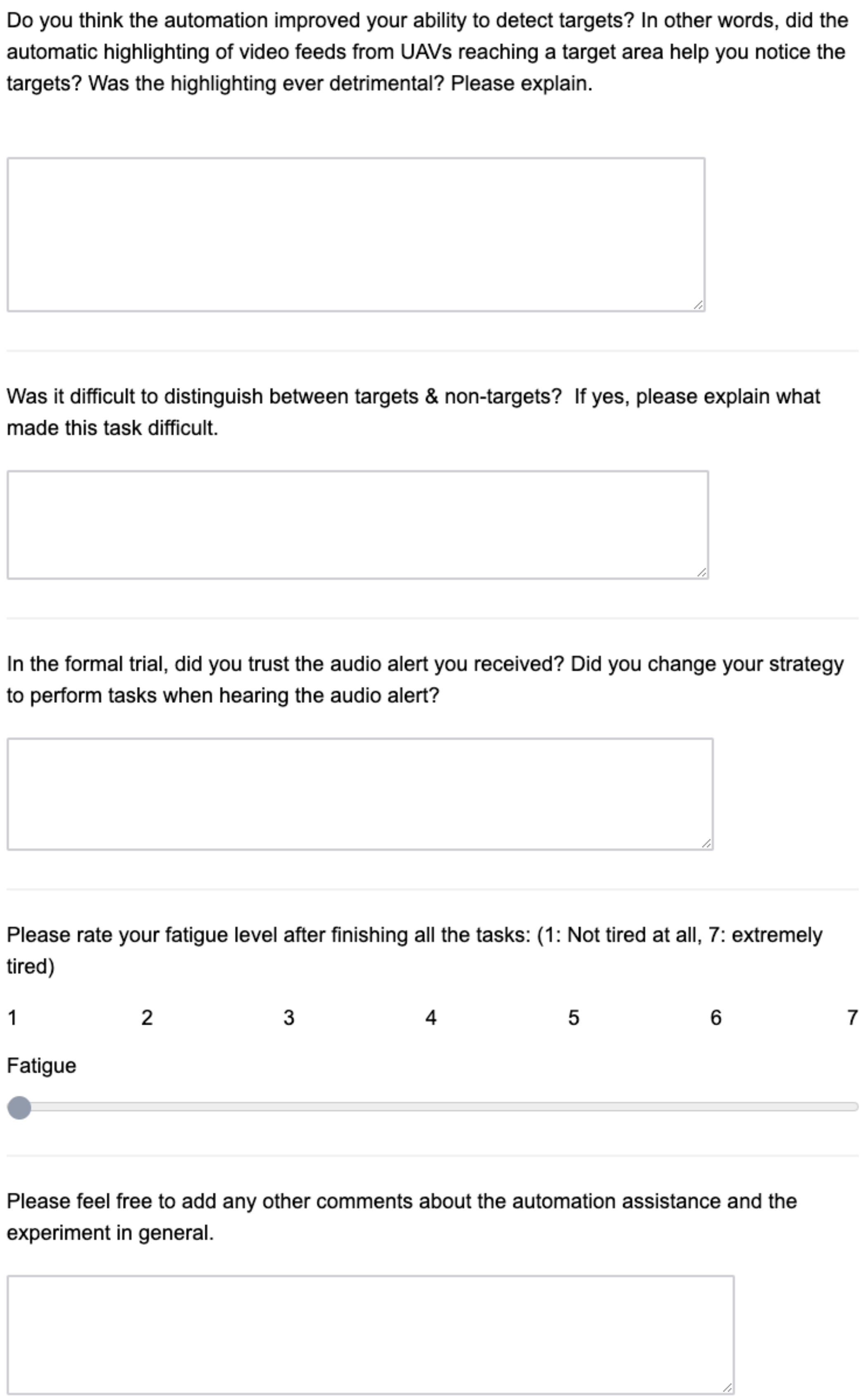
●

Timesharing of the above two tasks

●

What was your strategy for performing both tasks (monitoring the UAVs and doing the tracking task) at the same time in the informed baseline trials?

What was your strategy for performing both tasks (monitoring the UAVs and doing the tracking task) at the same time in the formal trial?



Do you think the automation improved your ability to detect targets? In other words, did the automatic highlighting of video feeds from UAVs reaching a target area help you notice the targets? Was the highlighting ever detrimental? Please explain.

Was it difficult to distinguish between targets & non-targets? If yes, please explain what made this task difficult.

In the formal trial, did you trust the audio alert you received? Did you change your strategy to perform tasks when hearing the audio alert?

Please rate your fatigue level after finishing all the tasks: (1: Not tired at all, 7: extremely tired)

1 2 3 4 5 6 7

Fatigue

Please feel free to add any other comments about the automation assistance and the experiment in general.