Supporting Trust Calibration and Attention Management in Human-Machine Teams Through Training and Real-Time Feedback on Estimated Performance

by

Kevin Marc Lieberman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Robotics)
in the University of Michigan
2022

Doctoral Committee:

Professor Ella M. Atkins, Co-Chair
Professor Nadine B. Sarter, Co-Chair
Associate Professor Leia A. Stirling
Assistant Professor Xi Jessie Yang

Kevin Marc Lieberman

klieberm@umich.edu

ORCID iD: 0000-0002-3136-2050

DEDICATION

To my family.

ACKNOWLEDGEMENTS

Thank you to the mentors, community, friends, and family who have supported my work and my growth throughout graduate school, cumulating in the completion of this dissertation.

I am very grateful to my advisor, Dr. Nadine Sarter for her mentorship, encouragement, and dedication to my success as a doctoral student and researcher. I also appreciate the support of my co-advisor, Dr. Ella Atkins, and committee members Dr. Leia Stirling and Dr. Xi Jessie Yang, who have offered meaningful feedback, their encouragement, and their time, to strengthen my research.

I would like to thank the Air Force Office of Scientific Research, the Department of Defense National Defense Science and Engineering Graduate (NDSEG) Fellowship, and a fellowship from the University of Michigan Robotics Institute for supporting my doctoral degree and this dissertation's line of research. In addition, grants from the Air Traffic Control Association (ATCA); University of Michigan College of Engineering and Rackham Graduate School; Human Factors and Ergonomics Society (HFES) at the University of Michigan; and the ROI Community Initiative of the Charles and Lynn Schusterman Family Philanthropies enabled me to attend and present at conferences and advance my research.

I have benefited from wonderful department staff who go out of their way to help students succeed. Thank you to Denise Edmund, Kimberly Mann, Dan Newman, and Damen Provost in the Robotics Institute, and Chris Konrad, Mint, Dr. Sheryl Ulin, Teresa Maldonado, Liz Michalski, Tina Picano Sroka, and Valerie Martin in the Industrial and Operations Engineering Department. They offered warmth and encouragement while removing logistical

roadblocks and creating opportunities. Li Morrow at the University of Michigan Institution Review Board (IRB) helped me navigate many challenges while conducting studies during the COVID-19 pandemic, and I appreciate her support in the completion of this research.

Thank you to my friends and labmates in The Human-automation Interaction and Cognition (THInC) Lab: Brandon Pitts, Julie Prinet, Yuzhi Wan, Yidu Lu, Karanvir Panesar, Robert Thomas, Akshay Bhardwaj, and Hannah Baez. They have offered essential encouragement and have been key thought partners in both the development of my work and in my development as a human factors researcher. This dissertation's studies would not have been possible without the THInC Lab's excellent programmers and research assistants, including Kevin Zhao, Aditya Mannari, Grace Miller, Vikram Mathew, William Hampton, Derek Witcpalek, Tejas Harith, Hanna Dong, and Chris Lee.

When I led the University of Michigan Jewish graduate student community and the Michigan Israel Engineering Trek, I would frequently find myself in situations that were high risk, high tempo, and complex. I grew significantly as a leader, thanks to conversations with Tilly Shemer, Or Shemer, Rav Lisa Stella, Naomi Solomon, Netanella Rafael, Avi Wolf, Ayal Beer, Benji Donitz, Eli Goldweber, Liz Livingston, Sheira Cohen, Joelle Abramowitz, Leah Josephson, Jordan Fruchtman, Ehud Har-Even, Meira Chefitz, Katie Forsythe, Hannah Brady, Benny Fisher, and Verena Klein. They were thought partners, collaborators, and friends, who also helped me reflect on how we could apply findings from the human factors literature to better lead and serve our communities. I am also grateful to Dr. Tiffany Ng and Dr. Pamela Ruiter-Feenstra for introducing me to the carillon and being role models for how caillonists can lift up their communities.

Thank you to Dr. Devendra Garg for introducing me to research, Katrina Wisdom for her friendship and always offering sound graduate school advice, Jane Christenson for sparking an interest in studying human behavior, and past colleagues at Metron Aviation for inspiring me to study human factors.

And finally, thank you to my family. I love you very much and I'm looking forward to coming home.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Trust, the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability (Lee & See, 2004), plays a critical role in supervisory control and human-machine teaming. Poor trust calibration, i.e., a lack of correspondence between a person's trust in a system and its actual capabilities, leads to inappropriate reliance on, or rejection of the technology. Trust also affects attention management and monitoring of increasingly autonomous systems. Overtrust results in excessive neglect time (the time the machine agent operates without human intervention) while distrust makes operators spend too much time supervising a system at the cost of performing other tasks.

To address these challenges, this research examined how training and real-time information about system confidence can support trust calibration and effective monitoring of modern technologies. Specifically, the aims of this research were (1) to compare the effectiveness of active, experiential training with more traditional forms of instruction on mental model development, trust resolution (i.e. the ability to distinguish contexts when a machine can be trusted versus when it requires close supervision), and attention management (experiment 1), and (2) to assess how various visual and auditory representations of a machine's confidence in its own ability (experiments 2 and 3) and the framing of a machine's estimated accuracy as confidence or uncertainty (experiment 3) affect trust specificity (i.e. shifts in trust based on incremental variations in machine capability over time), monitoring, and reliance on technology.

The research was conducted in the context of supervisory control of multiple unmanned aerial vehicles (UAVs).

The first, longitudinal study showed that participants who received experiential training had the fewest gaps in their mental model of the multi-UAV system, compared to participants who received more traditional training methods. They appropriately lowered their trust and monitored a UAV's health more closely when its environment reduced the UAV's capabilities. Findings from the second and third studies demonstrated that real-time feedback on a machine's estimated accuracy facilitates trust specificity and effective monitoring. Specifically, the second study compared visual and auditory representations of system confidence. It showed that the choice of display depends on the intended domain of use. Auditory confidence displays are preferable to visual indications in environments that suffer from visual data overload as the former avoid resource competition and support time sharing. The third study compared two different visual representations (hue- versus salience-based) of system confidence and examined the impact of framing a machine's estimated accuracy as confidence or uncertainty. Indicating a machine's uncertainty (rather than confidence) in its performance led to closer monitoring of UAVs and smaller trust decrements when the machine's estimated accuracy was low. Also, participants were better able to distinguish between levels of confidence and uncertainty with a hue-based representation that employed a familiar color scheme (red-yellow-green), compared with a monochrome salience-based representation.

At a conceptual level, this research adds to the knowledge base in trust, transparency, and attention management related to supervisory control and human-machine teaming in high tempo, complex environments. This line of research also makes significant contributions to the development and validation of subjective and eyetracking-based methods for assessing trust in

technology. Finally, from an applied perspective, the findings can inform the design of training and interfaces to support the safe adoption and operation of human-machine systems in a wide range of safety-critical domains.

Chapter 1

Introduction

The introduction of robotic technologies to complex, high tempo, and high risk domains has been greeted with much enthusiasm. However, the acceptance and utilization of these systems have been hindered by the effects of poor trust calibration, i.e., a poor mapping of operators' trust to the actual system capabilities. Trust has been defined in the context of human factors as the "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004). An incorrect mapping can lead to a system being "overtrusted," i.e., in trust exceeding its capabilities and performance marked by misuse. Conversely, a system may be distrusted if trust is below the system's actual capabilities, contributing to disuse of the technology (Parasuraman & Riley, 1997). "Trust resolution" and "trust specificity" further describe the mapping between an operator's trust level and a system's capabilities. A person with high trust resolution can distinguish the contexts when a machine can be trusted versus when it requires close supervision, whereas trust specificity refers to moment-to-moment adjustments in trust as a machine's capabilities fluctuate temporally (Lee & See, 2004). Trust not only affects the adoption of and reliance on technology; it also impacts attention management. Overtrust can lead to an operator failing to monitor and supervise a system, while distrust may result in a person overallocating their attention resources to a system at the cost of neglecting other tasks (Hergeth, Lorenz, Vilimek, & Krems, 2016; Lu & Sarter, 2019; Muir & Moray, 1996). These breakdowns in resource allocation and attention control continue to represent a major challenge for creating effective human-machine systems –

the ultimate goal in many application domains (such as UAV/UAS control) where currently 2-3 operators are needed to manage a single machine agent (Chen, Barnes, & Harper-Sciarini, 2011; Murphy & Burke, 2010).

To date, research has examined trust and attention management separately; in contrast, this line of research focuses on how trust calibration and attention management interact with each other. A better understanding of the mutual shaping of the two phenomena is important to support the safe adoption of human-machine systems where machine agents outnumber human supervisory controllers. Examples of such systems include Department of Defense concepts for manned-unmanned teaming (Mayer, 2020) and commercial drone-based transportation systems for people and cargo, both of which task an operator with supervising multiple robots simultaneously (Federal Aviation Administration, 2020). Single-operator multi-agent (SOMA) operations depend on proper trust calibration so that the human operator can divide their limited attention resources as effectively as possible and notice or anticipate when an intervention might be required.

This research project examines how the interaction between trust calibration and attention management evolves unaided over a prolonged period of time and how it can be shaped through design and training in the interest of improved collaboration and joint system performance. To this end, candidate approaches to training and mental model development, as well as interface designs that support system transparency and attention guidance, were developed and assessed in the context of a multi-UAV control simulation. Eye tracking, as well as traditional trust measures such as subjective ratings and performance measures, were used collectively to evaluate the efficacy of these interventions. The specific objectives of this research project were:

- To investigate and compare the impact of active, experiential training with more traditional forms of training on a person's mental model development, trust resolution, attention management, and joint system performance.
- To assess how various visual and multimodal representations of a machine's confidence in its own ability, and the framing of a machine's estimated accuracy as confidence or uncertainty, shape trust specificity and supports attention management and joint system performance.

This first chapter introduces the concept of trust in human supervisory control. The chapter then describes how candidate approaches to operator training and representations of a machine's confidence or uncertainty in its accuracy and abilities may support trust calibration, attention management, and joint system performance.

Trust in Human Supervisory Control

The Machines in Human-Machine Teams

There has been a growing conversation around whether advanced technologies such as artificial intelligence warrant a new approach to trust in human-machine systems. To better understand trust between humans and machines, such as robots, automation, and increasingly autonomous systems, it can be beneficial to first consider the structure of their relationship in safety critical contexts.

Definitions (and consequently boundaries) of what constitutes a robot continue to be debated to today. As noted by Goodrich and Schultz (2007), many cultures, from ancient to modern times, have envisioned artificial beings that were created by humans, to act like humans, and serve humans. Jordan (2016) reviews a variety of definitions that have been used to describe

robots, and he observes that many definitions describe robots as being "human-like." Alternatively, some definitions opt to define robots based on the tasks they perform (e.g. a machine that senses, thinks, and acts) (Jordan, 2016, p. 27). Commenting on the lack of consensus with regards to the definition of a robot, Jordan (2016, p. 4) quotes Bernard Roth, who bridges the gap between the human-qualities and the task-based definition of robots:

> My view is that the notion of a robot has to do with which activities are, at a given time, associated with people and which are associated with machines. If a machine suddenly becomes able to do what we normally associate with people, the machine can be upgraded in classification and classified as a robot. After a while, people get used to the activity being done by machines, and the devices get downgraded from "robot" to "machine" (B. Roth, 2008).

A different type of machine, automation, is being conceptualized by the human factors community as a system that performs tasks previously performed by people (e.g. Parasuraman and Riley, 1997). This characterization of automation is similar to Roth's definition of a robot. Finally, the term "autonomy" is commonly used to describe systems that are non-deterministic, adaptable (National Academcy of Engineering, 2014), and can achieve a goal without its internal processes being explicitly defined (Hager, Rus, Kumar, & Christensen, 2015). From a technical perspective, the juxtaposition of autonomy and automation is to highlight that automation is bound to fixed rules and processes for achieving a goal, whereas autonomous systems can both learn and extrapolate patterns in its environment to guide its behavior. Like automation, autonomy also suggests a design objective to transfer tasks from humans to machines, such as "driverless" cars and "unmanned" aerial vehicles.

Some researchers worry that referring to machines as "autonomous" misleads the general public about their capabilities and inflates the perception that the machines do not need human support. The recent National Academies of Engineering report on *Human-AI Teaming* (National Academies of Sciences Engineering and Medicine, 2021) opted to use the term "human-AI teams" rather than "human-autonomy teams," as it was felt that autonomy in the context of humans and machines would imply that they both were acting independently (i.e. "autonomously") from each other. While the report's committee acknowledged that the metaphor of a "team" to describe the relationship between a human and a machine might lead to expectations that a machine will perform similarly to a human, they felt that the metaphor was suitable since it emphasizes the need to consider the interactions and coordination between all "team members" that is required to support collective performance.

The Persistence of Humans in Human-Machine Systems

Robots, automation, and increasingly autonomous systems are frequently designed to perform tasks that people find undesirable or cannot perform with the same precision and reliability (Parasuraman, Sheridan, & Wickens, 2000). Machines also have been developed to perform tasks that require higher levels of control, planning, and cognitive decision making. However, these technologies have many times failed to achieve their goal to fully remove the human from the system (Bainbridge, 1983; Parasuraman & Riley, 1997).

Robots, automation, artificial intelligence agents, and other machines that sense and act can be limited in their plans and decision making because they operate in dynamic, partially-observable, complex, and unanticipated environments (Mason, 2012; Parasuraman & Riley, 1997; Russell & Norvig, 2009; Woods, Tittle, Feil, & Roesler, 2004). When these machines fall short of operational expectations, humans are re-introduced to the system as supervisors that

monitor the system and occasionally intervene when necessary (Sheridan & Parasuraman, 2005; Sheridan & Verplank, 1978). People are assigned the role of supervisory controller because they are perceived to be more resilient and adaptable than machines; also, humans can better ideate new solutions to problems when a machine is operated in unfamiliar contexts (Parasuraman & Riley, 1997).

However, this introduces the paradox that systems initially designed to "automate out" the human still end up requiring human involvement (Bibby et al., 1975) but now operators are typically not provided sufficient support to perform their tasks of monitoring and intervening (Bainbridge, 1983). Designers who approach a system with the goal of minimizing the role of a human operator many times do not recognize the increased demands for coordination and interaction required between the human and the machine (Bainbridge, 1983; Parasuraman & Riley, 1997; Parasuraman et al., 2000; Sarter, Woods, & Billings, 1997; Woods et al., 2004). In short, human supervisory control does not remove a person from the system but changes the relationship and interaction between a person and technology, an often overlooked aspect in system design (Dekker & Woods, 2018).

The issue becomes all the more problematic in operational contexts that place high attention demands on the human supervisor, such as single-operator multi-agent systems. Wiener (1989) is credited with coining the related term "clumsy automation," which refers to automation that performs at its best for simple, routine tasks that would require little cognitive effort from the human operator. However, when a situation arises that imposes considerable mental workload and attention demands on the supervisor, the automation adds to the challenge as it requires communication and coordination, and creates new pathways to complete tasks (Dekker & Woods, 2018; Sarter et al., 1997). Automation that is brittle (Roth, Bennett, & Woods, 1987),

meaning its performance can dramatically degrade outside of anticipated, nominal contexts can further complicate supervisory control. Automation that is both clumsy and brittle leads to the unfortunate predicament that it helps the operator manage tasks that they could easily perform on their own (without the automation's support) but in contexts where the operator could use its help, the automation is useless or even adds to the operator's cognitive load.

In order for humans and machines to work effectively and efficiently as a team, an operator needs to make decisions about when to trust and rely on technology. Trust calibration is critical for the proper delegation of tasks, it affects how closely the human monitors the machine's performance, and ultimately determines the joint performance of the human-machine system.

Trust and the Dynamic Nature of a Machine's Capabilities

A human supervisor's task of choosing when to rely upon a machine and how to monitor its performance is not trivial. Operators in human-machine systems use trust to cope with uncertainty and determine appropriate reliance when total comprehension of a machine's situation and behavior may not be achievable (Lee & See, 2004). Trust must be calibrated so that a person's trust matches a machine's capabilities (Muir, 1987) or else the system risks eventual misuse or disuse (Parasuraman & Riley, 1997).

Two important aspects of trust are its resolution and its specificity. Trust resolution describes the mapping between an operator's trust level and the range of a system's capabilities (Cohen, Parasuraman, & Freeman, 1998; Lee & See, 2004). In other words, a person with high trust resolution can distinguish between the contexts when a machine can be trusted versus when it requires close supervision. Trust specificity, on the other hand, includes both "temporal specificity" which describes temporally-aligned shifts in trust based on incremental variations in

machine capability, and "functional specificity," which refers to the correspondence of trust to the capabilities of subsystems (Lee & See, 2004). In cases of low functional trust specificity, localized capability decrements can lead to a "spread" of distrust to other functions in the system (Muir & Moray, 1996).

A wide range of human-related, machine-related and environmental factors have been shown to contribute to the development of trust (Hancock et al., 2011; Schaefer, Chen, Szalma, & Hancock, 2016). Most research to date has focused on machine-related factors, such as the agent's competence (Chiou & Lee, 2021), dependability and reliability (Hoff & Bashir, 2015; Lee & See, 2004), and consistency and predictability (Feltovich, Bradshaw, Clancey, & Johnson, 2007; Marble, Bruemmer, Few, & Dudenhoeffer, 2004). Less is known about the impact of human-related factors (such as expertise and attentional), environmental factors (such as multitasking), and the coordination required to support trust.

The sequence of changes in a machine's capability also impacts trust. Trust calibration develops, in part, during initial training for a new technology. It then continues to evolve over time, based on operational experience with the system. Few studies have examined this long-term development of trust in human-machine systems. They have shown that it is easier to lose trust in a system than to regain it, and that events contributing to a loss of trust are more impactful than events that increase trust. For example, a decision aid providing invalid recommendations was found to experience a loss of trust that was larger than the degree of trust it regained with valid recommendations (Yang, Wickens, & Hölttä-Otto, 2016). The timing of poor system performance appears to affect the trust development process also. Robot systems that exhibited low reliability early in experiments have caused greater losses of trust than when low reliability periods arose later in experimental runs (Desai et al., 2012). Therefore, it is

important to evaluate how a person's trust grows, decays, and evolves over longer durations of time (and over several sessions of interaction) as a member of a human-machine team in dynamic environments.

Mitigation Strategies for Trust Miscalibration

Training

A promising way to mitigate breakdowns in trust calibration and attention control is to support top-down information processing. Top-down processing refers to the voluntary allocation of attention based on goals, knowledge, experience, and expectations (Lee, Wickens, Liu, & Boyle, 2017). For example, by facilitating the development of a mental model of a system through training and operational experience, an operator can be aided in contextualizing machine behavior, anticipating how a machine will react to changes and events in its environment, and appropriately allocating their attention when monitoring robot team members.

Developing and implementing effective training methods to support the formation of mental models continues to be a major challenge for manned and unmanned aviation operations (see Murphy and Shields, 2012; United States Government Accountability Office, 2017, 2020). Two main reasons for this problem are the time available to train operators and current training approaches and content. Operators in many domains are taught primarily "how to operate the system" (i.e., how to use its interface, interpret video, or control the system) (Goodrich & Schultz, 2007), rather than "how the system operates" (i.e., the system's underlying logic, capabilities, and limitations) (e.g., Sarter et al., 1997; Strauch, 2017).

The study presented in Chapter 2 considers three training methods to prepare operators for the task of conducting multi-UAV control during reconnaissance missions. The first method

models traditional approaches to training, with a participant receiving instruction on "how to operate the system," i.e., on actions required to perform a task or achieve a given goal. A second training method supplements this training by instructing a participant "how the system works" based on the review of a PowerPoint presentation. For example, a participant receiving this supplemental training learns that a UAV's target identification algorithms recognize tanks more reliably than people because tanks operate at high temperatures and can therefore be seen more clearly using a thermal camera. Finally, in a third condition, participants obtain the identical, supplemental information but they do so through experiential learning.

Experiential learning provides trainees the opportunity to actively explore the system. One benefit of experiential learning is that it fosters firsthand experience in a range of contexts, leading to better memory association and better transfer to the actual performance domain (Molesworth, 2005). Additionally, the knowledge acquired through experiential activities is more likely to be activated and recalled when needed, compared to the same knowledge gained through passive learning and rote memorization (Sarter & Woods, 1997; Strater et al., 2004) which results in "inert" knowledge. By providing guided exploration, error management training, and critical reflection across a variety of situations (Bell & Kozlowski, 2008; McDermott, Gronowski, Carolan, & Fisher, 2013), experiential learning supports improved trust resolution.

Specifically, in this study, participants experienced a range of prototypical and off-nominal operational scenarios to help them build up experience with the system, observed its performance across a range of circumstances, and learned to recognize situations that are likely to challenge the system's capabilities. Acquiring this knowledge was expected to help operators make better decisions about when to rely on the unmanned system and focus on other

tasks/vehicles, versus when to carefully monitor its performance and action selections and intervene when necessary.

Furthermore, the impact of a participant's training on trust resolution was studied over the course of four weeks. Earlier work in the area of human-automation has shown that trust fluctuates significantly following errors made by systems that are less than 100% reliable (e.g., Lee & Moray, 1992; Muir & Moray, 1996), and the timing of when robot failures or unexpected behavior are encountered significantly impacts changes in trust and trust recovery (Desai, Kaniarasu, Medvedev, Steinfeld, & Yanco, 2013). However, most research to date has focused on the (assessment of the) momentary state of trust (Hancock et al., 2011), rather than its evolution over time. Longitudinal studies of trust in human-machine interaction are urgently needed to explore a range of issues related to trust development and calibration.

Visual Representations of Confidence and Uncertainty

In addition to training, feedback and representations of a system's inner logic can also support trust calibration (Gao & Lee, 2006; Hoff & Bashir, 2015; Schaefer, Straub, Chen, Putney, & Evans, 2016; Sheridan, 1989). For example, a mismatch between expected and actual robot behavior can result from uncertainty. Uncertainties in a human-machine system may be caused by sensing processes, computations and data fusion, and the way information is represented to an operator (Chung & Wark, 2016). In order to mitigate potential breakdowns in trust and performance due to uncertainties present in a system, McGuirl and Sarter (2006) were able to show that pilots using a decision support tool for detecting and handling in-flight icing events experienced fewer icing-related stalls if the automated tool provided continuously updated information about its confidence in its own performance. Seong and Bisantz (2008) similarly demonstrated in an aircraft identification task that meta-information, i.e., feedback qualifying the

estimated accuracy of a decision aid's recommendations, significantly increased performance and better supported trust calibration.

Findings from the few studies that examined the effectiveness of presenting confidence or uncertainty information have highlighted that the specific design and implementation of this information is critical (Du, Huang, & Yang, 2020; Helldin, Falkman, Riveiro, & Davidsson, 2013; Mercado et al., 2016; Seong & Bisantz, 2008; Sorkin, Kantowitz, & Kantowitz, 1988; Stowers et al., 2020; Wiczorek & Manzey, 2014; Zirk, Wiczorek, & Manzey, 2020). Bisantz (2013) and Chung and Wark (2016) provide an extensive review of visualization techniques for conveying uncertainty, primarily in contexts without high attention demands. Their review highlights that further research is needed to better understand how representations of uncertainty may impact trust and attention management – one of the questions addressed in the present project. Another important question relates to the effects of framing information about the system in terms of uncertainty about, or confidence in, its own performance and abilities. For example, framing a decision's outcome in terms of risks and loses, rather than gains, may cause a person to overestimate the risk, and prompt a person to be risk averse by making a decision based on an inaccurate perception the consequences (Kahneman & Tversky, 1979; Sheridan, 2008; Tversky & Kahneman, 1986, 1992).

To date, a major research focus in the design of confidence and uncertainty (visual) representations has been whether there is a natural mapping between the degree of uncertainty/confidence and the visual properties of an uncertainty representation. For example, blurriness and low levels of brightness and color saturation have been found to effectively represent high uncertainty (Bisantz, 2013; Bisantz et al., 2009). However, Bisantz et al. (2009) concluded that the mapping between color saturation and uncertainty may depend on the task. In

their study, participants were given sets of four colors that had the same hue but varied in saturation. Participants were presented with a map that needed to convey *highly certain information* and were tasked with assigning the sets of four colors to regions of the map that had been labeled with varying degrees of certainty. This was followed by a similar task to support the design of a map that needed to show regions with *less certain information*; participants assigned the sets of four colors with varied saturation to regions of the map that also had been labeled with varying degrees of certainty. Trials were completed with maps of two different backgrounds; a background map with a very saturated color (that matched the hue of the color set) or a background map with a neutral color (e.g. white). The study found that participants would assign colors such that the most relevant information would have the greatest contrast with the background. If a map was intended to identify which geospatial regions had *more* certain information, participants would assign colors that had the greatest contrast with the background to those areas. If a map was intended to identify regions with *less* certain information, participants would also assign colors that had the greatest contrast with the background to those areas. These findings suggest that meta-information should be designed such that the information that is most relevant to the task should be the most visually salient.

The findings discovered through the color ranking tasks align with the SEEV model (Wickens, Goh, Helleberg, Horrey, & Talleur, 2003), which proposes that the salience and value of information (in addition to expectancy and effort) impacts visual scanning and attention allocation. However, there remains a need to empirically assess the impact of the salience of confidence and uncertainty information on attention management, as well as trust calibration and performance, in an operational context. Most of the studies exploring the effectiveness of natural mappings to uncertainty were conducted in the absence of time pressure; the ability of an

operator to assess system confidence or uncertainty at-a-glance is thus unknown. Finally, there have been few studies that have assessed how inaccurate confidence information (such as a robot indicating that it has high confidence in its abilities during periods of low reliability) impacts operator trust and overall system performance. The study reported in Chapter 4 will address those shortcomings.

Auditory Representations of Confidence

The literature reviews by Bisantz (2013) and Chung and Wark (2016) highlight that further research is needed to determine how uncertainty should be indicated when operators' visual attention is heavily taxed, as is the case in domains such as aviation and medicine.

Specifically, the efficacy of auditory displays of system confidence remains relatively unexplored. Past research has evaluated the sonification of the uncertainty associated with position-based data in cartographic applications (Basapur, Bisantz, & Kesavadas, 2003; Bearman, 2011; Bearman & Lovett, 2010). For example, a study conducted by Basapur, Bisantz, and Kesavadas (2003) tasked participants to create paths through a virtual minefield that minimized path length and proximity to mines. Participants could select locations in the virtual minefield to trigger a visual, auditory, or haptic cue that represented the probability that a mine was present. The researchers found that a visual display of uncertainty information led to faster task completion times than auditory and haptic displays. Krygier (1994) describes various properties of sound that may be mapped to data, and notes that varying the pitch of a tone is an effective way to represent ordinal data; however, there is no consensus on how to map pitch to uncertainty, confidence, and other types of meta-information. Fisher (1994) observed that higher reliability might be represented by a higher pitched tone, but such a mapping could be at odds with a common semantic understanding that interprets high pitch tones to represent warnings.

Some studies have represented high uncertainty with a high pitched tone (Bearman, 2011) while others have mapped high confidence (i.e. low uncertainty) to a high pitched tone (Basapur, Bisantz, & Kesavadas, 2003).

The study described in Chapter 3 examines the use of auditory representations of confidence in the context of human supervisory control. As suggested by Multiple Resource Theory (Wickens, 2008) and past research in multimodal displays (Riggs et al., 2017; Wickens, 2008), distributing information across different sensory channels (in this case, vision and hearing) reduces resource competition and allows operators to process more tasks and information in parallel, in a given time period. However, the efficacy of auditory confidence representations may be hindered by crossmodal interaction effects, such as modality shifting (e.g. switching attention from the heavily engaged visual channel to the less used auditory channel), and a constrained capacity to maintain auditory information in working memory for a prolonged period of time (Wickens et al., 2003).

The Relationship Between Attention Management, Mental Models and Trust Calibration

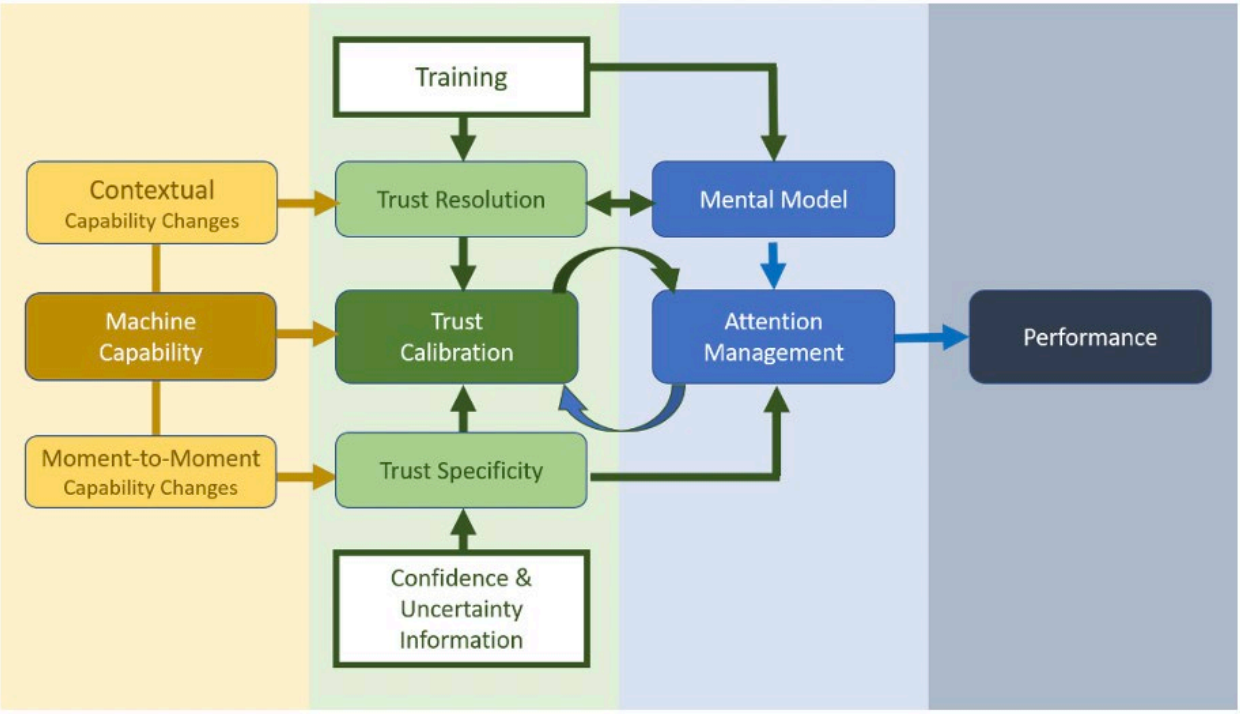
There is a considerable body of research on the assessment and evaluation of trust in human-machine systems. However, few studies have explored the relationship between trust calibration, attention management and operators' understanding of technology, especially during off-nominal conditions and in high-risk domains. Aviation accidents such as Airbus Industrie Flight 129 (Lee & See, 2004), Turkish Airlines Flight 1951 (Hoff & Bashir, 2015), and Asiana Flight 214 (Hergeth et al., 2016; National Transport Safety Board, 2014) illustrate how an operator's improper trust in a system can lead to machine misuse and disastrous outcomes. These accidents exemplify three prototypical contributors to mismatches between an operator's

expectations of a system's capabilities and the system's true capabilities, namely a poor mental model, low system observability, and highly dynamic and/or nonroutine situations (Sarter & Woods, 1995, 1997). In all accidents, the pilots lacked awareness of the active automation state, due to poor display design and inadequate attention management, and they did not understand the implications of the active mode for automation behavior, due to low trust resolution given their flawed mental model of the automation. The result was a breakdown in human-automation coordination at a time when the system operated at its safety boundary.

Figure 1.1 illustrates a conceptual framework of the relationships between trust calibration, attention management, mental models and joint system performance. This framework guides and forms the basis for the present line of research which seeks to develop and evaluate candidate methods to support trust calibration and increase a machine's transparency in a highly dynamic environment. Training was explored as a means to support trust resolution and top-down attention allocation through facilitating the development of an operator's mental model. Representations of a machine's confidence or uncertainty were designed to support bottom-up attention management and trust specificity. The link between trust and attention management was exploited to measure or infer trust based on eye movements. For example, Lu and Sarter (2019) evaluated how system reliability impacted a person's monitoring behavior and trust in a simulated UAV target identification task using eye tracking. They found that several eye tracking metrics, such as increased total fixation durations and greater backtrack and transition rates, corresponded to periods of lower automation reliability and an operator indicating, through ratings, that they had lower trust in the system. Infrequent monitoring was also correlated with higher trust in studies of automated driving (Hergeth et al., 2016; Petersen,

Robert, Yang, & Tilbury, 2019) and the supervision of an automatic pump in a pasteurization plant (Muir & Moray, 1996).

Figure 1.1
The Relationship Between a Machine's Capability, Trust Calibration, Attention Management, and Performance.



At the highest level, this line of research aims to improve the safe adoption of, as well as appropriate compliance and reliance in human-robot systems in attention-demanding, high risk environments. The specific aims of this research were (1) to compare the effectiveness of active, experiential training with more traditional forms of instruction on mental model development, trust resolution, and attention management (Chapter 2) and (2) to assess how various visual and multimodal representations of a machine's confidence in its own ability (Chapters 3 and 4) and the framing of a machine's estimated accuracy as confidence or uncertainty (Chapter 4) shape

operator monitoring of and reliance on technology. A summary of findings and proposals for future work are presented in Chapter 5.

References

Bainbridge, L. (1983). Ironies of Automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Basapur, S., Bisantz, A. M., & Kesavadas, T. (2003). The Effect of Display Modality on Decision-Making with Uncertainty. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 558–561). https://doi.org/https://doi.org/10.1177/154193120304700364

Bearman, N. E. (2011). Using sound to represent uncertainty in future climate projections for the United Kingdom. *Proceedings of the 17th International Conference on Auditory Display (ICAD 2011)*.

Bearman, N., & Lovett, A. (2010). Using Sound to Represent Positional Accuracy of Address Locations. *Cartographic Journal*, *47*(4), 308–314. https://doi.org/10.1179/000870410X12911302296833

Bell, B. S., & Kozlowski, S. W. J. (2008). Active Learning: Effects of Core Training Design Elements on Self-Regulatory Processes, Learning, and Adaptability. *Journal of Applied Psychology*, *93*(2), 296–316. https://doi.org/10.1037/0021-9010.93.2.296

Bibby, K. S., Margulies, F., Rijnsdorp, J. E., Withers, R. M. J., Makarov, I. M., & Rijnsdorp, J. E. (1975). Man's Role in Control Systems. *IFAC Proceedings Volumes*, *8*(1), 664–683. https://doi.org/10.1016/s1474-6670(17)67612-2

Bisantz, A. M. (2013). Uncertainty Visualization and Related Techniques. In J. D. Lee & A. Kirlik (Eds.), *The Oxford Handbook of Cognitive Engineering* (pp. 1–25). https://doi.org/10.1093/oxfordhb/9780199757183.013.0040

Bisantz, A. M., Stone, R. T., Pfautz, J., Fouse, A., Farry, M., Roth, E. M., … Thomas, G. (2009). Visual Representations of Meta-Information. *Journal of Cognitive Engineering and Decision Making*, *3*(1), 67–91. https://doi.org/10.1518/155534309X433726

Chen, J. Y. C., Barnes, M. J., & Harper-Sciarini, M. (2011). Supervisory Control of Multiple Robots: Human-Performance Issues and User-Interface Design. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *41*(4), 435–454. https://doi.org/10.1109/TSMCC.2010.2056682

Chiou, E. K., & Lee, J. D. (2021). Trusting Automation: Designing for Responsivity and Resilience. *Human Factors*. https://doi.org/10.1177/00187208211009995

Chung, J., & Wark, S. (2016). *Visualising Uncertainty for Decision Support*. Victoria, Australia.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. *In Proc. Command and Control …*, 1–37. https://doi.org/10.1.1.90.2591

Dekker, S. (2009). *Report of the Flight Crew Human Factors Investigation: Conducted for the Dutch Safety Board into the Accident of TK1951, Boeing 737-800 Near Amsterdam Schipol Airport, February 25, 2009*.

Dekker, S., & Woods, D. D. (2018). Automation and its impact on human cognition. *Coping*

with Computers in the Cockpit, 7–27. https://doi.org/10.4324/9780429460609-2
Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. A. (2013). Impact of Robot Failures and Feedback on Real-Time Trust. *ACM/IEEE International Conference on Human-Robot Interaction*, 251–258. https://doi.org/10.1109/HRI.2013.6483596
Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., … Yanco, H. A. (2012). Effects of Changing Reliability on Trust of Robot Systems. *Proceedings of the Seventh Annual ACMIEEE International Conference on HumanRobot Interaction HRI 12*, 73–80. https://doi.org/10.1145/2157689.2157702
Du, N., Huang, K. Y., & Yang, X. J. (2020). Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and Reliance, and Task Performance in Human-Automation Teaming. *Human Factors*, *62*(6), 987–1001. https://doi.org/10.1177/0018720819862916
Federal Aviation Administration. (2020). Urban Air Mobility (UAM) Concept of Operations v1.0.
Feltovich, P. J., Bradshaw, J. M., Clancey, W. J., & Johnson, M. (2007). Toward an ontology of regulation: Socially-based support for coordination in human and machine joint activity. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 4457 LNAI, pp. 175–192). https://doi.org/10.1007/978-3-540-75524-1_10
Fisher, P. F. (1994). Hearing the Reliability In Classified Remotely Sensed Images. *Cartography and Geographic Information Systems*, *21*(1), 31–36. https://doi.org/10.1559/152304094782563975
Gao, J., & Lee, J. D. (2006). Effect of Shared Information on Trust and Reliance in a Demand Forecasting Task. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 215–219).
Goodrich, M. A., & Schultz, A. C. (2007). Human-Robot Interaction: A Survey. *Foundations and Trends in Human-Computer Interaction*, *1*(3), 203–275. https://doi.org/10.1561/1100000005
Hager, G. D., Rus, D., Kumar, V., & Christensen, H. (2015). Toward a Science of Autonomy for Physical Systems : A white paper prepared for the Computing Community Consortium committee of the Computing Research Association., 1–10.
Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254
Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting System Uncertainty in Automotive UIs for Supporting Trust Calibration in Autonomous Driving. *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, AutomotiveUI 2013*, 210–217. https://doi.org/10.1145/2516540.2516554
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: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744
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. https://doi.org/10.1177/0018720814547570

Jordan, J. M. (2016). *Robots*. Cambridge, Massachusetts: MIT Press.
Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Anlysis of Decision under Risk. *Econometrica*, *47*(2), 263–292.
Krygier, J. B. (1994). Sound and Geographic Visualization. In *Modern Cartography Series* (Vol. 2, pp. 149–166). Elsevier Science Ltd. https://doi.org/10.1016/B978-0-08-042415-6.50015-6
Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392
Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392
Lee, J. D., Wickens, C. D., Liu, Y., & Boyle, L. N. (2017). *Designing for People: An Introduction to Human Factors Engineering* (Third). Charleston, SC: CreateSpace.
Lu, Y., & Sarter, N. B. (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. https://doi.org/10.1109/THMS.2019.2930980
Marble, J. L., Bruemmer, D. J., Few, D. A., & Dudenhoeffer, D. D. (2004). Evaluation of Supervisory vs. Peer-Peer Interaction with Human-Robot Teams. *Proceedings of the Hawaii International Conference on System Sciences*, *37*(C), 2067–2076. https://doi.org/10.1109/hicss.2004.1265326
Mason, M. T. (2012). Creation Myths: The Beginnings of Robotics Research. *IEEE Robotics and Automation Magazine*, *19*(2), 72–77. https://doi.org/10.1109/MRA.2012.2191437
Mayer, D. (2020). AFLCMC Awards Skyborg contract. Wright-Patterson AFB, Ohio: AFLCMC Public Affairs.
Mayer, R. C., Davis, J. H., & Schoorman, D. F. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734.
McDermott, P. L., Gronowski, M. R., Carolan, T. F., & Fisher, A. (2013). Error Training and Adaptive Remediation: The Impact on Transfer Performance in a Complex Planning Task. *Proceedings of the Human Factors and Ergonomics Society*, 2096–2100. https://doi.org/10.1177/1541931213571467
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*, *48*(4), 656–665. https://doi.org/10.1518/001872006779166334
Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors*, *58*(3), 401–415. https://doi.org/10.1177/0018720815621206
Molesworth, B. R. C. (2005). *Experiential Training and Risk Management Behaviour amongst Pilots*. University of Western Sydney.
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. https://doi.org/10.1016/S0020-7373(87)80013-5
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. https://doi.org/10.1080/00140139608964474
Murphy, R. R., & Burke, J. L. (2010). Human-Robot Interactions in Future Military Operations. In F. Jentsch & M. Barnes (Eds.), *Human-Robot Interactions in Future Military Operations*

(1st ed., pp. 31–49). London. https://doi.org/10.4324/9781315587622
Murphy, R. R., & Shields, J. (2012). *The Role of Autonomy in DoD Systems*. *DoD Defense Science Board*. https://doi.org/10.1016/S0140-6736(02)11924-6
National Academcy of Engineering. (2014). *Autonomy Research for Civial Aviation: Toward a New Era of Flight*. National Academy of Sciences.
National Academies of Sciences Engineering and Medicine. (2021). *Human- AI Teaming: State of the Art and Research Needs*. *Human-AI Teaming*. Washington, DC. https://doi.org/10.17226/26355
National Transport Safety Board. (2014). *Descent Below Visual Glidepath and Impact with Seawall, Asiana Airlines Flight 214, Boeing 777-200ER, HL7742, San Francisco, California July 6, 2013*. Washington, DC.
Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/https://doi.org/10.1518/001872097778543886
Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354
Petersen, L., Robert, L., Yang, X. J., & Tilbury, D. (2019). Situational Awareness, Driver's Trust in Automated Driving Systems and Secondary Task Performance. *SAE International Journal of Connected and Automated Vehicles*, *2*(2), 1–26. https://doi.org/10.4271/12-02-02-0009
Riggs, S. L., Wickens, C. D., Sarter, N. B., Thomas, L. C., Nikolic, M. I., & Sebok, A. L. (2017). Multimodal Information Presentation in Support of NextGen Operations. *International Journal of Aerospace Psychology*, *27*(1–2), 29–43. https://doi.org/10.1080/10508414.2017.1365608
Roth, B. (2008). Forward. In B. Sicilano & O. Khatikb (Eds.), *Springer Handbook of Robotics* (pp. v–ix). Berlin / Heidelberg.
Roth, E. M., Bennett, K. B., & Woods, D. D. (1987). Human Interaction with an "Intelligent" Machine. *International Journal of Man-Machine Studies*, *27*(5–6), 479–525. https://doi.org/10.1016/S0020-7373(87)80012-3
Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson.
Sarter, N. B., & Woods, D. D. (1995). How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 5–19. https://doi.org/10.1518/001872095779049516
Sarter, N. B., & Woods, D. D. (1997). Team Play with a Powerful and Independent Agent: Operational Experiences and Automation Surprises on the Airbus A-320. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(4), 553–569. https://doi.org/10.1518/001872097778667997
Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). *Automation Surprises*. *Handbook of Human Factors and Ergonomics*. https://doi.org/10.1207/s15327108ijap0204_5
Schaefer, K. E., Chen, J. Y. C., 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: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schaefer, K. E., Straub, E. R., Chen, J. Y. C., Putney, J., & Evans, A. W. (2016). Communicating Intent to Develop Shared Situation Awareness and Engender Trust in Human-Agent Teams. *Cognitive Systems Research*. https://doi.org/10.1016/j.cogsys.2017.02.002

Seong, Y., & Bisantz, A. M. (2008). The Impact of Cognitive Feedback on Judgment Performance and Trust with Decision Aids. *International Journal of Industrial Ergonomics*, *38*(7–8), 608–625. https://doi.org/10.1016/j.ergon.2008.01.007

Sheridan, T. B. (1989). *Trustworthiness of Command and Control Systems*. *Analysis, Design and Evaluation of Man–Machine Systems 1988*. IFAC. https://doi.org/10.1016/B978-0-08-036226-7.50076-4

Sheridan, T. B. (2008). Risk, Human Error, and System Resilience: Fundamental Ideas. *Human Factors*, *50*(3), 418–426. https://doi.org/10.1518/001872008X250773

Sheridan, T. B., & Parasuraman, R. (2005). Human-Automation Interaction. *Reviews of Human Factors and Ergonomics*, 89–129.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators*. Cambridge.

Sorkin, R. D., Kantowitz, B. H., & Kantowitz, S. C. (1988). Likelihood Alarm Displays. *Human Factors*, *30*(4), 445–459. https://doi.org/10.1177/001872088803000406

Stowers, K., Kasdaglis, N., Rupp, M. A., Newton, O. B., Chen, J. Y. C., & Barnes, M. J. (2020). The IMPACT of Agent Transparency on Human Performance. *IEEE Transactions on Human-Machine Systems*, *50*(3), 245–253. https://doi.org/10.1109/THMS.2020.2978041

Strater, L. D., Reynolds, J. P., Faulkner, L. A., Birch, D. K., Hyatt, J., Swetnam, S., & Endsley, M. R. (2004). PC-Based Tools to Improve Infantry Situation Awareness, 668–672.

Strauch, B. (2017). The Automation-by-Expertise-by-Training interaction: Why Automation-Related Accidents Continue to Occur in Sociotechnical Systems. *Human Factors*, *59*(2), 204–228. https://doi.org/10.1177/0018720816665459

Tversky, A., & Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *The Journal of Business*, *59*(4), S251–S278.

Tversky, A., & Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. https://doi.org/10.15358/0340-1650-2006-6-331

United States Government Accountability Office. (2017). *Air Force and Army Should Improve Strategic Human Capital Planning for Pilot Workforces*.

United States Government Accountability Office. (2020). *Air Force Should Take Additional Steps to Improve Aircrew Staffing and Support*.

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., & Talleur, D. A. (2003). Attentional Models of Multitask Pilot Performance using Advanced Display Technology. *Human Factors*, *45*(3), 360–380. https://doi.org/10.4324/9781315092898-10

Wiczorek, R., & Manzey, D. (2014). Supporting Attention Allocation in Multitask Environments: Effects of Likelihood Alarm Systems on Trust, Behavior, and Performance. *Human Factors*, *56*(7), 1209–1221. https://doi.org/10.1177/0018720814528534

Wiener, E. L. (1989). Human Factors of Advanced Technology (Glass Cockpit) Transport Aircraft. *(Nasa-Cr-177528)*, (June), 222.

Woods, D. D., Tittle, J., Feil, M., & Roesler, A. (2004). Envisioning Human-Robot Coordination

in Future Operations. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *34*(2), 210–218. https://doi.org/10.1109/TSMCC.2004.826272

Yang, X. J., Wickens, C. D., & Hölttä-Otto, K. (2016). How Users Adjust Trust in Automation: Contrast Effect and Hindsight Bias. *Proceedings of the Human Factors and Ergonomics Society*, 196–200. https://doi.org/10.1177/1541931213601044

Zirk, A., Wiczorek, R., & Manzey, D. H. (2020). Do We Really Need More Stages? Comparing the Effects of Likelihood Alarm Systems and Binary Alarm Systems. *Human Factors*, *62*(4), 540–552. https://doi.org/10.1177/0018720819852023

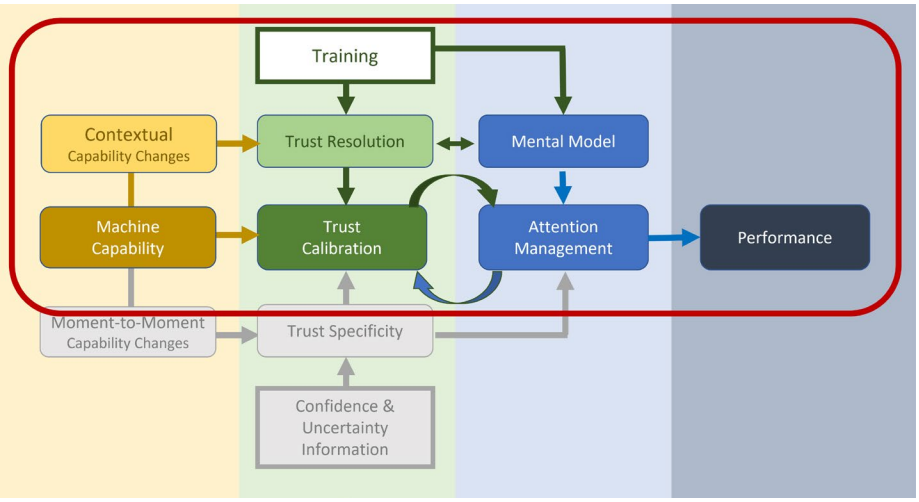
Chapter 2

Learning When to Trust: A Longitudinal Study on the Impact of Training on Trust Calibration, Attention Management and Joint System Performance

One way to support trust calibration and top-down attention management based on a proper mental model of a system is through training, as shown in Figure 2.1. The purpose of this study was to assess how well different types of training achieve these goals and how training outcomes may be affected by actual operational experience with a system over an extended period of time (4 weeks). The experiment complements earlier research which has mostly taken 'snapshots' of trust (over few hours or days) even though the phenomenon is known to take time to develop and recover after it is lost due to a system failure or violations of expectations (Desai et al., 2013).

Figure 2.1

The Study Presented in this Chapter Focuses on the Part of the Conceptual Framework that Relates to How Training Supports Trust Resolution and Top-Down Attention Management



The study was set in a multi-UAV military target classification simulation. Participants in the baseline group were taught “how to work the system”, i.e., how to operate the interface of the multi-UAV system to perform a mission. The second group of participants received additional training using a set of PowerPoint slides to teach them “how the system works”, i.e., which processes and environmental factors affect UAV behavior/performance. The third group received the same information through experiential learning, i.e., by providing feedback and coaching to participants as they completed interactive training simulations. Note that the second and third groups both received the baseline training, in addition to their supplemental training.

The study’s main expectations were that:

1. Learning how a highly autonomous system works in an experiential context, versus in a more passive instructional context (through viewing PowerPoint slides), would lead to improved trust resolution and more timely and appropriate monitoring of UAVs.

2. Observations of actual UAV performance over the course of the study would further improve trust calibration, thus reducing differences in performance seen between training groups at the start of the data collection.

3. During high-workload phases, operators would – by necessity – intervene less often and give consent to UAV assessments more quickly, even though their trust ratings may not change.

4. Erroneous target assessments by the UAV at the beginning of the longitudinal study would result in larger decrements of trust than when the same errors would occur at the end of the study.

Method

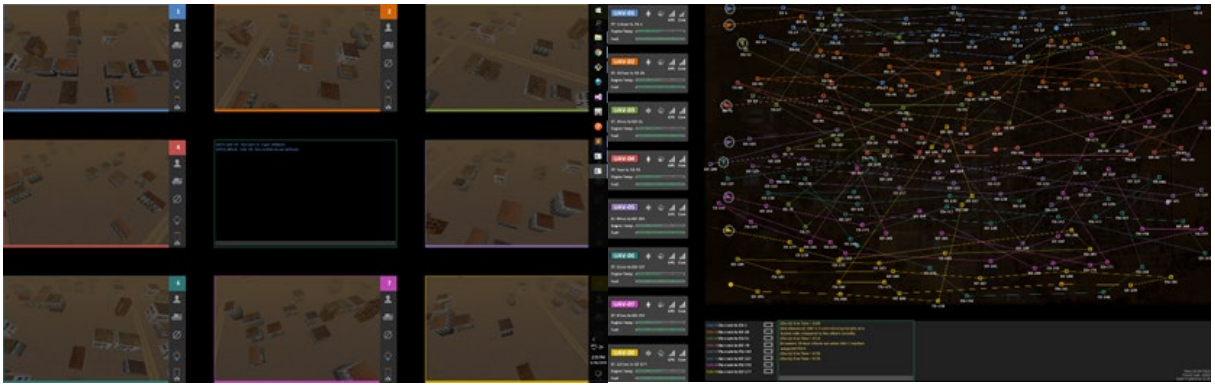
Participants

Twenty-eight University of Michigan students between the ages of 18-30 years old (M = 23.3 years old, SD = 2.11) completed the experiment. An Air Force SME confirmed that the participants' age range was comparable to that of Air Force UAV pilots, and that it would be appropriate to use students in this study since it simulates a futuristic concept with which current Air Force UAV pilots do not have experience. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University of Michigan (UM IRB: HUM00162932). Informed consent was obtained from each participant.

Task and Apparatus

Participants were tasked with monitoring the simulated video feeds of eight unmanned aerial vehicles (UAVs) in a military reconnaissance simulation with the purpose of detecting and classifying adversarial tanks and personnel. The simulator interface (shown in Figure 2.2) was presented on two side-by-side monitors. Each UAV followed a pre-planned route, shown on a map on the right-side monitor. Simulated camera video feeds of the ground from each UAV's vantage point were shown in a 3x3 grid on the simulator's left monitor. The center window of the grid contained a chat room that simulated real-time messages being exchanged with peer UAV operators and air traffic controllers.

Figure 2.2
Simulator Interface

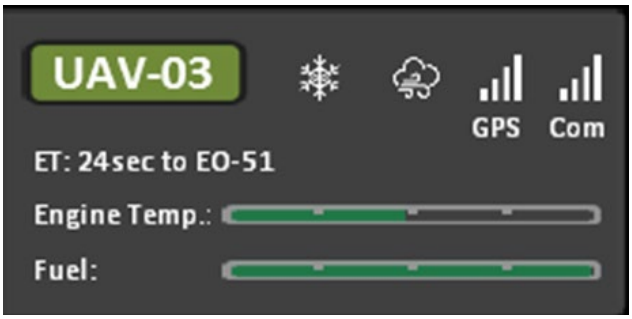


During each scenario, a UAV's video feed became highlighted when the UAV reached a pre-determined waypoint and the UAV momentarily loitered. The UAV analyzed the field directly below it and informed the participant if it had identified either a person, a tank, or neither a person nor a tank at the waypoint. The recommended classification of the scene was expressed through highlighting the corresponding button (e.g., a button with the icon of a person or a button with the icon of a tank; see Figure 2.3). The participant then reviewed the scene and pressed the button corresponding to their belief of the correct classification. A participant had 12 seconds to respond to the recommended classification before the UAV continued to the next waypoint. The recommendations provided by the UAV were hard coded into the simulation itself, rather than being calculated in real-time, to ensure that there was an overall recommendation reliability rate of 80%; past research has suggested that lower levels of reliability may cause monitoring, trust, and performance to suffer (Wickens & Dixon, 2006; Dixon & Wickens, 2007).

Figure 2.3
The UAV Simulator Recommends a "Person" Classification

In addition to the target identification task, participants were asked to monitor two chatrooms and provide information when prompted. One chatroom was reserved for simulated Air Traffic Control (ATC) communication while the second chatroom represented a simulated Mission Room which supported communication between troops, an off-site person known as a "Screener" who reviewed and interpreted surveillance footage, and the participant (in the role of the UAV ground control station operator). Participants were also responsible for monitoring each UAV's health display (see Figure 2.4), which was comprised of a wing icing warning icon, a turbulence warning icon, indicators of the vehicle's GPS and communication signal strength, and fuel and temperature gauges. The identifier of the UAV's next waypoint and its estimated time of arrival (ETA) at the waypoint were included in the health display. During the initial baseline training ("how to work the system"), participants were advised how to respond to decrements in a UAV's health. For example, a participant was trained to press a UAV's "Return to Base Icon" when it was low on fuel.

Figure 2.4
UAV Health Display



Finally, the map shown on the right monitor (see Figure 2.2) provided an aerial view of all waypoints and updated with each UAV's position in real-time. Participants were advised that UAVs would occasionally deviate from their preprogrammed paths. When participants noticed a deviation, they were expected to click on the UAV's icon on the map. The UAV would then return to its pre-planned flight path.

Experiment Design

The experiment employed a mixed factorial design. The two independent variables were (1) training method (3 levels; between-subjects) and (2) operational tempo (two levels; within-subjects).

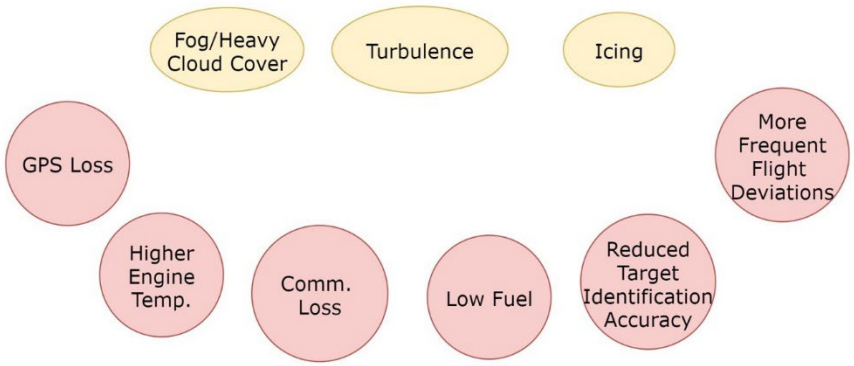
Training Method

Participants were randomly assigned to one of three groups (baseline, passive training, experiential training) that each received a different method of training during the first week of the experiment. The baseline group was trained only on how to use the simulator interface "to work the system", i.e., to complete the target classification task, maintain system health successfully, and send messages in chat rooms. This group received no guidance, feedback, or

interaction from the experimenter while completing training exercises with the simulator. The second group of participants reviewed a supplemental PowerPoint presentation that taught them "how the system works." Participants were taught the dependencies, limitations, and inter-relationships of system components and the operational environment (e.g., how weather limited the performance of a sensor which, in turn, could affect the accuracy of automation recommendations that were dependent on that sensor). This training method was a form of passive learning. The third group learned "how the system works" through experiential learning – an active process. In experiential training, participants completed a training scenario while receiving guidance and feedback on their actions, learning from the researcher why the system behaved in certain ways and how to correctly respond to certain situations. Both the second and third group of participants received the baseline training too. Care was taken to ensure that the exposure time to UAV operations was the same for all three groups. To this end, all three groups completed the second training scenario, but received different (or no) instruction with it.

Participants in the experiential training group performed a reflection exercise at the conclusion of their training, before starting the actual experiment. They were provided a handout and asked to complete a concept map (shown in Figure 2.5). The handout showed three environmental conditions (fog, turbulence and icing) and 6 operational events, such as "GPS Loss." Participants were asked to draw lines between the environmental conditions and the operational events to indicate how they believed the environment could affect the UAV's performance. The concept map was completed by participants in all three groups at the conclusion of the longitudinal study to assess/score and compare their mental models of the system.

Figure 2.5
Concept Map Worksheet



Operational Tempo

Operational tempo, which refers to the number of events in a time period, was varied within subjects. During a ten-minute high operational tempo period, there were five UAV malfunctions (such as flight deviations) and 27 prompts in chatrooms (such as requests from Air Traffic Control for UAVs to descend to lower altitudes). Ten-minute low operational tempo periods contained two or fewer UAV malfunctions and approximately seven prompts in chatrooms.

Procedure

Data for each participant was collected over four weeks, with two days of data collection occurring each week. During the first week, a participant received training each day, followed by a 15-minute training scenario and 30 minutes of actual data collection. A debrief questionnaire was completed after each data collection session. During the remaining three weeks, twice each week, participants completed a 30-minute scenario with no additional training.

Dependent Measures

The dependent measures in this study included self-reported trust scores, target classification compliance and performance, secondary task performance (health monitoring, chatroom participation), eye tracking data, and debrief questionnaires. Target classification performance (e.g., accuracy, response time, and compliance), as well as responses to health monitoring information and chat room messages, were all recorded by the simulator. Eye movement data was collected and analyzed using the commercially available Tobii Pro Glasses 2 and accompanying Tobii Pro Lab software.

During the experiment, the simulation paused every ten minutes to collect subjective trust ratings from the participant. A pop-up window appeared, and participants were prompted to provide trust ratings on a scale from 1 (low) to 10 (high). These subjective ratings of the system included their overall trust in the system, as well as their trust in each of the eight UAVs. Once the ratings were entered, the pop-up window disappeared and the simulation resumed.

Results

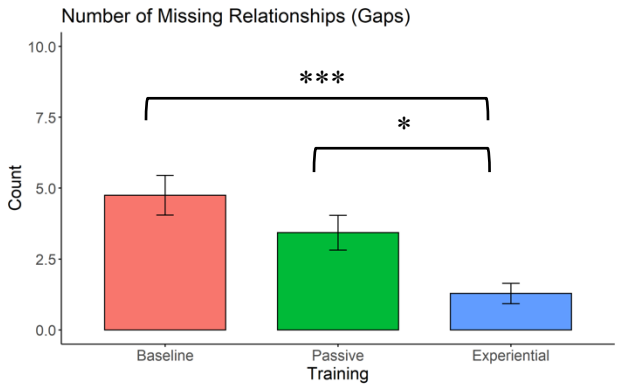
The performance data and trust ratings were analyzed using linear mixed models with random effects. Analyses were performed using the statistics software R, the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015), and the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017). Likelihood ratio tests of the model with the examined fixed effect and the model without the examined fixed effect were used to obtain p-values.

Training and Mental Model

It was expected that participants who received experiential training, as compared to passive training, would have a more accurate mental model of the multi-UAV system, leading to improved trust resolution and more timely and appropriate monitoring of UAVs (Expectation 1).

To assess this expectation, participants' concept maps (Figure 2.5) were analyzed. Relationships between environment factors and UAV health decrements that were not drawn on the concept map worksheet suggested a gap in a participant's mental model. A generalized linear model analysis found that participants who received the experiential training had, on average, 73% fewer gaps in their concept map than participants in the baseline training group (z = 3.53, 95% CI [0.47 0.88], $p < 0.001$), and an estimated 63% fewer gaps than participants who received the passive training (z = -1.25, 95% CI [0.22 0.84], p = 0.012; see Figure 2.6). The difference in missing relationships between the baseline and passive training groups was not significant.

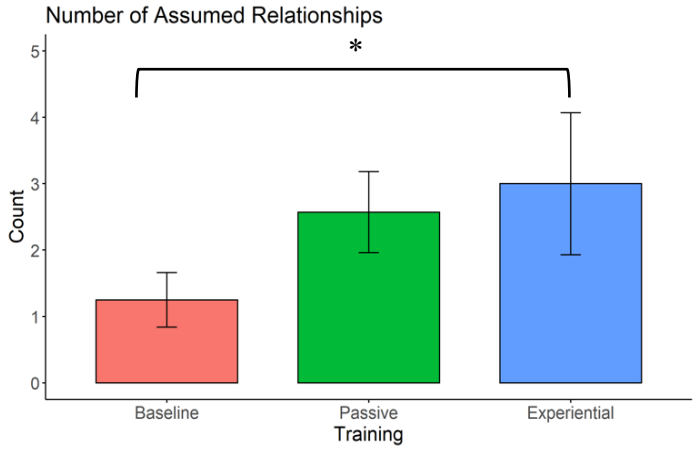
Figure 2.6
Number of Missing Relationships (Gaps) in Concept Map as a Function of Training



If a participant incorrectly associated an environment factor with a UAV health decrement, such as drawing a line on the worksheet to incorrectly suggest that fog might lead a UAV to run out of fuel faster, this was counted as an "assumed relationship" (similar to a "false positive"). Figure 2.7 shows the number of assumed relationships in a participant's concept map as a function of their training. A generalized linear model analysis found that the concept maps of participants in the experiential training group included significantly more (approximately 2.4

times as many) assumed relationships than the maps of participants who received just the baseline training (z = 2.28, 95% CI [1.2 5.3], p = 0.023). Participants in the passive training group assumed approximately twice as many incorrect relationships as the baseline group (z = 1.83, 95% CI [1.0 4.6], p = 0.067).

Figure 2.7
Number of Assumed Relationships in Concept Map as a Function of Training



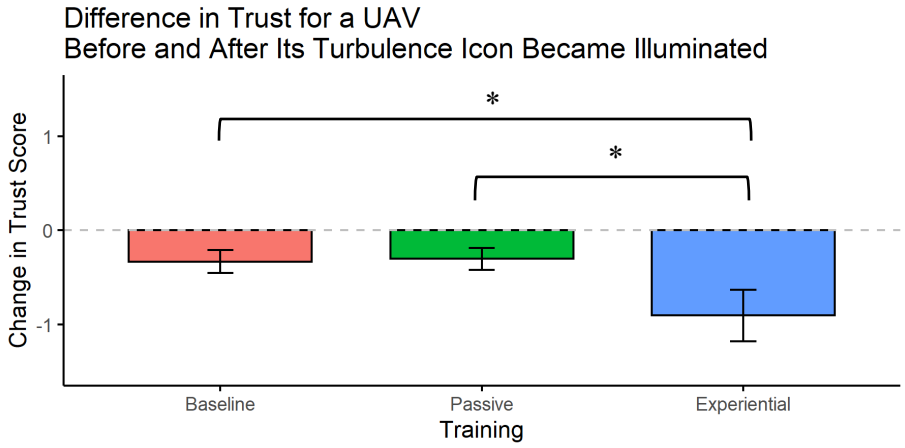
Training and Trust

One way to assess the impact of training on trust resolution is by examining the impact of turbulence on a participant's UAV trust scores. This study simulated that a UAV experiencing turbulence would provide less reliable classifications, have higher engine temperatures (leading to a UAV's engine to become too hot), and experience faster fuel burn (causing a UAV to run out of fuel). Participants in the passive and experiential training groups were informed of this causal relationship as part of their supplemental training which the baseline group did not receive. It was anticipated that participants who were aware of the impact of turbulence on UAV health and target classification capabilities would provide lower trust scores for a UAV once the

turbulence icon became illuminated. Figure 2.8 shows the change in trust scores by training group for a UAV before it entered turbulence and after it entered turbulence.

Figure 2.8

Change in Trust Score in Response to Turbulence Icon Illumination in Sessions 2, 4, and 8



A linear model analysis with training as a fixed effect and the data collection session and participants as random effects found a marginally significant main effect on the change in trust scores for a UAV whose turbulence icon had become illuminated ($\chi^2(2) = 5.62$, $p = 0.060$). Trust scores for UAVs significantly dropped more among participants who received experiential training than the participants who received the baseline training (M =-0.6, $t(36.67) = -2.06$, $p = 0.046$) and passive training (M = -0.6, $t(37.27) = -2.20$, $p = 0.034$). An additional linear model analysis found that the differences in trust decrements between training groups that were associated with a UAV's turbulence icon becoming illuminated did not significantly change over the course of the experiment (Expectation 2).

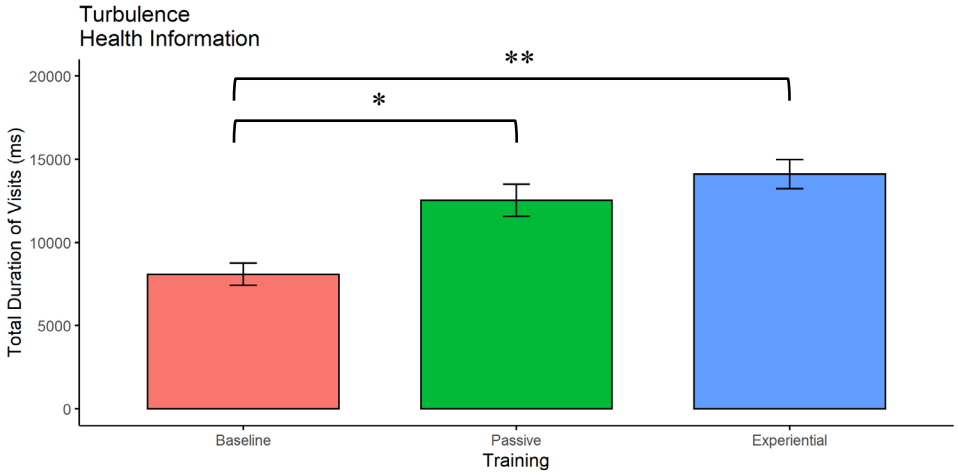
Training and Monitoring

Eye tracking data was evaluated to assess the impact of training on how participants monitored a system based on their mental model and trust resolution (Expectation 1). Only participants in the passive and experiential training groups were informed that a UAV was more likely to fail during turbulence due to faster fuel burn and engine temperature increases.

Figure 2.9 illustrates how participants in all three training groups monitored a UAV's health information (e.g., engine temperature and fuel) when it experienced turbulence.

Figure 2.9

Total Duration of Visits to Health Information AOI During Turbulence, Aggregated Across All Scenarios



Training had a significant impact on the total duration of visits to the health information display ($\chi^2(2) = 7.87$, $p = 0.020$), and a marginally significant impact on the number of visits to the health information display ($\chi^2(2) = 5.26$, $p = 0.072$). On average, participants in the passive training group cumulatively spent 44% more time than the baseline group monitoring the health information during turbulence ($t(36.01) = 2.13$, $p = 0.040$), and participants in the experiential

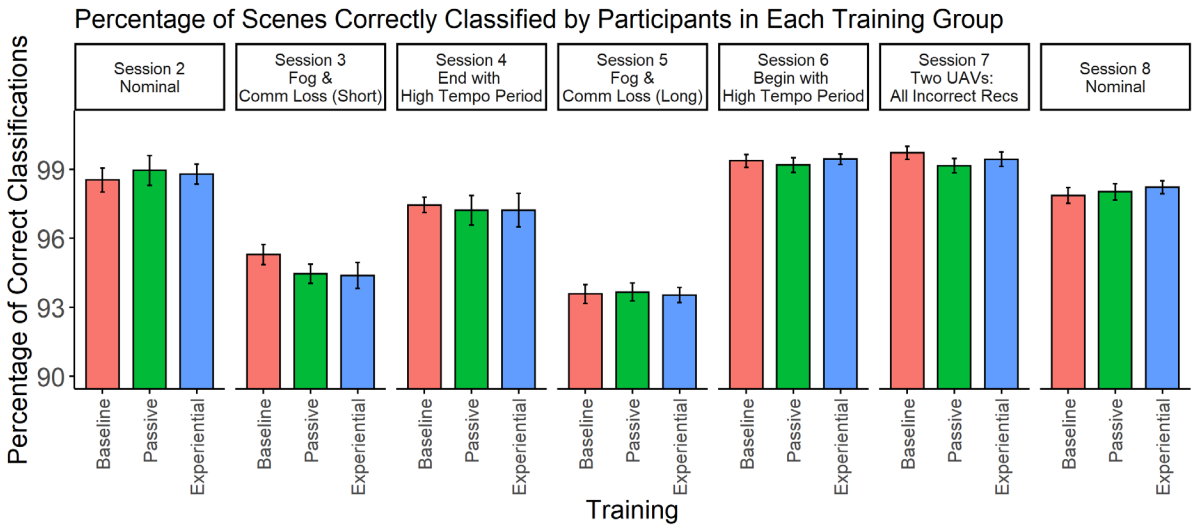
training group cumulatively spent 61% more time than the baseline group monitoring a UAV's health information during turbulence ($t(34.94) = 2.87$, $p = 0.007$).

Training and Performance

The impact of training and observing UAV performance over the course of the study (Expectation 2) was assessed by evaluating the accuracy of scene classifications and the time required to complete the classification task, as well as participants' responses to UAV health failures. The bar graphs in Figure 2.10 show the percentages of scenes that were correctly classified by participants in each training group. Note that a ceiling effect was observed, and the y-axis has been zoomed-in to a range of 90% to 100%. A generalized linear model showed that the type of training was not a significant predictor of classification accuracy, and accuracy did not change significantly over the course of the experiment. As shown in Figure 2.10, the percent of correct classifications was slightly lower in scenarios the contained fog, resulting in a greater number of incorrectly *recommended* classifications and ultimately incorrect classifications provided by the joint human-UAV team.

Figure 2.10

Accuracy of Scene Classifications as a Function of Training and Scenario



Linear mixed models evaluated the impact of training on participants' response time to flight deviations and on the time required to classify a scene. Only targets that were classified correctly by participants were included in the analysis; a scene classified *incorrectly* and quickly would not be a direct comparison of performance if it were compared with a scene that was classified correctly and more slowly. Training did not significantly impact the time it took to correctly classify targets nor their response time to flight deviations. Additional linear mixed effects analyses evaluated the impact of training on a person's response time to "low fuel" and hot "engine" warnings; again, training was not found to have a significant effect on the response time to either of these warnings.

High Tempo Operations

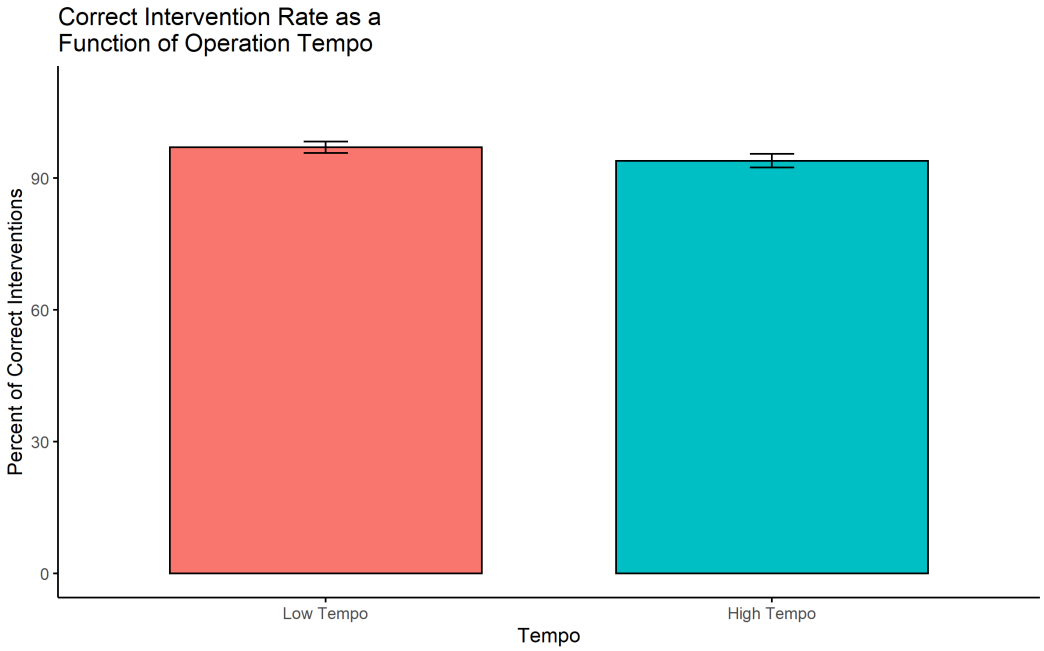
Expectation 3 predicted that participants would intervene less often and comply with UAV classification recommendations more quickly during higher tempo periods (i.e., when

workload and attention demands were higher). Participants completed two scenarios which included both low tempo and high tempo periods.

To distinguish between cases when participants complied with a UAV's recommendation due to high operational tempo and high workload, and instances when participants complied with a UAV's recommendation because it was indeed correct, participant compliance in cases when a UAV provided incorrect recommendations in high and low tempo periods was evaluated. Figure 2.11 plots the percentage of scenes that were incorrectly classified by a UAV but were correctly classified by participants. It was expected that participants would comply with the UAV recommendations and incorrectly classify the scenes when the operational tempo and attention demands were high. However, a generalized linear model showed that participants correctly intervened and did *not* comply when a UAV provided incorrect recommendations, regardless of operational tempo. There was no interaction between training and operational tempo.

Figure 2.11

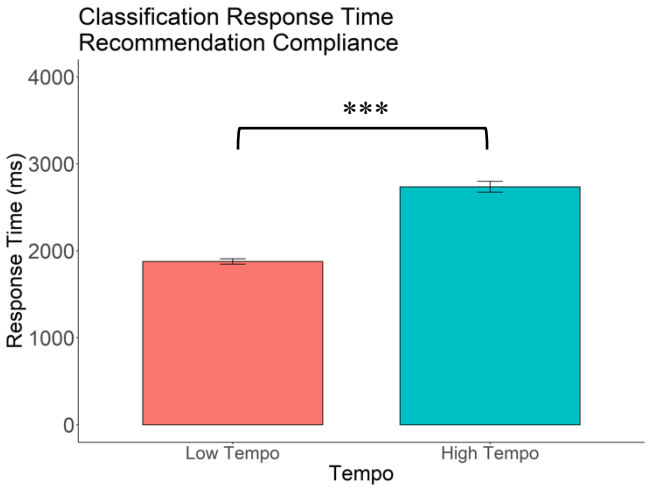
Percentage of Scenes Incorrectly Classified by a UAV but Correctly Classified by Participants as a Function of Operational Tempo



Response times to comply (i.e., agree) with correct recommended classifications are shown in Figure 2.12. A linear mixed model analyzed the impact of tempo on the response time to (correctly) comply with the UAV's recommended classification. An increase in tempo was found to significantly increase the response time ($\chi^2(1) = 264.56$, $p < 0.001$) by an average of 864 milliseconds.

Figure 2.12

Response Time to Comply with a Correct Recommended Classification



Longitudinal Evolution of Trust

The study's fourth expectation predicted that erroneous target recommendations provided by a UAV at the beginning of the longitudinal study would lead to larger trust decrements than incorrect target classifications towards the end of the study. This expectation was examined using the scenarios performed during the first and last data collection sessions which were identical in design; participants were not informed that they were completing the same scenario in both sessions. Figure 2.13 shows the difference in a UAV's trust scores before and after it provided an incorrect recommendation as a function of whether the scenario was completed at the start or end of the longitudinal study. A linear mixed model analysis was conducted, with the timing of the data collection session (i.e. first or final data collection session) as a fixed effect, and the participant and the waypoint that was classified as random intercepts. Likelihood ratio tests of the model with the fixed effect and the model without the session fixed effect were used to obtain p-values. Incorrect recommendations at the beginning of the longitudinal study did not

lead to more significant trust decrements than the same incorrect recommendations at the conclusion of the study. A second linear mixed model analysis that also included the training group as a fixed effect did not find that there was a statistical difference in trust decrements between training groups for incorrect recommendations at the beginning of the study and at the conclusion of the study.

Figure 2.13

Difference in Trust Scores for Incorrect Classification Recommendations Provided by a UAV in the Study's First and Last Data Collection Sessions

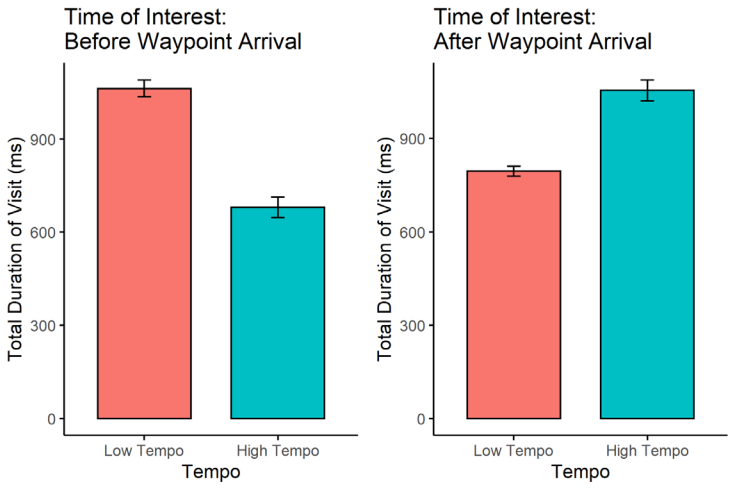
Previewing

An unexpected result was that participants took more time to classify targets during high tempo operations (with higher attention demands and workload) than during periods with lower tempo operations, fewer attention demands, and less workload (Expectation 3, see Figure 2.12). Debrief responses and eye tracking data were reviewed to explain this unexpected finding.

Twenty-five out of the 28 participants who completed the four week study reported in the debrief questionnaire that, even though they were not instructed to do so, they referenced a countdown clock, which displayed the number of seconds remaining until a UAV arrived at a

waypoint, or used the Navigation Map (shown on the right display in Figure 2.2) to anticipate when a UAV would arrive at a waypoint. This allowed participants to "preview" and mentally classify the scene at a waypoint before the automation's recommended classification was provided. Figure 2.14 shows the total visit duration for a UAV video feed before and after the recommended classification was presented. The figure suggests that participants spent more time "previewing" the scenes during low tempo periods, whereas participants spent more time reviewing the scene after the UAV had arrived at a waypoint during high tempo periods.

Figure 2.14
Monitoring Behavior, Before and After Waypoint Arrival, as a Function of Operation Tempo



Trust Definition

Since the previous analyses of trust ratings yielded unexpected findings, the trust ratings were further examined to assess their validity. Specifically, there was a concern that, as participants progressed through the study, they would no longer base their trust ratings on the definition of trust provided at the start of the experiment, during training, but rather on their own interpretation of the concept prior to or as it evolved throughout the study.

While reviewing the self-paced PowerPoint based training on the first day of the experiment, all participants were informed that, "In this experiment, trust is defined as *the attitude that a machine will help achieve a person's goals in a situation characterized by uncertainty and vulnerability.*" This was based on Lee and See's (2004) definition of trust in automation. Participants were then prompted every ten minutes in each scenario to rate their trust in each individual UAV. After completing the scenarios at the midpoint and at the conclusion of the multi-week study, each participant completed a second debrief questionnaire (separate from the debrief questionnaire that was completed each day) that asked participants to recall the definition of trust in this experiment.

Only two of the participants could closely recall the trust definition. All other participants provided different interpretations of trust. It was also observed that the trust definition presented in training was broad and did not refer to specific UAV capabilities. However, the trust definitions provided by 14 participants who completed the cumulative debriefs suggest that they adopted a narrower definition of trust which reflected only their trust in a UAV's classification abilities; other capabilities, such as a UAV's ability to manage its health, were not considered in their trust scores.

Effects of Agreement Between Proposed and Actual Target on Attention Management

Past research has suggested that eye tracking metrics may be an effective method to infer a person's trust in automation technology (Hergeth et al., 2016; Lu & Sarter, 2019). To assess the validity of this approach, the eye tracking data was examined to determine whether factors other than trust, such as the agreement between automation recommendations and actual target presence/type at each waypoint, influenced a person's gaze behavior.

Eye tracking data was recoded to distinguish the four types of pairings between a UAV's recommendation and the type of target a scene contained:

- *Target*: The UAV correctly indicated that a target was present at a waypoint.
- *Different Target*: A target was present in the video feed, but the UAV indicated either that a different target was present or suggested classifying the scene as not having a target at all. For example, the UAV suggested that a "tank" classification should have been provided, but a person could be seen in the video feed.
- *No Target:* The UAV correctly suggested that a "none" classification should be provided; there was no target at the waypoint.
- *Target Missing:* The UAV indicated that a tank or person classification should have been provided, but no target could be seen in the video feed.

The Total Visit Duration and the Number of Fixations for a scene, as a function of the congruence between the recommended and actual target, are shown in Figure 2.15 and Figure 2.16, respectively.

Figure 2.15

Total Visit Duration for a Scene's AOI based on the Congruence between Recommended and Actual Target

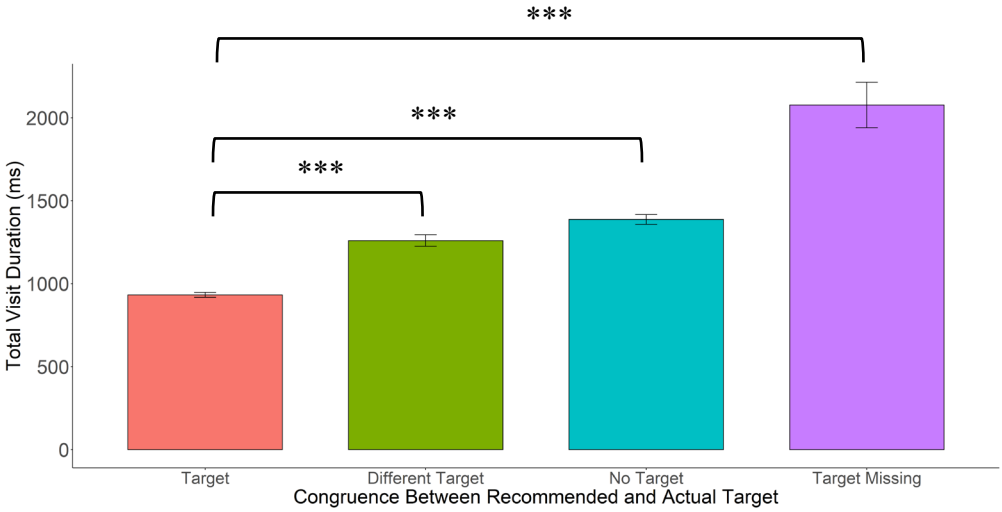
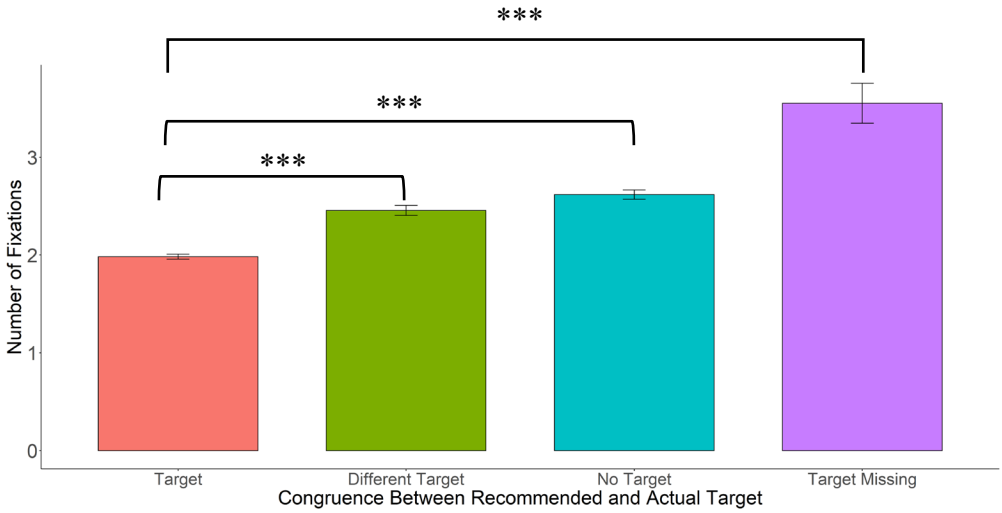


Figure 2.16

Number of Fixations in a Scene's AOI based on the Congruence between Recommended and Actual Target



A linear mixed model analysis assessed the impact of the congruence between recommended and actual target on the duration of a participant's visit to a video feed during the

classification task, and a generalized linear mixed model analyzed the impact of congruence on the number of fixations in the video feed. In both models, congruence was a fixed effect and the participant was modeled as a random intercept. Likelihood ratio tests found that congruence significantly impacted the total visit duration ($\chi^2(3) = 361.88$, $p < 0.001$) and the number of fixations ($\chi^2(2) = 73808$, $p < 0.001$) in a video feed when classifying a waypoint's scene. There was no interaction between the type of training and the congruence between the recommended and actual target. Pairwise comparisons further showed that people and tanks that were misclassified by a UAV had an average total visit duration that was 334ms (39%) longer than the average total visit duration for people and tanks that were classified correctly ($t(5605.95) = 10.23$, $p < 0.001$), and there was a mean increase of 25% more fixations in such cases ($z = 9.40$, $p < 0.001$). Scenes that were classified correctly as not having any targets had an average increase of 31% more fixations ($z = 13.39$, $p < 0.001$) and an average total visit duration that was 449ms (53%) greater than scenes with targets that were classified correctly ($t(5605.61) = 15.74$, $p < 0.001$). Finally, when a tank or person recommendation was provided for scenes without targets, the total visit duration increased by 1178ms (138% increase) and there was an average increase of 86% more fixations ($z = 9.95$, $p < 0.001$) than instances when a tank or person classification was correctly recommended ($t(5604.72) = 11.22$, $p < 0.001$).

Discussion

This study assessed the effects of training on trust resolution, attention management and performance. Participants were assigned to one of three groups that each received a different method of training. Participants in all groups, including the baseline group, received instruction on how to operate the multi-UAV system to complete the mission's tasks. Two of the groups received additional training that emphasized how a UAV's environment and internal processes

impact a UAV's performance; one of these groups performed the supplemental training with a set of PowerPoint slides (i.e., "passive learning") while the remaining group was coached and asked reflection questions via an interactive training simulation (known as "experiential" or "active" learning). Participants completed eight data collection sessions over the course of four weeks. Task performance, eye movement data, and self-reported trust in the multi-UAV system over the course of the study were evaluated.

Training and Mental Model Development

Participants completed a worksheet on the final day of data collection to externalize their mental model of the UAV system. Training significantly impacted the quality of a participant's mental model. Specifically, experiential training reduced the number of gaps in the model, compared to passive training. Reflection questions that were incorporated into the experiential training may have better supported the encoding of information about the UAVs' internal processes into participants' memory. Also, the experimenter "coached" participants in the experiential (but not the passive) training group during the training scenario. This allowed participants to become aware of and correct inaccuracies in their mental model.

At the same time, the experiential training group showed the largest number of (incorrectly) assumed relationships in their mental model. This may be explained by the fact that the supplemental training they received "revealed" some causal relationships between a UAV's environment and its health, which were not emphasized in the baseline training. This may have led participants to wonder if there were additional relationships that they had not been informed about. One way to avoid this problem in future real-world operations may be to gather data from UAV operators at the end of their experiential training, using a concept map (like the one in

Figure 2.5), to identify common misconceptions and explicitly inform subsequent trainees that these relationships do not exist.

Participants in the baseline training had more gaps in their mental models than participants who received experiential training at the start of the experiment. This suggests that these participants were not able to infer over time, as a result of operational experience with the system, how a UAV's processes and environment caused its performance to decrease. One reason why participants in the baseline group may have had trouble inferring the relationships between a UAV's environment and its performance is that they had to cope with high task demands and thus may not have had the attention resources required to engage in hypothesis testing to deduce the inner workings of the UAVs.

Trust Resolution, Monitoring and Trust Scores

Eye tracking metrics indicated that training also impacted how participants monitored a UAV's health information. Specifically, compared to the baseline group, participants in the passive training group and the experiential training group spent more time cumulatively gazing at the health information when the turbulence icon for a UAV was illuminated. This can be explained by the fact that they were informed during training that turbulence could lead to vehicle health failures. Thus, training resulted in better trust resolution leading to more effective attention allocation. This improved trust resolution also manifested in more appropriate and dynamic trust ratings. Participants in the experiential training group lowered their trust in a UAV to a greater extent than the rest of the participants when a UAV experienced turbulence.

Surprisingly, there were no significant differences in trust between the baseline and passive training groups when a UAV entered turbulence. Furthermore, for all training groups, the magnitude of trust decrements when a UAV provided an incorrect recommendation was the

same independent of whether the incorrect classification was made at the start or towards the end of the study. Though it is possible that there were indeed no significant differences in trust, it is also worth considering that the limitations of the trust measurement method may have obscured changes in a person's trust. Only two participants could recall the Lee and See (2004) definition of trust that was provided at the beginning of the study. Lee and See's rather broad definition of trust was the attitude that *a UAV* would be helpful; however, half of the participants more narrowly defined trust in the debrief questionnaire as an attitude that *a UAV's classification capabilities* would be helpful. Thus, participants' trust scores may reflect only their trust in a UAV's classification abilities, and not include their trust in a UAV's health management capabilities. This explanation is further supported by the responses to the concept map worksheet (see Figure 2.5); while 57% of the participants in the experiential training group indicated their (correct) belief that turbulence would reduce a UAV's classification capabilities, only 25% of participants in the baseline group and 14% of participants in the passive training group indicated this relationship. Therefore, participants in the passive training group may have more closely monitored the health information display to detect fuel and engine warnings caused by turbulence but did not believe that turbulence warranted adjusting their trust in a UAV's recommended classifications.

Training and Performance

Joint human-machine performance on the waypoint classification task was extremely high. This ceiling effect meant that no significant differences in accuracy, response time, and associated monitoring behavior were observed as a function of training. While extensive pilot testing was conducted in advance of the study to ensure that task difficulty was high enough and thus a ceiling effect would be avoided, participants in the actual experiment adopted an

unexpected strategy. Debrief responses, as well as the eye tracking data, indicate that participants began viewing a scene as it entered the edge of the video feed's display (moments before the UAV would arrive at its waypoint, before the scene would appear in the center of the video feed, and before the automation's recommended classification would be provided). Participants would then classify the scene on their own before the automation recommended a classification. The eye tracking data suggests that participants were able to engage in this behavior when the operational tempo was low and there were fewer competing attention demands. While the previewing strategy was not expected nor desired from an experiment design perspective, it may be an effective strategy in operational contexts when an operator has to cope with high competing attention demands in a short period of time. Previewing may allow operators to spread out these demands and better balance their workload so that they can attend to multiple tasks over a longer time interval.

Another reason why performance did not differ as a function of type of training may be the low cost of not trusting recommended classifications. Participants needed to just briefly glance at the video feed if they did not trust (or wanted to validate) the classifications provided by the UAVs. More involved reviewing and/or response tasks might increase the cost of not relying on a machine's recommendations, leading to greater longitudinal effects and an impact of training on performance.

Contingent Orienting

Participants' monitoring behavior was significantly influenced by whether the UAV's recommended classification was correct and matched the imagery seen in the UAV's video feed. The time spent visually scanning the video feed was the shortest when the target identified by the UAV was present, followed by instances when a target was present but a different classification

was recommended. This suggests that participant gaze behavior was influenced by an attentional phenomenon known as "contingent orienting" (Folk, Remington, & Johnston, 1992). Contingent orienting refers to the situation where a stimulus captures a person's attention, highly reliably and involuntarily, if it matches the person's top-down control settings (e.g. Folk et al. (1992)). Based on the eye tracking metrics, participants were able to identify targets more quickly if the target matched the recommended classification provided by the automation. This suggests that participants incorporated the UAV's recommendations into their decision-making when determining how to classify waypoints.

These results also suggest that incorrect recommendations can have significant implications on the monitoring and gaze behavior of an operator when a target (or an expected stimulus) is not present. Most notably, participants spent more than twice the amount of time inspecting a video feed and had nearly twice as many fixations when a target was missing than when the expected target could be found. It is possible that participants conducted a more thorough and extensive review of the video feed to ensure that no target was missed. The increased allocation of visual attention resources when an expected stimulus is not present needs to be further studied in applications with greater attention demands, as it may restrict the ability of an operator to sufficiently monitor a system in a high tempo, multi-tasking environment.

Finally, this study's observation that gaze behavior is influenced by the accuracy of a machine's recommendation, and by congruence between recommended and actual target, complicates the use of eye tracking and response times to measure trust. Future research is needed to determine to what extent trust shapes how a person monitors a system differently from typical visual search tasks. For example, if a machine sequentially provides a series of inaccurate recommendations, will a person visually inspect and review the display more because the person

has low trust in the machine? Or could the person still have high trust in the machine and just naturally review the display more thoroughly because the expected target stimulus cannot be found?

Conclusion

Findings from this study suggest that experiential training may facilitate the development of an operator's mental model of a system and thus improve trust calibration. Furthermore, results indicate that operational exposure to a system may not be sufficient for an operator to successfully infer causal relationships in a system. This might be because a high tempo, high risk operational environment like the one simulated in this study does not afford a person the opportunity to observe and explore a system's capabilities, reflect on their experience and thus improve their mental model – as can be done in experiential training. Therefore, instructors and managers tasked with designing training programs may find experiential training to be more beneficial than traditional, simplified training methods to address trust miscalibration and breakdowns in attention management.

Researchers should be aware that participants may adopt different semantic understanding of prompts requesting them to self-report their trust in a system, and eye tracking measures that are employed to infer trust may be influenced by visual search phenomena (such as the congruence between an expected visual target and the stimulus). Accordingly, future research might benefit from asking participants to rate their trust through more narrowly defined prompts and at a higher frequency. For example, participants might be asked to rate their trust in the UAV's capability to perform one task (rather than a person's trust in all the UAV's capabilities) or in each situational context. However, limitations and tradeoffs must be considered when using subjective ratings to assess trust. Asking participants to rate their trust in

only a subset of a machine's capabilities may not provide a complete picture of how a participant trusts a machine. Furthermore, more frequent prompts for a participant to rate their trust in a machine may disturb a participant's engagement in the mission, affect short term memory, perturb reflective cognitive processes, and degrade joint human-machine performance.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. A. (2013). Impact of Robot Failures and Feedback on Real-Time Trust. *ACM/IEEE International Conference on Human-Robot Interaction*, 251–258. https://doi.org/10.1109/HRI.2013.6483596

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors*, *48*(3), 474–486. https://doi.org/10.1518/001872006778606822

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary Covert Orienting Is Contingent on Attetional Control Settings. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 1030–1044.

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: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models . *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lu, Y., & Sarter, N. B. (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. https://doi.org/10.1109/THMS.2019.2930980

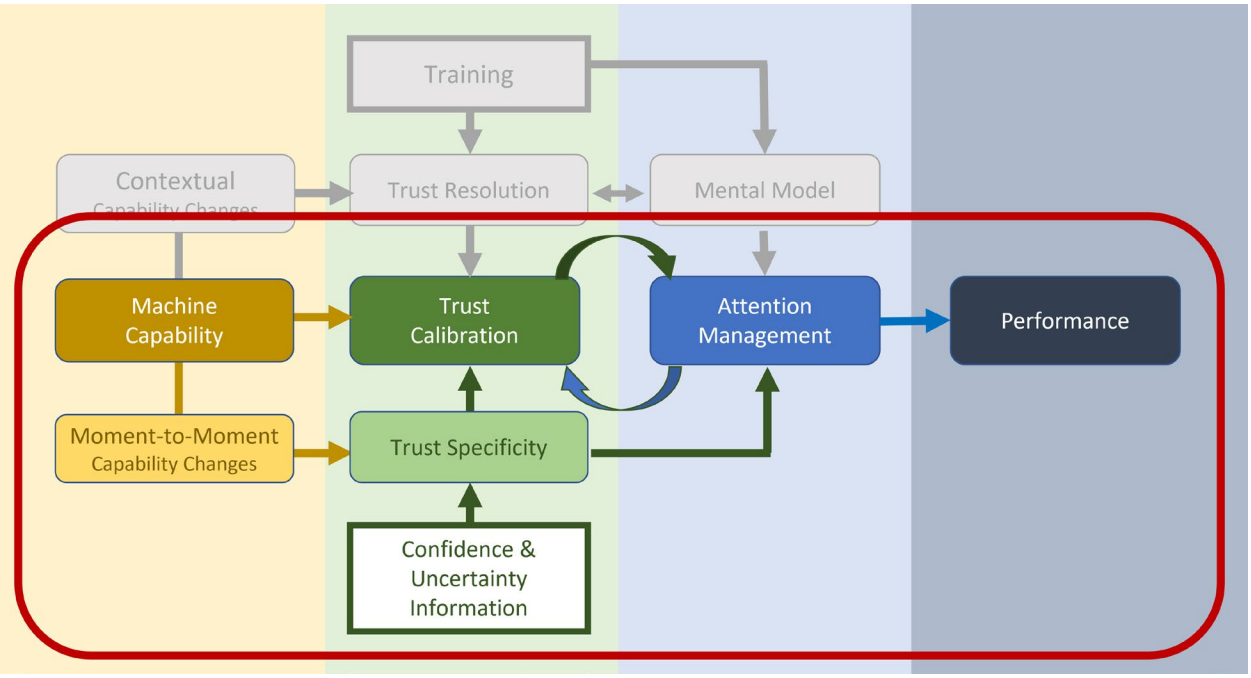
Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. https://doi.org/10.1080/14639220500370105

Chapter 3

A Comparison of Auditory and Visual Representations of System Confidence to Support Trust Specificity, Attention Management, and Joint Performance in Human-Machine Teams

The previous study examined how supporting mental model development in advance of operations, through three different training approaches, might improve trust resolution and calibration, as well as top-down attention management. The next two studies in this line of research focus on the impact of system transparency, i.e. real-time feedback on system confidence (or uncertainty) in its own performance, on operators' trust specificity, monitoring behavior and joint system performance; see Figure 3.1. Providing information about moment-to-moment fluctuations in estimated system accuracy is intended to guide attention in a bottom-up fashion, prompting the operator to more closely monitor the automation and review its recommendations in instances of low reliability.

Figure 3.1
The Study Presented in this Chapter Focuses on the Part of the Conceptual Framework that Relates to How Confidence or Uncertainty Information Supports Trust Specificity and Bottom-Up Attention Management



The first of the two studies examined the effects of auditory and visual representations of system confidence on trust specificity and attention management in the context of supervision of a multiple unmanned aerial vehicle (UAV) target classification system. This system, like many other high-risk application domains, imposes significant competing visual attention demands on human operators. As suggested by Multiple Resource Theory (Wickens, 2008) and past research in multimodal displays (Lu et al., 2013; Riggs et al., 2017; Sarter, 2013; Wickens, 2008), it is beneficial to distribute information across different sensory channels (in this case, vision and hearing) in such domains as this reduces resource competition and allows operators to process

simultaneously multiple tasks and sources of information. Therefore, auditory representations of system confidence were expected to reduce response times, lead to a supervisor noticing the recommended classifications more reliably and improve joint system performance across the entire task set when compared to visual representations. In general, providing information about system confidence was expected to improve trust specificity, with high system confidence leading to an operator complying with UAV assessments faster and more often while low confidence should lead to slower response times since an operator would allocate more resources to monitor and review the automation's recommendations.

Method

Participants

Eighteen engineering students from the University of Michigan, including 9 males and 9 females between the ages of 18-30 *(M = 22.6, SD = 3.79),* were recruited for this study via school mailing lists. Participants were required to have normal or corrected-to-normal vision without color-deficiencies ("color blindness"), and normal hearing abilities. This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board at the University of Michigan (UM IRB: HUM00136276). Informed consent was obtained from each participant.

Task and Apparatus

The task in this experiment was to monitor the simulated video feeds of six UAVs for the presence of targets. Participants were informed that automation onboard the UAV would scan pre-defined regions to assist with the detection of the target, which was a green sedan. During the 25-minute scenario, the UAVs followed preplanned trajectories. Simulated camera video feeds

of the ground from each UAV's vantage point were displayed in a 2x3 grid on a single 30" monitor display. A UAV's video feed became highlighted when it had possibly identified a desired target. The participant then reviewed the scene and pressed one of two buttons to either confirm or reject the presence of a target (see Figure 3.2). In addition, participants needed to scan the various UAV video feeds on a continuing basis to make sure all targets were detected.

Figure 3.2
Simulator Screenshot

Design

This experiment employed a within-subjects 3x2 factorial design. The two independent variables were the modality in which confidence information was presented (none, visual, auditory) and the confidence with which an object was identified by the UAV (high, low). The visual confidence information consisted of a green border that appeared around UAV windows containing a potential target. High confidence was represented by a highly saturated border while low confidence was represented by a less saturated border. In the auditory condition, a high pitch tone represented high confidence that a potential target had been detected, and a low pitch tone

corresponded to low confidence. During each block of trials, only one modality (none, visual or auditory) was used to present confidence information. In the auditory condition, the respective UAV video feed was highlighted to indicate which UAV the tone was associated with.

Procedure

Each experiment session started with participants being informed about the goals of the experiment. Participants received training on the simulation and the search tasks for five minutes before each block of trials.

A total of three blocks of trials were completed. Each block lasted 25 minutes during which system confidence was either not provided, or encoded in visual or auditory form, respectively. During each run, the system of UAVs identified a total of 163 potential targets. Seventy-eight of the items were identified with high confidence and 85 of the items were identified with low confidence. Participants were informed that objects identified as targets with high confidence were indeed targets approximately 90% of the time, while objects identified as targets with low confidence were actual targets only 60% of the time. The order of blocks was counterbalanced, and while not known to the participants, each scenario had the exact same events in the same order (and differed only with respect to the modality that confidence information was presented in). Pilot testing in advance of the study confirmed that participants did not realize that identical scenarios were performed.

The entire experiment lasted approximately two hours. Participants completed an online debrief questionnaire at the conclusion of all trials, and the participants were paid a total of $30 for compensation.

Dependent Measures

The dependent measures included performance on the target detection task (response time and detection rate/accuracy) and subjective trust ratings. Target detection performance was captured and recorded by the simulator. During the experiment, the simulation paused once every two minutes and participants were prompted to rank their trust in each UAV on a scale from 0-9 (with '9' being the highest possible trust rating). Unlike the first study, the UAVs in this study performed only the target detection task and did not need to simultaneously manage their own health. Therefore, the trust prompts in this study were not ambiguous with regards to which task was being trusted. Once the participants had completed entering their trust ratings, the simulation automatically resumed.

Results

The performance data and trust ratings were analyzed using linear mixed models with random effects. Analysis was performed using the statistics software R, the *lme4* package (Bates, et al., 2015), and the *lmerTest* package (Kuznetsova et al., 2017). Likelihood ratio tests of the model with the examined fixed effect and the model without the examined fixed effect were used to obtain p-values.

Response Times

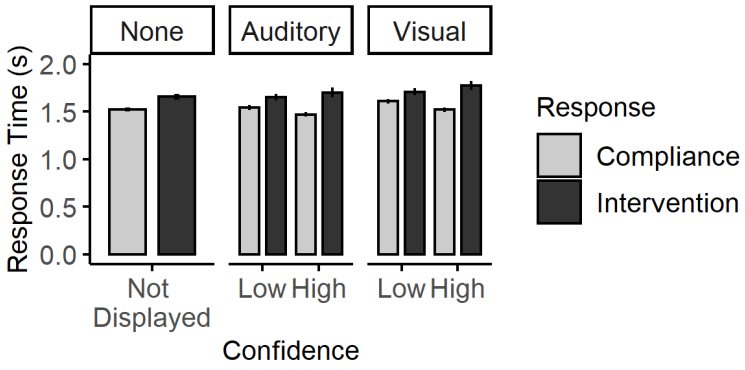
The means and standard deviations of response times to system recommendations as a function of representation modality and confidence level are shown in Table 3.1. A linear mixed model compared the response times between targets identified with high confidence and low confidence. Among the trials with visual and auditory representations of confidence, participants on average responded 0.08 seconds slower when a UAV had low confidence than high

confidence (χ^2 (1) = 61.9, $p < 0.001$). It is interesting to note that objects that were detected with high confidence led to faster response times not only with auditory and visual confidence representations but also when no confidence information was provided. In an effort to explain this outcome, a linear mixed model analyzed the impact of interventions (i.e. when the automation incorrectly indicated a target and the participant needed to determine and respond that a target was not present) on participant response times. The model included the participant response (to comply or reject the recommended detection), the modality of the confidence representation, and the confidence level associated with a waypoint (high or low) as fixed effects; the participant was considered as a random effect. The results show that the need to intervene and reject the UAV recommendation, irrespective of modality and confidence level, increased the response time by an average of 0.14 seconds (χ^2 (1) = 192.8, $p < 0.001$); see Figure 3.3.

Table 3.1
Response Times

		Representation Modality		
		None	Auditory	Visual
Confidence	High	*M* = 1.528s *SD* = 0.363	*M* = 1.499s *SD* = 0.386	*M* = 1.553s *SD* = 0.380
	Low	*M* = 1.579s *SD* = 0.431	*M* = 1.577s *SD* = 0.432	*M* = 1.641s *SD* = 0.439

Figure 3.3
Response Time as a Function of Confidence, Modality, and Response Type



A linear mixed model assessed whether low confidence auditory representations led to longer response times than high confidence auditory representations. Since erroneous detections by the automation led to longer participant response times, the full model included both the confidence level and the response type (i.e. compliance or intervention) as fixed effects and the participant as a random effect. The reduced model excluded the confidence fixed effect and showed that low confidence auditory representations increased the response time by 0.046 seconds (χ^2 (1) = 10.0, p = 0.002).

A linear mixed model compared the response times for each modality. The modality was included as a fixed effect and each waypoint and participant as random effects. Modality impacted response times with statistical significance (χ^2 (2) = 62.7, $p < 0.001$). Specifically, pairwise comparisons showed that auditory representations of confidence resulted in response times that were 0.017 seconds faster than no confidence information (t(8223) = -2.22, p = 0.026, d = 0.06), and response times for visual representations of confidence were 0.043 seconds slower than no confidence information (t(8224) = 5.49, $p < 0.001$, d = 0.15). A second

linear model with the modality and confidence level included as fixed effects did not a find a significant interaction between the two factors.

Accuracy

Accuracy on the detection task was evaluated by summing the number of times that a participant correctly complied with the automation's identification of the target and the number of times that the participant correctly intervened and reported that a target was not present at the waypoint. Overall, accuracy did not differ significantly as a function of modality ($\chi^2(2) = 4.23$, $p = 0.121$). However, during the trial block with visual representations of confidence, participants were more likely not to respond to UAV target detections (on average, one percentage point more of the waypoints had no responses than the trial with no confidence information ($t(89) = 2.54$, $p = 0.013$)), and they incorrectly complied with the automation's recommended detections at a higher rate (on average, incorrect compliance was 3% higher than during the "no confidence representations" trials ($t(90) = 1.97$, $p = 0.051$)).

Trust Ratings

Trust ratings were collected to determine whether the availability of confidence information improved participants' trust calibration and to examine the relationship between subjective trust and monitoring behavior. Since trust is likely more calibrated towards the end of a block of trials (based on observed system performance), the analysis compared the median of the last three trust scores for each UAV. The UAV and the participant were random effects in the linear mixed model. Two hypotheses regarding trust outcome variables were tested using Bonferroni adjusted alpha levels of 0.025. Auditory, but not visual, confidence information led to a slight but not significant increase in trust (by 0.21 points, on a ten-point scale) over no confidence information (χ^2 (1) = 4.4527, $p = 0.035$). While participants were provided the

opportunity to rate their trust for each (individual) UAV, half of the participants reported equivalent trust scores for all UAVs.

Debrief Responses

After completing the study's three blocks of trials, participants completed a debrief questionnaire. Eleven out of 18 participants responded that the information about the automation's confidence affected their own assessment of whether a target was presented. Only 8 out of 18 participants thought the confidence information impacted how they monitored the system. Nine participants considered the auditory confidence representations to be the most useful, followed by 7 participants who indicated that the visual representations were most useful, and two participants who did not find either the visual or auditory confidence representations helpful. Twelve participants felt that high confidence should be represented by a high pitch tone whereas two participants thought high confidence should be represented by a low pitch tone; four participants had a neutral response when asked how confidence should be mapped to a representation's pitch.

Discussion

The purpose of this experiment was to determine whether providing information about a machine's confidence in its ability to detect a target would improve trust calibration and attention management, as well as overall system performance. In addition, two different representations of confidence information - auditory and visual - were compared to assess their effectiveness for supporting operators in evaluating a system's trustworthiness quickly and reliably.

The Impact of Confidence on Performance

The observed difference in response times between high confidence and low confidence targets suggests that participants incorporated the confidence information into their signal evaluations. Longer response times for the automation's lower confidence detections indicate that participants spent more time inspecting and monitoring the video feeds independently to avoid that they incorrectly complied with the automation's recommendations. For auditory representations of confidence, the natural mapping of a high pitch tone to a high level of confidence resulted in faster response times, suggesting that participants intuitively and correctly mapped the pitch of a tone to the machine's reliability level, in parallel with performing their visual tasks, and then allocated their attentional resources accordingly.

Task accuracy did not differ significantly as a function of confidence. Low confidence detections, while leading to slower response times as they likely prompted more monitoring of the respective video feed, did not increase attention demands to the extent that participants were less responsive to low confidence detections than high confidence detections.

Visual Representations of Confidence Added to Attention Demands

In contrast to Basapur et al. (2003) who observed faster response times to visual representations of confidence, the use of this modality in this study led to longer response times and more instances where the participant did not respond at all to the automation's target recommendation. This may be explained, in part, by the specific implementation of visual confidence information in our experiment. The intent of the design was to allow participants to process the information pre-attentively, in peripheral vision, to avoid interference with the visual target detection task. However, despite extensive pilot testing of the design, participants' feedback following the actual experiment suggests that the visual cues were not salient enough

and actually increased visual processing demands by requiring focal visual attention. The fact that visual representations were not salient enough may be related to a limitation of this study. We did not perform cross-modal matching, i.e. asking participants to adjust the intensity of the visual and auditory cues to match their perceived intensities (Pitts, Riggs, & Sarter, 2016), in advance of the experiment.

The auditory representation of confidence information did not increase response time, suggesting that participants were able to successfully attend to both the auditory confidence information and the visual detection task at the same time. This finding is in line with Multiple Resource Theory, which predicts better time sharing when information is distributed across sensory channels, due to reduced competition for attentional resources.

Providing confidence information did not improve the accuracy of the joint human-machine team, and statistically significant differences in mean response times were on the magnitude of hundredths of seconds and thus had a small effect. Future research might consider assessing the impact of confidence representation on performance with a testbed that incorporates additional (secondary) tasks and adds more substantial attention demands on a participant; this may result in larger effects being observed and better illustrate how the provision of confidence information may critically impact performance in high tempo operational contexts.

Incorrect Detections Led to Longer Response Times

The same scenario (with different confidence representations) was presented to participants in each trial. Unexpectedly, targets that were classified as “high confidence” in the visual and auditory conditions led to faster response times even in trials with no confidence information. Further analysis revealed that this finding can be explained by the fact that participants generally took more time to respond when the automation incorrectly indicated the

presence of a target, suggesting that participants may have been concerned that they initially missed seeing the target, and spent more time inspecting a video feed when the target was not immediately found.

Multi-UAV Trust Ratings

It was expected that subjective trust ratings would reflect better trust calibration when confidence information was provided. Instead, auditory (but not visual) confidence information resulted in slightly higher, but not necessarily more appropriate trust ratings. One possible reason why the expected benefit of providing confidence information was not observed is that, in contrast to most previous experiments on trust, participants had to provide individual ratings for multiple UAVs whose reliability varied throughout the study. Participants reported in the debrief questionnaire that it was difficult for them to track and remember the performance of each UAV, a challenge that was likely exacerbated by the high tempo of operations.

Conclusion

In summary, the findings from this study indicate that providing information about a system's confidence in its own abilities may be an effective technique for improving the performance of human-machine teams, especially in high tempo operations with high attentional demands. Operators in this experiment showed faster response times when the machine identified a target with high confidence and an intervention was not necessary. However, the results also highlight that performance effects critically depend on the specific design of confidence information. Visual representations increased the demands on an operator's attentional resources which were taxed also by the visual detection task and thus resulted in slower response times due to interference. In contrast, auditory representations resulted in faster response times as they avoided task interference and employed a natural mapping between high

pitch tones and high confidence levels. Future research might consider comparing how different auditory design methods for representing confidence impact performance. Studies might also consider using eye tracking to compare how visual and auditory confidence representations impact gaze behavior and system monitoring.

References

Basapur, S., Bisantz, A. M., & Kesavadas, T. (2003). The Effect of Display Modality on Decision-Making with Uncertainty. In *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting* (pp. 558–561). https://doi.org/https://doi.org/10.1177/154193120304700364

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models . *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lu, S. A., Wickens, C. D., Prinet, J. C., Hutchins, S., Sarter, N. B., & Sebok, A. L. (2013). Supporting interruption management and multimodal interface design: Three meta-analyses of task performance as a function of interrupting task modality. *Human Factors*, *55*(4), 697–724. https://doi.org/10.1177/0018720813476298

Riggs, S. L., Wickens, C. D., Sarter, N. B., Thomas, L. C., Nikolic, M. I., & Sebok, A. L. (2017). Multimodal Information Presentation in Support of NextGen Operations. *International Journal of Aerospace Psychology*, *27*(1–2), 29–43. https://doi.org/10.1080/10508414.2017.1365608

Sarter, N. B. (2013). Multimodal Displays: Conceptual Basis, Design Guidance, and Research Needs, (September 2017), 1–19. https://doi.org/10.1093/oxfordhb/9780199757183.013.0038

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Chapter 4

Comparing the Effectiveness of Hue- Versus Salience-Based Representations of Confidence and Uncertainty for Supporting Trust Calibration and Attention Management

The findings from the study described in Chapter 3 suggested that visual confidence representations may degrade performance as they compete for attentional resources with operators' visual tasks and therefore result in longer response times to automation classifications. However, it is not clear whether this finding was the result of the specific implementation of confidence information (i.e., an additive representation with high confidence corresponding to a highly saturated border), and whether a different visual representation of confidence might lead to a better outcome. Additionally, the attention demands in the study may have been too low, as suggested by the ceiling effect for classification accuracy, which may have obscured performance benefits of providing confidence information. The study described in this chapter therefore compared two different visual representations of confidence-related information in the context of more demanding tasks.

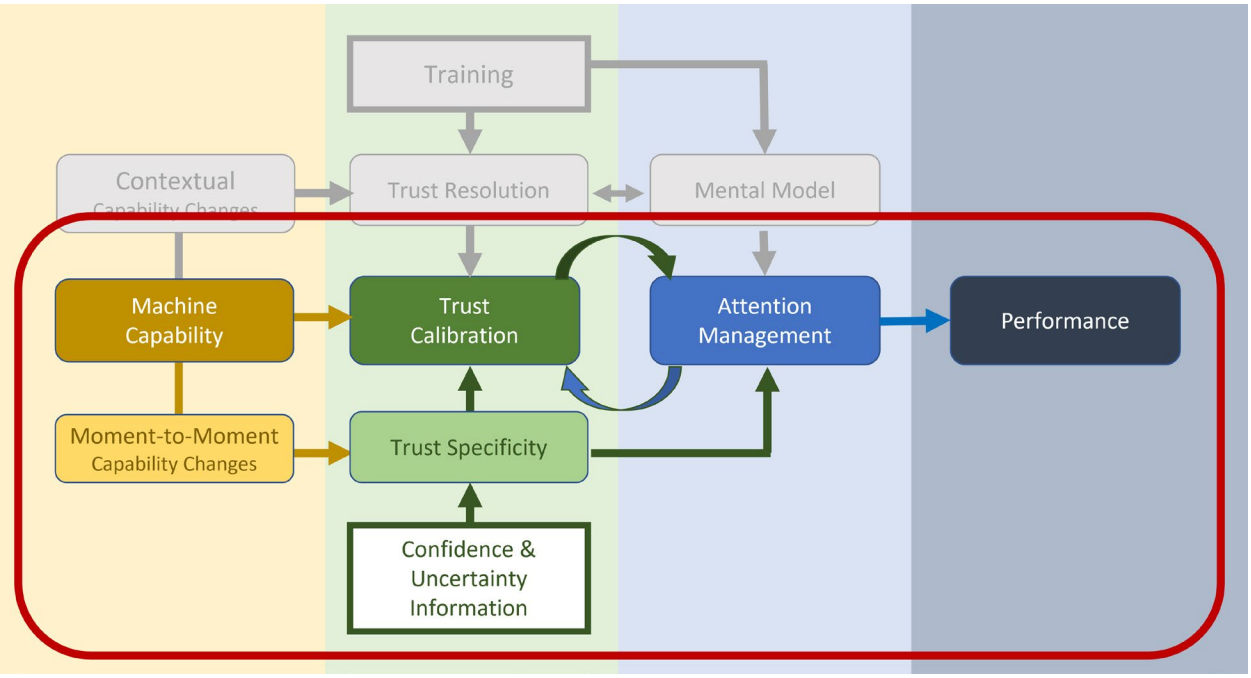
The study also examined how the framing of confidence information as confidence or uncertainty may affect trust calibration and attention management, as shown in Figure 4.1. Past research has evaluated visualizations of confidence and uncertainty separately to assess how well they enable a person to determine a system's estimated accuracy quickly and accurately. In contrast, this experiment directly compared the two framing techniques. Finally, the study included both valid and invalid estimates of system performance to examine whether and how

inaccurate confidence information affects trust levels and monitoring behavior. The following four outcomes were expected:

1. Salience-based representations of uncertainty, as opposed to salience-based representations of confidence, would capture attention faster and more reliably when the estimated accuracy of a classification was low. This prediction was because uncertainty representations would become brighter with low assumed accuracy while the opposite would happen with confidence information.
2. Participants would find it easier to distinguish between different levels of confidence and uncertainty with hue-based (as opposed to salience-based) representations of a machine's estimated accuracy which, in turn, would lead to better attention management and better performance on the classification and secondary tasks.
3. Participants would monitor recommended classifications more closely with uncertainty (as opposed to confidence) framing. This expectation was based on past research which suggested that a negative framing of impact (i.e. uncertainty) can influence attitudes and behavior more than a positive framing of benefits and gains (i.e. confidence; Tversky & Kahneman, 1986). In addition, uncertainty framing would likely cause a participant to adopt more risk averse behavior (Sheridan, 2008), leading to slower classification times, initially lower trust scores, better classification accuracy, and worse performance on the mission's secondary tasks, compared to confidence information.
4. As a corollary to the third expectation, it was anticipated that participants in the uncertainty condition would be more likely to notice and subsequently adjust their

monitoring and trust in the multi-UAV system when there was a mismatch between a system's true and estimated accuracy (Expectation 4).

Figure 4.1
The Study Presented in this Chapter Focuses on the Part of the Conceptual Framework that Relates to How Confidence or Uncertainty Information Supports Trust Specificity and Bottom-Up Attention Management



Method

Participants

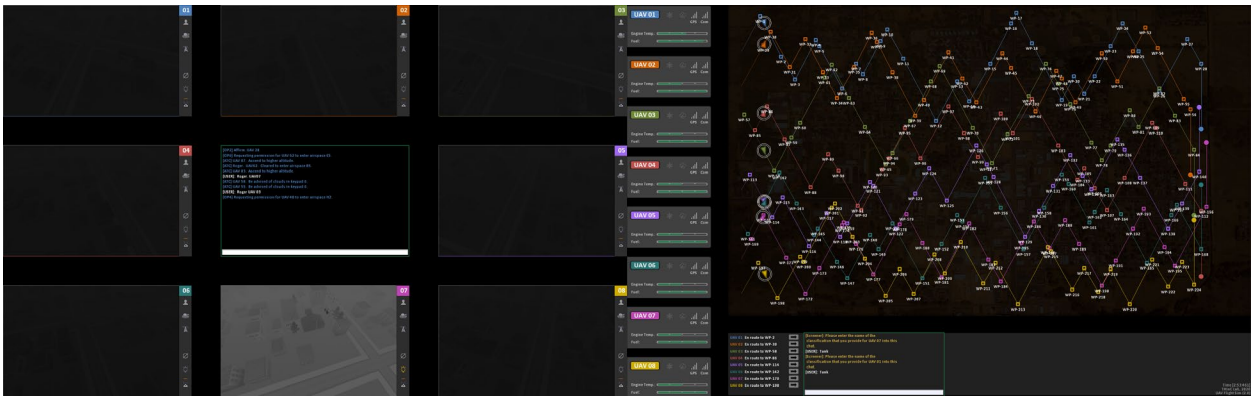
Sixty University of Michigan students between the ages of 18-30 years old (M 22.8 years old, SD = 2.48) completed the study. An Air Force Subject Matter Expert confirmed that the age range of the participants was analogous to that of Air Force UAV pilots. This research complied with the American Psychological Association Code of Ethics and was approved by the

Institutional Review Board at the University of Michigan (UM IRB: HUM00197449). Informed consent was obtained from each participant.

Task and Apparatus

Participants supervised eight unmanned aerial vehicles that classified imagery in a military reconnaissance task. The experiment's testbed was an augmented simulator based on the Air Force Vigilant Spirit Control Station. The simulator's interface spanned two monitors (see Figure 4.2); the left monitor displayed the video feeds for each of the UAV's thermal cameras and the right monitor displayed an aerial map that depicted the position and health of each UAV in real-time.

Figure 4.2
Simulator Interface



Once a UAV reached a target area, it analyzed the field directly below itself and informed the participant if it had identified either a person, a tank, or communication equipment in the scene. The automation's classification was expressed through the illumination of an icon of a person, a tank, or a radio tower to the right of the video feed (see Figure 4.3). In addition, the UAV displayed its confidence or uncertainty in its recommendation via a border that appeared

around the entire video feed. The Experiment Design section will provide more detail on the various confidence/uncertainty representations. Initially, the infrared imagery at each waypoint was too dark to be interpretable by the participant; participants had to either rely on the recommended classification or elect to brighten the video feed to inspect the imagery more closely before classifying the scene. A participant could either choose to comply with the automation's recommended classification of the scene, select a different classification that they believed was correct, or press a button to indicate that they believe no target was present at the waypoint. Participants had twelve seconds to classify the scene waypoint arrival before the UAV would proceed to its next waypoint.

Figure 4.3
The UAV Simulator Recommends a "Person" Classification with a Border Indicating that the UAV has High Confidence in its Classification

In addition to identifying possible targets at each waypoint, participants needed to attend to three secondary tasks. They had to monitor and respond to messages from an Air Traffic Control chatroom, located in the center of the video feed displays (as shown in Figure 4.2), as well as a Mission Room chat room that simulated inquiries from other military personnel about how the participant interpreted and classified the video feed imagery. Participants were also

tasked with monitoring the aerial map, which displayed the preprogrammed flight path and position of each UAV in real-time. During training, participants were cautioned that UAVs would occasionally deviate from their preplanned trajectories. They were told to click the corresponding UAV icon which would return the vehicle to its route.

Experiment Design

This experiment employed a mixed design. The four independent variables were (1) framing of estimated accuracy information (3 levels; between-subject), (2) method of representing confidence/uncertainty (2 levels; between-subject), (3) operational tempo (2 levels; within-subject), and (4) representation accuracy (2 levels; within-subject).

Framing of Estimated Accuracy Information (Between-Subjects)

Participants were randomly assigned to one of three groups which were presented with either confidence information, uncertainty information, or neither confidence nor uncertainty information for each recommended classification.

Method of Representing Confidence/Uncertainty (Between-Subjects)

The two groups who received either confidence or uncertainty information were further subdivided into two groups each. One group was informed that the border around a video feed would stay a constant hue but vary in brightness, such that high uncertainty/confidence recommendations were represented with the more salient (i.e. brighter border; see Figure 4.4). This additive representation can be considered a form of natural mapping. In the other group, the border used a red-yellow-green color scheme, as shown in Figure 4.5. A red border corresponded to assumed low accuracy (i.e. low confidence; high uncertainty), a yellow border corresponded to a medium level of accuracy (i.e. medium confidence; medium uncertainty) and a green border corresponded to high accuracy (i.e. high confidence; low uncertainty). This substitutive mapping

tries to exploit conventions and participants' familiarity with similar color coding (e.g., traffic lights, (Wickens & Hollands, 2000, p. 101)). Participants were informed during training that 95% of recommendations with high estimated accuracy, 80% of recommendations with medium estimated accuracy, and 65% of recommendations with estimated low accuracy were likely correct.

Figure 4.4
A Video Feed Displaying its Level of Confidence or Uncertainty with a Constant-Hue, Varying-Salience Border

Figure 4.5
A Video Feed Displaying its Level of Confidence or Uncertainty Using a Red-Yellow-Green Color Scheme

Operational Tempo (Within-Subjects)

Operational tempo refers to the number of events in each time period. During low tempo periods, a UAV arrived at a new waypoint every 18 seconds and there was a flight deviation once every five minutes. In addition, participants were prompted to respond to one ATC

command each minute. A high tempo period was created by increasing the number of chat messages and flight deviations, as well as shortening the periods between waypoint classification prompts. During high tempo periods, new imagery at a waypoint had to be classified approximately every 14 seconds, there was one flight deviation every three minutes, and participants were prompted to respond to three ATC commands each minute. A prompt in the Mission Room chat room appeared every minute throughout both high and low operational tempo periods.

Representation Accuracy (Within-Subjects)

There were two five-minute periods during the last 15 minutes of the scenario where a machine overestimated the accuracy of its recommendations; a five-minute period of overestimated accuracy was followed by a five-minute period of valid accuracy, which was followed by a five-minute period of overestimated accuracy. Outside of these periods, the confidence and uncertainty representations were appropriately mapped (e.g. high confidence recommendations were correct 95% of the time).

Procedure

Participants were trained how to use the multi-UAV simulator and how to interpret the confidence/uncertainty representations, according to their randomly assigned group. The training included reviewing an instructional PowerPoint and completing a ten-minute training scenario that introduced participants to the study's tasks. At the end of training, participants completed a quiz via Google Forms that assessed their ability to recognize targets and interpret confidence/uncertainty information correctly; the quiz informed participants about any incorrect answers and presented the correct answers. The participant then took a five-minute break before the one-hour experiment began. Participants received $30 for their participation in the

cumulatively two-hour study. An additional $10 bonus was awarded to the participants with the first- and second-best performance scores in each group.

Dependent Measures

The dependent measures in this study included target classification performance, secondary task performance (chatrooms and monitoring for flight deviations), eye tracking metrics, self-reported scores of trust and one's own monitoring capabilities, and a debrief questionnaire. Primary and secondary task performance (response time and accuracy) was recorded by the simulator. The commercially available eye tracker Tobii Pro Glasses 2 and the software Tobii Pro Lab measured gaze behavior and reported eye tracking metrics for each data collection session. A person's trust was also inferred based on a person's gaze behavior (Lu & Sarter, 2019).

At seven points during the study, the simulator momentarily paused and participants were prompted to rate their agreement with the following statements on a scale of 1 (low) to 10 (high):

1. I have enough time to monitor all UAVs as much as I desire
2. I trust the high confidence recommendations
3. I trust the low confidence recommendations

The prompt wording was revised for participants who were provided classification recommendations with uncertainty displays (e.g. "I trust the low uncertainty recommendations"). Participants in the baseline group, who were not shown any representations of confidence or uncertainty, were simply asked whether they generally trusted the classification recommendations in lieu of the final two questions. The prompts occurred 10, 20, 40, 45, 50, 55, and 60 minutes into the simulation scenario, corresponding to events when either the operational tempo or the representational accuracy changed.

Results

The impact of confidence and uncertainty representations on trust calibration, attention management, and performance was evaluated in two stages. During the first stage, the potential for confounds in subsequent analyses was eliminated. Specifically, the data for participants who ignored a task was excluded from later analyses, as their neglect of a task would affect attention management and performance across the entire multitask testbed. Also, the study reported in Chapter 2 found that gaze behavior and response times were affected by some participants visually inspecting a video feed prior to a UAV arriving at a waypoint and prior to a recommended classification being displayed (this behavior was referred to as "previewing"). As described in more detail below, analyses were conducted during this first stage to evaluate the extent of previewing in the present study and identify whether subsequent analyses needed to exclude the corresponding data.

During the second stage, the impact of the two types of visual representations of system confidence and uncertainty information on trust specificity and attention management was analyzed. This analysis followed Parasuraman, Sheridan, & Wickens' (2000) four-stage model of human information processing which includes 1) information acquisition; 2) information analysis; 3) decision and action selection; and 4) action implementation. Attention capture (Expectation 1) was examined first, followed by assessments of how participants analyzed the confidence and uncertainty representations (Expectations 2 and 3). Since it was anticipated that (the validity of) representations of confidence or uncertainty would impact whether a person would decide to review a UAV's recommendation, video feed illumination and monitoring was

then evaluated (Expectation 4). Finally, classification performance was assessed (Expectations 2 and 3).

The statistics software R, and the *lme4*, (Bates et al., 2015) *lmerTest* (Kuznetsova et al., 2017), *multcomp* (Hothorn, Bretz, & Westfall, 2008), and *rstatix* (Kassambara, 2021) packages were used to analyze the data. Participants were modeled as random effects in the linear mixed models and generalized linear mixed models to distinguish between variation that was contributed by fixed effects (e.g. representation of confidence information) and variation attributed to an individual person's general ability to multi-task and cope with high attention demands (i.e. their attentional capacity).

Task Adherence

Since one aim of this study was to assess confidence and uncertainty representations in a multi-tasking environment with high attention demands, data for eight participants were excluded from the analysis because they ignored one of their tasks. This was done to reduce a potential confound where better monitoring and performance on one task might simply be due to fewer attention demands. Some of these participants may not have intended to neglect tasks but were compelled to do so because the remaining tasks imposed attention demands beyond their capacity. This is reflected by participants' responses to the statement, "I have enough time to monitor all UAVs as much as I desire," after the simulation's first high tempo period. Participants rated their agreement on a scale of 1 (low) to 10 (high), and an independent samples t-test was used to compare the group participants who had adhered to all the study's tasks and the group of participants who had neglected one of the study's tasks. Participants who neglected one of the study's tasks rated their agreement 2.3 points higher than participants who attended to all of the study's tasks ($M = 7.9$, $t(9.65) = 3.92$, 95% CI [1.0 3.7], $p = 0.003$). Participants who

attended to all tasks, albeit poorly, were still included in the analysis. Ultimately, the analyses included data from eight participants in the group that was not provided confidence or uncertainty information, twelve participants in each of the confidence information groups, and ten participants in each of the uncertainty information groups.

Previewing and Monitoring

This dissertation's first study found that, ignoring their instructions, participants began inspecting a UAV's video feed as the vehicle approached its waypoint and before the recommended classification was provided. While this is an effective strategy for sampling in an environment with high attention demands (Wickens & Hollands, 2000), this behavior leads to participants mentally classifying the imagery without support of the UAV. To determine whether similar behavior occurred in this experiment, a generalized linear model compared the number of waypoints previewed by the baseline and four experimental groups. Pairwise comparisons found that participants in the baseline group previewed more waypoints than any other group (see Figure 4.6 and Table 4.1) and participants who were presented hue-based representations of uncertainty previewed waypoints significantly less than all groups (see Table 4.2).

Figure 4.6

Number of Waypoint Images Previewed as a Function of Display Group

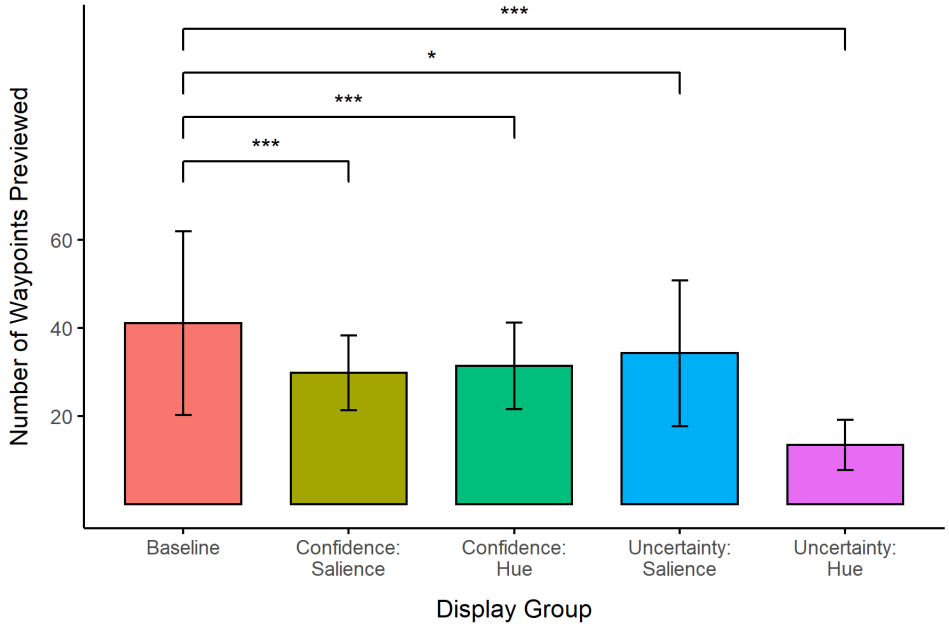


Table 4.1

Pairwise Comparisons of the Previewing Generalized Linear Model to the Baseline Group

	exp(Estimate)	*z*	*p*	95% CI
Confidence: Salience	0.73	-4.20	< .001	[0.62 0.84]
Confidence: Hue	0.76	-3.57	< .001	[0.66 0.88]
Uncertainty: Salience	0.83	-2.35	0.019	[0.72 0.97]
Uncertainty: Hue	0.32	-10.90	< .001	[0.27 0.40]

Table 4.2

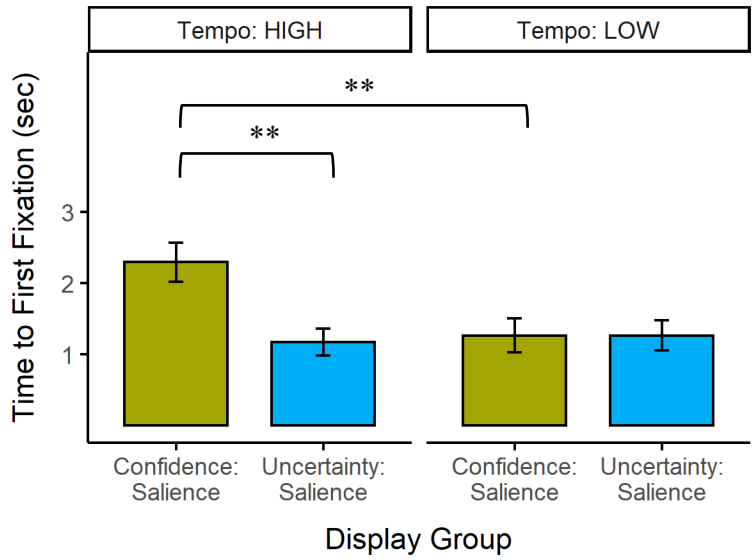
Pairwise Comparisons of the Previewing Generalized Linear Model to the Hue-Based Uncertainty Group

	exp(Estimate)	*z*	*p*	95% CI
Baseline	3.05	10.90	< .001	[2.50 3.73]
Confidence: Salience	2.21	7.85	< .001	[1.81 2.70]
Confidence: Hue	2.33	8.42	< .001	[1.92 2.84]
Uncertainty: Salience	2.54	9.18	< .001	[2.09 3.11]

Attention Capture

Eye tracking data were examined to assess whether salience-based visualizations of uncertainty captured attention faster than salience-based visualizations of confidence when a recommendation's estimated accuracy was low (Expectation 1). A linear mixed model analyzed the impact of representing confidence or uncertainty recommendations with a salience-based visualization method, operational tempo, and their interaction as fixed effects; the participant was modeled as a random effect (see Figure 4.7). During high tempo periods, salience-based uncertainty representations captured attention, on average, 1.145 seconds faster than salience-based confidence representations ($t(54.64) = -3.14$, 95% CI [-1.882 -0.423], $p = 0.003$). Salience-based confidence representations also captured attention 1.040 seconds faster in low tempo periods than in high tempo periods ($t(164.77) = -3.33$, 95% CI [-1.656 -0.424], $p = 0.001$). There was no significant difference in attention capture during low tempo periods and for salience-based uncertainty representation between tempos.

Figure 4.7
Attention Capture of Salience-Based Representations of Low Estimated Accuracy as a Function of Framing and Tempo



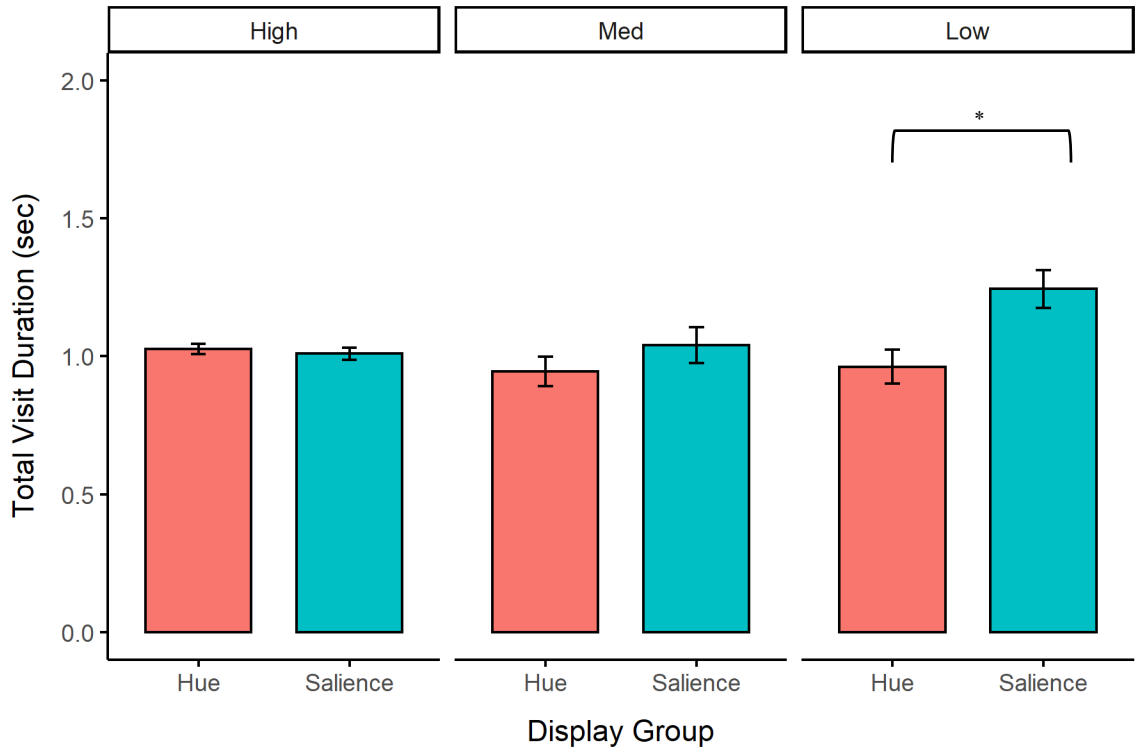
Interpretations of Confidence and Uncertainty Representations

Gaze Behavior While Evaluating Representations

A linear mixed model analysis evaluated the impact of salience-based and hue-based representations of a UAV's estimated accuracy on total visit duration (i.e. the time a participant gazed at the representation before choosing to illuminate the video feed; Expectation 2). The representation method (hue-based or salience-based representation) and the estimated accuracy (i.e. high, medium, or low) were fixed effects and the participant was a random effect. As shown in Figure 4.8, salience-based representations led to significantly longer gaze durations of the border only when estimated accuracy was low ($M = 0.246$, $t(191.19) = 2.22$, 95% CI [0.027 0.464], $p = 0.028$).

Figure 4.8

Total Visit Duration to Evaluate a Representation of a UAV's Estimated Accuracy as a Function of Representation Method and the Level the Estimated Accuracy



Intuitiveness Rating

To contextualize the results of Expectations 2 and 3 (i.e. the impact of framing and the representation method of a machine's estimated accuracy on attention management, trust calibration, and performance), the debrief questionnaire asked participants to rate their agreement with the statement, "the mapping of colors to confidence [or uncertainty (depending on the group)] was intuitive." A score of one indicated strong disagreement and five was associated with strong agreement. A linear model ($F(3,42) = 5.04$, $p = 0.005$) and pairwise comparisons indicated that participants who were presented salience-based trust scores of uncertainty provided lower intuitiveness ratings than those in the other groups (M =3.2; see Figure 4.9 and Table 4.3) .

Figure 4.9
Intuitiveness Rating as a Function of Display Group

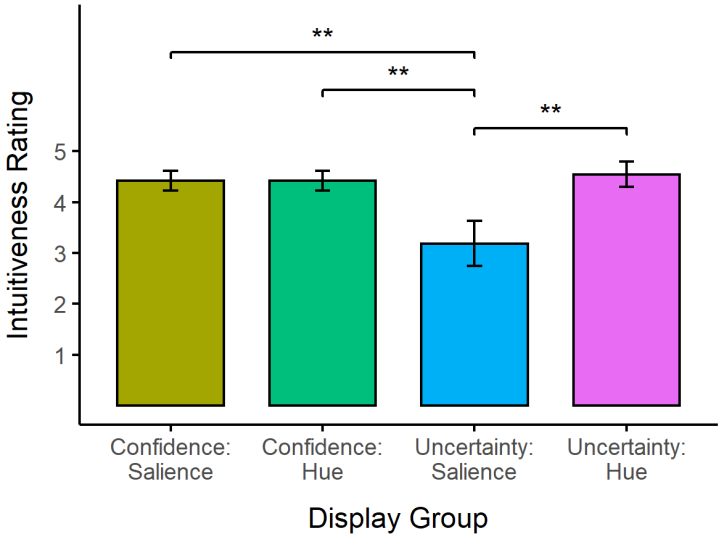


Table 4.3
Pairwise Comparisons of Intuitiveness Ratings to Salience-Based Representations of Uncertainty

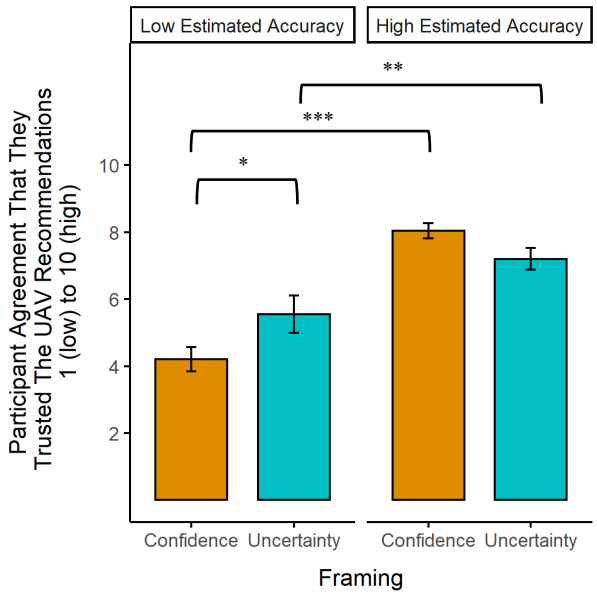
	Estimate (as compared to Uncertainty: Salience)	*t*	*p*	95% CI
Confidence: Salience	1.2	3.10	0.003	[0.4 2.0]
Confidence: Hue	1.2	3.10	0.003	[0.4 2.0]
Uncertainty: Hue	1.4	3.35	0.002	[0.54 2.2]

How Framing Modulated Trust in High and Low Confidence and Uncertainty Representations

It was expected (Expectation 3) that uncertainty information would initially cause participants to report lower trust in the automation than participants who were provided confidence information. A linear mixed model assessed the impact of framing and whether imagery was classified with a high or low estimated accuracy on the first trust scores provided by participants; the model included an interaction, and each participant (participants in the baseline group were excluded) was treated as a random effect. As shown in Figure 4.10, high uncertainty recommendations initially received trust ratings that were 1.3 points higher ($t(84.98) = 2.57$, 95%

CI [0.3 2.4], $p = 0.018$) than low confidence recommendations ($M = 4.2$, $SE = 0.3$); there was not a significant difference between the initial trust ratings for high confidence and low uncertainty recommendations. The model also found that participants rated their trust higher for classification recommendations with high estimated accuracy than low estimated accuracy for both confidence ($M = 3.8$ points higher, $t(44) = 8.56$, 95% CI [2.9 4.7], $p < 0.001$) and uncertainty ($M = 1.7$ points higher, $t(44) = 3.36$, 95% CI [0.7 2.6], $p = 0.001$).

Figure 4.10
The Impact of Framing on Initial Self-Reported Trust Ratings



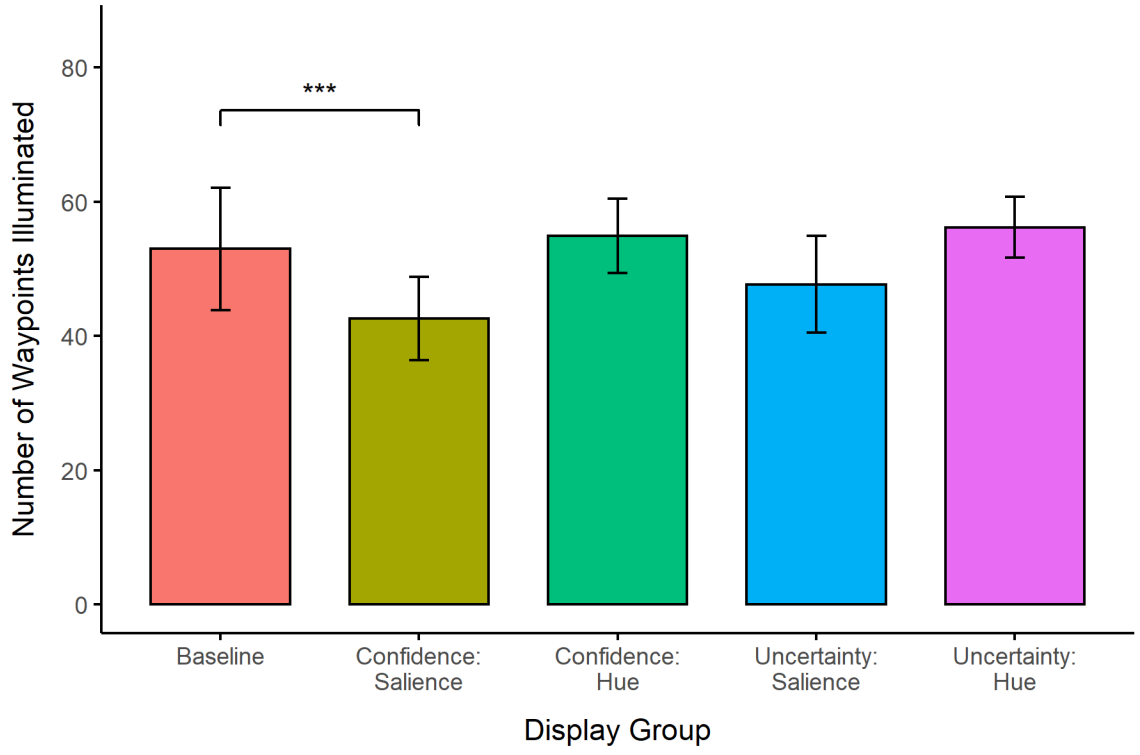
Video Feed Monitoring

Video Feed Illumination and Representation Methods

A generalized linear model analyzed how the display of confidence or uncertainty information (i.e. Expectation 3) influenced the number of waypoints illuminated by a participant (prior to the final 15 minutes of the scenario when the UAVs made incorrect estimates of their classification accuracies). Waypoints that were previewed (i.e. illumination was not determined

based on the estimated accuracy) or unclassified by the participant were excluded from the analysis. Pairwise comparisons found that only salience-based representations of high-confidence led to significantly fewer waypoints being illuminated than the number of waypoints illuminated by the baseline group (M = 26%, z = -3.90, 95% CI [12% 32%], $p < 0.001$). There were no statistically significant differences in video feed illumination between the baseline group and other representations of high confidence or low uncertainty, as shown in Figure 4.11. Furthermore, there were no statistically significant differences in video feed illumination between the baseline group and any representation of *low* confidence or *high* uncertainty.

Figure 4.11
Number of Waypoints that were Estimated to have High Accuracy and were Illuminated as a Function of Display Group



Reviewing Recommended Classifications

Eye tracking data were analyzed to determine how framing impacted a participant's scanning behavior when reviewing video feed imagery (Expectation 3). A linear mixed model analysis evaluated gaze duration for a video feed after it was illuminated (and the imagery could be seen clearly), and a generalized linear mixed model was used to analyze the fixation count within the video feed. Framing was modeled as a fixed effect and the participants and waypoints were modeled as random effects. By analyzing the waypoints as a random effect, the models differentiated between effects of framing and effects that were unique to each waypoint (e.g. imagery discernability, operational tempo, estimated accuracy of recommendation).

Overall, images that were classified correctly by the UAV and reviewed by the participant had a longer total visit duration with marginal significance ($M = 0.243$, $t(43.90) = 1.88$, 95% CI [-0.015 0.499], $p = 0.066$) in case of uncertainty (as opposed to confidence) framing; the difference in fixations was not significant. Framing did not have a significant impact on the total visit duration or fixation count for targets that were classified *incorrectly* by the UAV.

Mismatch Between Estimated and Actual System Accuracy

Forty minutes into the scenario, there was a five-minute high tempo period where valid accuracy estimates were provided, followed by a second high tempo period where the UAVs overestimated the accuracy of their classifications. Next, there was a five-minute *low* tempo period with valid accuracy estimates, followed by a second low tempo period where the UAVs again overestimated the accuracy of their classifications. It was expected (i.e. Expectation 4) that participants who were presented uncertainty recommendations would better adjust their monitoring and trust in the multi-UAV system during periods with invalid accuracy estimates,

compared to participants who were presented confidence information. Linear mixed models compared how framing (i.e. confidence or uncertainty) and the validity of accuracy estimates impacted trust scores (Figure 4.12) and gaze behavior (Figure 4.13). Only scenes that were classified by the UAVs correctly were included in this analysis, since incorrect classifications have been shown to have a bottom-up effect on gaze behavior, confounding whether longer visit durations are due to less trust or a mismatch between actual and expected imagery (see Chapter 2).

Figure 4.12
Gaze Behavior as a Function of Framing, Tempo, and Estimated Accuracy

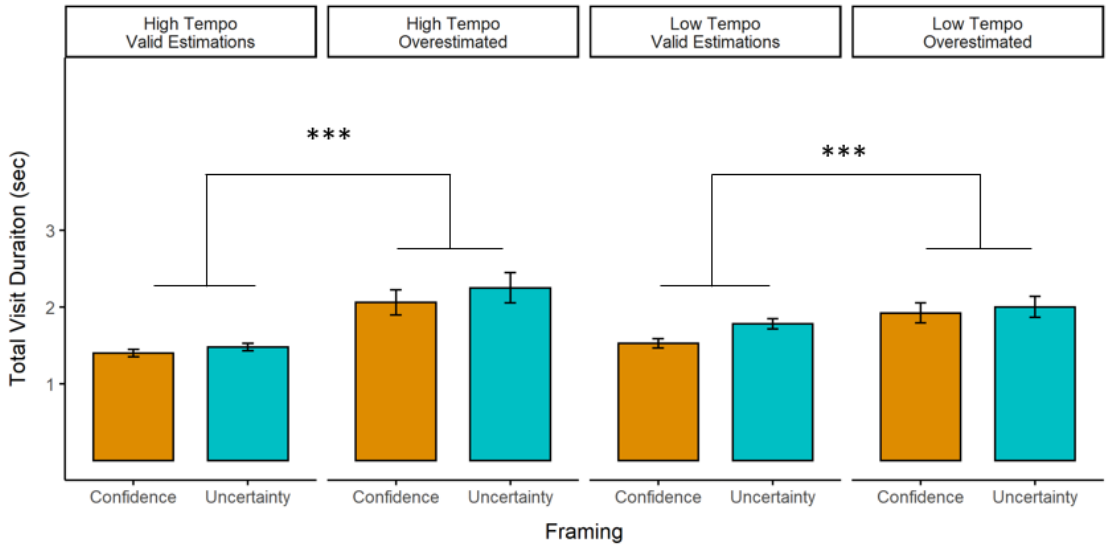
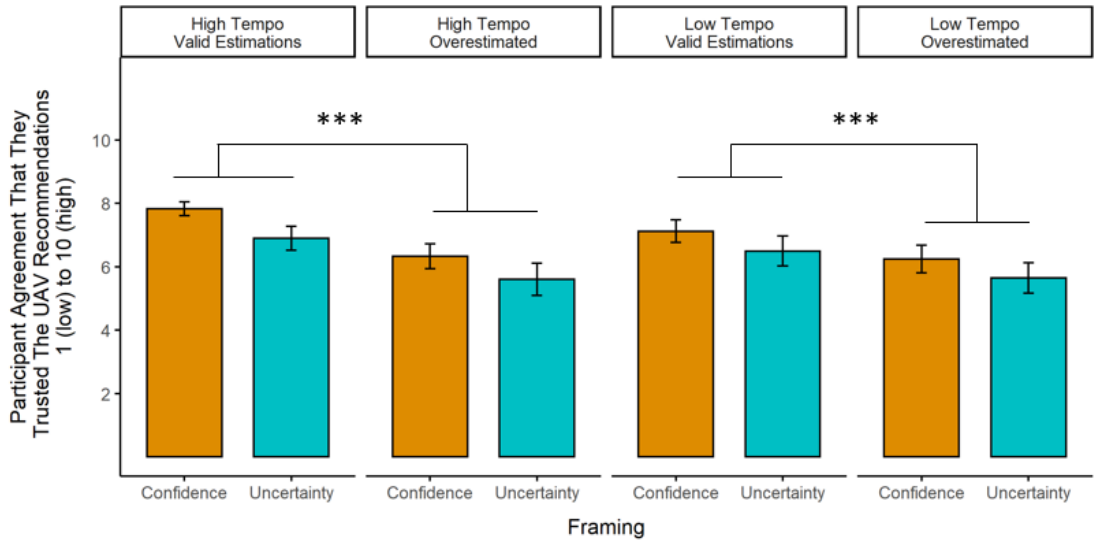


Figure 4.13

Trust Rating as a Function of Framing, Tempo, and Estimated Accuracy



Likelihood ratio tests found that framing did not significantly affect a participant's gaze behavior or trust scores when UAVs overestimated the accuracy of their classifications during high tempo operations (the first two periods in Figure 4.12 and Figure 4.13). However, invalid accuracy estimates led to a significant increase in the total visit durations to analyze the video feed imagery ($M = 0.706$, $t(650.42) = 7.57$, 95% CI [0.523 0.890], $p < 0.001$) and a significant decrease in trust scores ($M = -1.4$, $t(44) = -5.74$, 95% CI [-1.9 -0.9], $p < 0.001$).

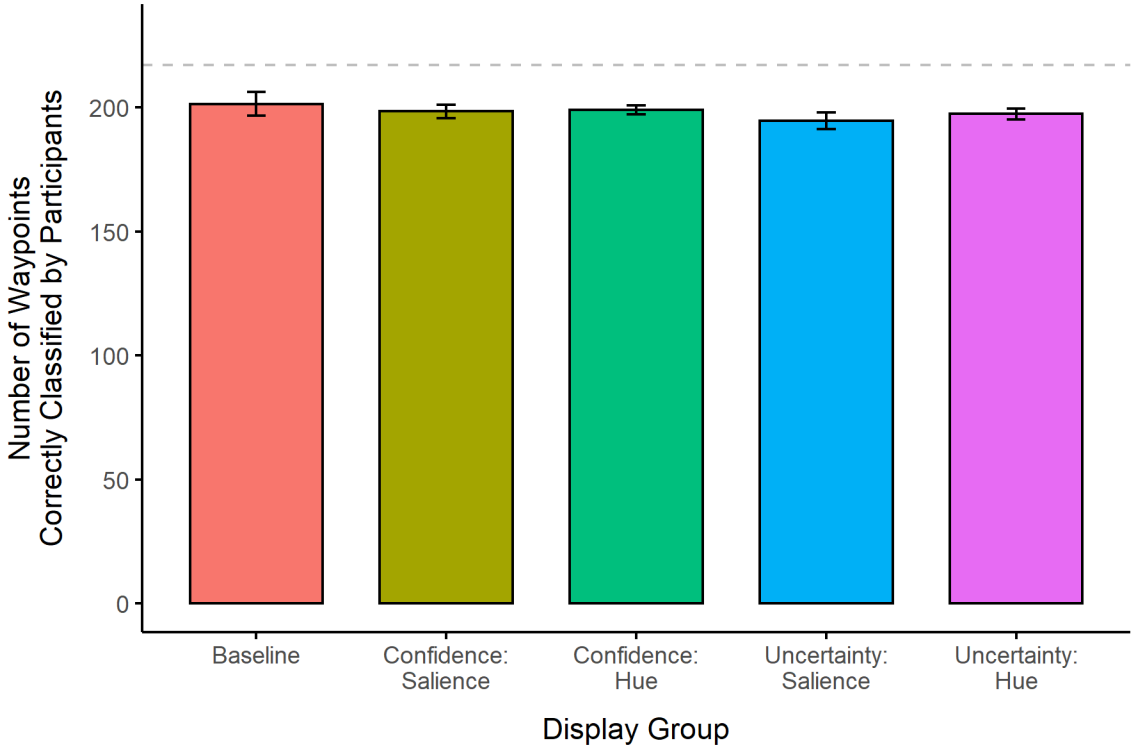
Likelihood ratio tests also did not find significant differences in how framing impacted gaze behavior and trust scores during the two final *low* tempo periods when the UAVs began to overestimate the accuracy in their classifications (the third and fourth periods in Figure 4.12 and Figure 4.13). Overall, participants gazed for longer durations at the imagery ($M = 0.310$, $t(554.21) = 3.64$, 95% CI [0.143 0.478], $p < 0.001$) and provided lower trust scores ($M = -0.9$, $t(44) = -4.636$, 95% CI [-1.2 -0.5], $p < 0.001$) when the UAVs overestimated their classification accuracies in the low tempo periods.

Performance

A likelihood ratio test indicated that a participant's display group did not significantly impact the number of waypoints that the participant ultimately classified correctly. Figure 4.14 suggests that a ceiling effect may have been observed; on average, participants classified 91.2% of waypoints correctly during the hour-long scenario.

Figure 4.14
Waypoint Classification Accuracy as a Function of Display Group

The dashed gray line represents the total number of classification tasks in a scenario.



Linear mixed models assessed the extent that display design impacted classification times; the display group was a fixed effect and the waypoints were random effects. Classification times were measured by the simulator as the total time elapsed from when a UAV arrived at a waypoint and displayed its recommended classification, to when the participant provided their

classification of the video feed's imagery. Therefore, classification times included both the amount of time that participants spent attending to the primary classification task, as well as the time elapsed while the participant attended to other tasks. It is worth noting that classification times may have been affected by the congruence between the expected and actual targets, the estimated accuracy's magnitude and validity, the operational tempo, and previewing. Therefore, the focused analysis evaluated classifications that were performed when there were greater attention demands and performance was more critical (i.e. waypoints during high operational tempo periods that were not previewed). Since there were relatively fewer instances when the UAVs had medium or low estimated accuracy levels, and even fewer cases when the UAVs provided wrong classifications, the analysis was restricted to the larger sample size of correct classifications that were highly estimated to be accurate. As shown in Figure 4.15 and the pairwise comparisons in Table 4.4, hue-based uncertainty representations led to longer classification times than the other visualization methods, but the effect was small.

Figure 4.15
Waypoint Classification Time as a Function of Display Group.

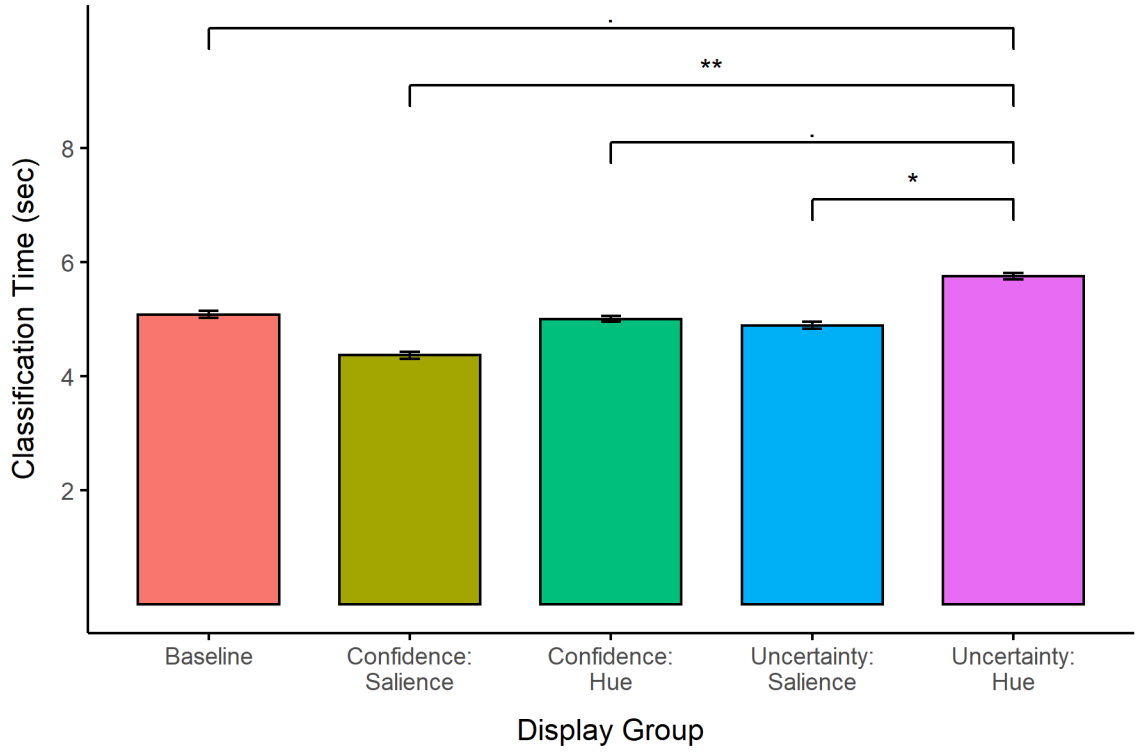


Table 4.4
Pairwise Comparisons of Classification Times to the Hue-Based Uncertainty Group

	Estimate	*t*	*df*	*p*	95% CI	*d*
Baseline	-0.932	-1.88	51.56	0.065	[-1.920 0.055]	0.078
Confidence: Salience	-1.375	-3.08	51.22	0.003	[-2.265 -0.485]	0.111
Confidence: Hue	-0.863	-1.93	51.25	0.058	[-1.754 0.026]	0.070
Uncertainty: Salience	-0.983	-2.11	51.50	0.039	[-1.915 -0.053]	0.081

Secondary task performance was evaluated by comparing the response times to flight deviations and chat messages between display groups. Since the operational tempo may have impacted response times, these analyses focused on high tempo operations when performance maintenance was more critical. A linear mixed model found that salience-based representations of confidence (M = -148.80, t(44.87) = -2.60, 95% CI [-263.69 -34.81], p = 0.012), hue-based

representations of confidence (M = -144.89, t(44.87) = -2.54, 95% CI [-259.78 -30.89], p = 0.015), and salience-based representations of uncertainty (M = -124.48, t(44.82) = -2.05, 95% CI [-2642.73 -3.13], p = 0.047) had significantly shorter response times to flight deviations than the baseline group (see Figure 4.16). The number of correct responses to ATC chat messages during high tempo periods as a function of display group is shown in Figure 4.17. A generalized linear model analysis found that only salience-based representations of uncertainty had a significantly different quantity (14.1% fewer) of correct responses than the baseline group (z = -2.24, p = 0.025).

Figure 4.16
Flight Deviation Response Time as Function of Display Group

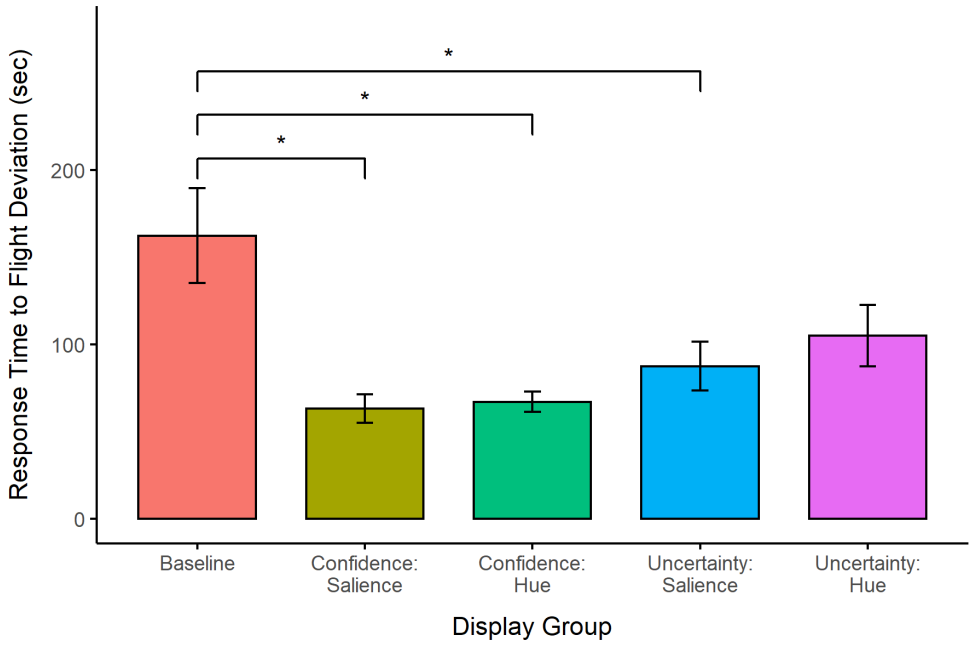
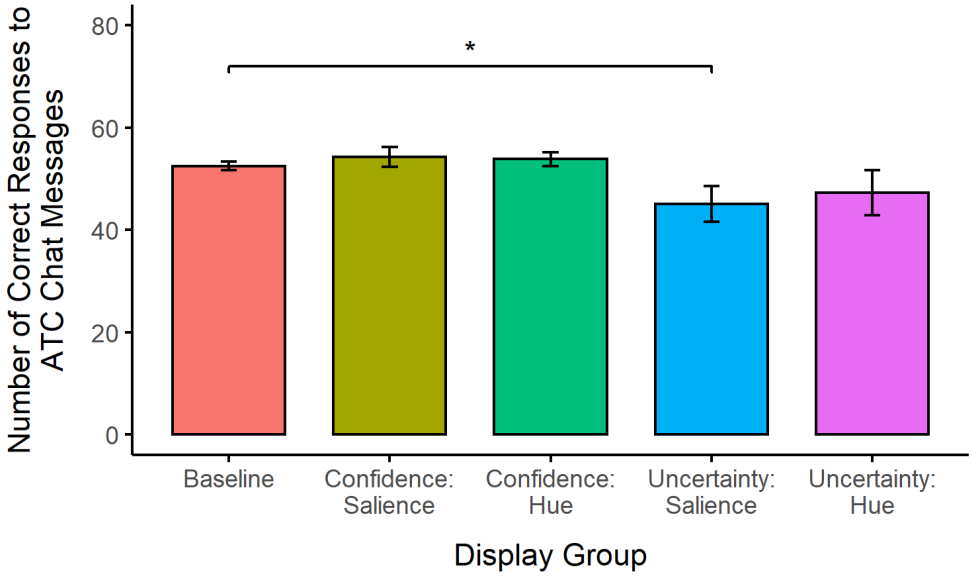


Figure 4.17

Number of Correct Responses to ATC Chat Messages in High Tempo Operations as a Function of Display Group



Summary of Findings

This study assessed the impact of two different visual representations (salience- versus hue-based) and two types of framing (confidence versus uncertainty) on trust calibration, attention management, and performance. Table 4.5 presents a summary of the study's main expectations and findings.

Table 4.5
Summary of Expectations and Findings

Legend: ✔ Supports Expectation ✘ Contradicts expectation • Partially, but not fully, supports expectation

Expectations	**Findings**	**Supported?**
1 - Additive, salience-based representations of uncertainty, as opposed to salience-based representations of confidence, will capture attention faster and more reliably when the estimated accuracy of a classification is low	✔ Salience-based representations of high uncertainty captured attention significantly faster than salience-based representations of low confidence in high tempo periods (but not low tempo periods).	(✔)
2 - Participants will find it easier to distinguish between levels of confidence and uncertainty with hue-based representations (compared to salience-based)	✔ Participants spent significantly longer gazing at salience-based (as opposed to hue-based) representations of high uncertainty and low confidence before choosing to illuminate a video feed. ✔ Participants reported that salience-based representations of uncertainty were less intuitive than other visualization methods.	(✔)

3 - Uncertainty framing will cause a participant to adopt more risk averse behavior, such as more closely monitoring and scrutinizing a UAV's recommendations. This will lead to slower classification times, initially lower trust scores, better classification accuracy, and worse performance on the mission's secondary tasks (compared to confidence framing).	✔ Representations of uncertainty were associated with longer video feed gaze durations when UAVs classified imagery correctly. ✔ Hue-based representations of uncertainty had slower classification times than the other visualization methods. ✘ Representations of low confidence led to significantly greater trust decrements than high uncertainty. ✘ There were no significant differences in the number of waypoints illuminated between the confidence and uncertainty framing groups. • Participants had significantly greater initial trust in waypoints that were classified with a higher estimated accuracy than lower estimated accuracy.	(•)
4 - Displays of uncertainty will make it more likely that participants notice a mismatch between a system's true and estimated accuracy and subsequently adjust their monitoring and trust.	• Framing did not lead to significant differences in gaze behavior or trust scores when UAVs overestimated their accuracy. However, participants gazed longer at the video feeds and provided lower trust scores during these overestimation periods.	(•)

Discussion

This study assesses the impact of a system providing real-time feedback regarding its confidence in or uncertainty about its own performance on human operators' trust calibration, attention management, and joint system performance. Specifically, the framing of a machine's estimated accuracy as confidence or uncertainty, the method for visualizing the estimated accuracy level, and the impact of temporary mismatches between a machine's estimated and actual accuracy were evaluated.

Representing Confidence and Uncertainty

During high-tempo operations, when a UAV's estimated classification accuracy was low, salience-based representations of uncertainty (as compared to confidence) captured participants' attention significantly faster. This partially supports our first expectation and may be explained by the fact that, when accuracy was low, uncertainty representations were brightest (i.e. high uncertainty) whereas confidence representations were at their lowest salience level. In contrast, during low-tempo operations, attention capture rates did not differ between the two types of framing, confidence or uncertainty. The latter finding qualifies past research conducted by Bisantz et al. (2009) on additive representations and reinforces the need for researchers to contextually evaluate confidence and uncertainty representations.

While salience-based representations of uncertainty had the advantage of leading to faster attention capture during high-tempo periods, participants considered them to be less intuitive than all the other representations of a system's estimated accuracy in this study. Two participants reported that they had trouble making the absolute judgments required to distinguish between the different salience levels (and corresponding levels of uncertainty). Another two participants expressed that, intuitively, they would have expected greater brightness and salience to

correspond with greater certainty, rather than uncertainty. Thus, additive representations of uncertainty may have been confusing to participants because an increase in one dimension – salience– correlated with a decrease in another dimension – certainty or confidence. Unaware that participants in other groups were presented a hue-based representation of uncertainty, two participants suggested that a red-yellow-green representation of uncertainty would have been more intuitive.

Difficulties with distinguishing salience levels and the unintuitive mapping of salience to uncertainty may explain why participants gazed at additive representations of low accuracy for significantly longer before illuminating a video feed, compared to substitutive hue-based representations. The above findings highlight that, while unfamiliar hue-based palettes may not map naturally to the additive dimension of uncertainty (Bisantz et al., 2009), the red-yellow-green palette employed in this study was effective likely due to the widely familiar symbolic meaning of these hues (e.g., their use in alarm design).

Confidence and Uncertainty Framing

Representations of a machine's confidence or uncertainty were both effective methods for supporting trust specificity. Prior to gaining substantial experience with the simulator, participants at the beginning of the study provided higher trust ratings for recommended classifications that were associated with greater estimated accuracy (i.e. high confidence, low uncertainty) and lower trust ratings for recommended classifications that had a lower estimated accuracy. However, in contrast to Expectation 3, participants provided lower trust scores for low confidence recommendations than equivalent high uncertainty recommendations. Participants were informed during training that recommendations with a low estimated accuracy were correct for 65% of classifications. This should have resulted in a trust score of approximately 6.5.

However, low confidence representations received a mean initial trust score of 4.2, and high uncertainty representations had a mean initial trust score of 5.5. One explanation for this finding might be that, because a confidence orientation focuses on gains, performance decrements ("losses") were perceived more acutely by operators, whereas operators presented with an uncertainty framing expected the machine to be imperfect and their trust was both better calibrated and more resilient to a machine's lower capabilities. This finding may also be explained by results from a study conducted by Yang, Wickens, & Hölttä-Otto (2016) which concluded that the expected validity of automated recommendations may influence the magnitude of trust decrements. Since recommendations presented with high uncertainty were anticipated to be incorrect, they may not have triggered negative trust feedback loops to the same extent as unexpected violations of confidence. Yang et al. (2016) connected their findings to prospect theory (Kahneman & Tversky, 1979) which posits that losses are assessed more negatively than gains.

Eye tracking data support our expectation that participants would gaze at a video feed for a longer period of time when presented with uncertainty representations, compared to confidence representations. This suggests that the uncertainty framing indeed promoted a risk-averse mindset and caused participants to review a UAV's recommendations more closely. The UAV's estimated accuracy level (i.e. high confidence versus low confidence) did not affect how long participants reviewed the video feed. While trust scores were sensitive to changes in estimated accuracy, the difference between the three reliability levels (15% each) may not have been sufficient to affect participants' reliance and monitoring behavior. In comparison, Lu and Sarter (2019) reported longer gaze durations correlated with reliability decrements in their study of a system that changed from being 95% reliable to 50% reliable.

Salience-based high confidence indications were the only type of representation that led to greater reliance on UAV classifications than the baseline condition. As shown in Figure 4.11, participants in this group illuminated the video feeds less frequently to check if the recommended high confidence classifications were correct. One explanation might be that the other display groups drew attention to the UAVs' classification capabilities being imperfect and needing to be monitored, whereas the salience-based confidence representations de-emphasized the likelihood that interventions may be necessary. For example, the hue-based confidence and uncertainty representations used a red border to indicate when monitoring was needed, and the salience-based uncertainty recommendations were brightest when monitoring would have been beneficial too. In comparison, salience-based representations of confidence were least salient when monitoring would have been beneficial. This may have led to higher expectations of reliability, and greater trust (or even overtrust) for salience-based confidence representations.

Overestimation of Accuracy

When UAVs overestimated their performance, high confidence and low uncertainty recommendations dropped in accuracy from 95% to approximately 30%. Participants noticed this drop in performance, lowered their trust in the multi-UAV system's classifications and gazed at the video feed imagery for longer periods of time. This significant difference in gaze behavior suggests that eye tracking may be better suited for assessing trust resolution than trust specificity; large changes in capabilities led to significant changes in gaze behavior, but the incremental 15% difference between each of the three reliability levels (e.g. low, medium and high confidence) did not lead to significant changes in monitoring behavior (see also Lu and Sarter (2019) and the study described in Chapter 2).

While there were no significant differences in trust score decrements and gaze behavior between confidence and uncertainty representations when UAVs overestimated their accuracies, gaze durations (but not trust scores) in the uncertainty group were slower to return to their initial value when the UAVs accuracy returned to 95% (see Figure 4.12). When participants encountered a second period of overestimated accuracies, they lowered their trust scores and increased their monitoring again. However, the magnitude of these changes was smaller than during the first period of incorrect accuracy estimates. The total visit durations and trust ratings shown in the final period of Figure 4.12 and Figure 4.13 suggest that participants approached an average level of trust and monitoring as they were increasingly exposed to periods of lower system capabilities and learned how to manage both the increased monitoring needs while still attending to the secondary task demands.

Performance

There was no significant difference in participants' classification accuracy between groups. Note, however, that this may be due to a ceiling effect. During pilot testing in advance of the experiment, participants felt that the tasks were sufficiently difficult and that further increasing attention demands might cause them to neglect tasks. There was also a concern that increasing attention demands beyond a participant's capacities might compel them to comply with a UAV's recommended classifications because they had no choice due to excessive workload, rather than compliance reflecting their trust in the system.

While differences in classification accuracy were not observed, hue-based representations of uncertainty led to slower classification times during high tempo operations, compared to all other display groups (including the baseline). This was a small effect, and it may also have been a cumulative effect. First, participants who were presented with hue-based representations of

uncertainty tended to illuminate the video feeds at more waypoints than the salience-based representations groups (see Figure 4.10). This finding may be explained by the fact that salience-based uncertainty representations were considered the least intuitive, thus potentially leading to misinterpretations of system accuracy. With salience-based *confidence* representations, there was a negative correlation between the salience of the signal and the need for monitoring (i.e. high salience equals high confidence implying little need for reviewing the video feed). Second, participants gazed at video feeds for significantly longer durations when presented with an *uncertainty* (versus confidence) framing, perhaps due to its orientation to the potential for loss and violations of a machine's anticipated accuracy. This highlights a tradeoff between the different designs. On the positive side, hue-based representations of uncertainty were considered more intuitive by participants, and they led to closer supervision of the UAVs. However, this resulted in slower classifications. In a low-tempo environment, or in a context where inaccurate classifications can have disastrous consequences, the relatively minor delay of 1-1.5 seconds for hue-based representations (see Table 4.4) may make them the design of choice.

There are indications that participants who were provided information about the UAVs confidence or uncertainty were able to better manage attention demands, compared to the baseline. Participants in the latter group took more time to respond to the secondary task's flight deviations and previewed waypoints to a significantly greater extent, perhaps because they needed to temporarily spread out the demands on their attention since they were not provided confidence or uncertainty representations to guide their monitoring.

Conclusion

Findings from this study confirm that representing a system's confidence or uncertainty in its recommendations is an effective technique to support trust specificity. Overall, representing

levels of uncertainty with a familiar red, yellow, and green palette were found to best support trust specificity and attention management. They were also interpreted faster by participants than salience-based representations. Additive, salience-based representations of confidence captured attention more slowly than salience-based representations of uncertainty when attention demands were high, but salience-based representations of uncertainty were counterintuitive to participants. Uncertainty representations prompted operators to review imagery more closely. Furthermore, classifications that were recommended with high uncertainty corresponded with smaller trust decrements, and self-reported trust scores more closely approximated a machine's true accuracy, than when the same classifications were recommended with low confidence. However, while hue-based recommendations of uncertainty were found to improve trust specificity and attention management, it hurt performance by leading to slower task completion times.

There are two important design and research implications of the fact that representations of uncertainty that employed the hues red, yellow, and green best supported trust specificity and attention management. While past research (Bisantz et al., 2009) has suggested that designers should avoid using hue-based substitutive representations of uncertainty, an additive variable, this study demonstrates that, if a familiar hue-based color palette is used, these representations are actually intuitive and effective because they do not interfere with attention management. Future research evaluating additional candidate methods to represent confidence and uncertainty in human supervisory control might benefit from assessing candidate visualizations in an attention-demanding context with eye tracking. Furthermore, while the symbolic palette of red, yellow, and green is might be familiar to certain users, operators with color deficiencies (i.e. color blindness) might have trouble distinguishing the red and green hues. Additional symbolic color palettes that would be accessible to larger proportion of the population should be

considered and evaluated. Finally, modern neural networks predominantly qualify their outputs by reporting confidence, not uncertainty. Future work is needed to properly map systems that might internally estimate their accuracies using confidence to interfaces that display uncertainty if trust specificity, attention management, and performance are to be optimized.

References

Bisantz, A. M., Stone, R. T., Pfautz, J., Fouse, A., Farry, M., Roth, E. M., … Thomas, G. (2009). Visual Representations of Meta-Information. *Journal of Cognitive Engineering and Decision Making*, *3*(1), 67–91. https://doi.org/10.1518/155534309X433726

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Anlysis of Decision under Risk. *Econometrica*, *47*(2), 263–292.

Lu, Y., & Sarter, N. B. (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. https://doi.org/10.1109/THMS.2019.2930980

Norman, D. A. (2013). *The Design of Everyday Things*. Basic Books.

Sheridan, T. B. (2008). Risk, Human Error, and System Resilience: Fundamental Ideas. *Human Factors*, *50*(3), 418–426. https://doi.org/10.1518/001872008X250773

Tversky, A., & Kahneman, D. (1986). Rational Choice and the Framing of Decisions. *The Journal of Business*, *59*(4), S251–S278.

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., & Talleur, D. A. (2003). Attentional Models of Multitask Pilot Performance using Advanced Display Technology. *Human Factors*, *45*(3), 360–380. https://doi.org/10.4324/9781315092898-10

Wickens, C. D., & Hollands, J. G. (2000). *Engineering Psychology and Human Performance* (3rd ed.). Prentice-Hall.

Yang, X. J., Wickens, C. D., & Hölttä-Otto, K. (2016). How Users Adjust Trust in Automation: Contrast Effect and Hindsight Bias. *Proceedings of the Human Factors and Ergonomics Society*, 196–200. https://doi.org/10.1177/1541931213601044

Chapter 5

Conclusion

Modern, and envisioned, robots and increasingly autonomous machines operate in complex, partially-observable, and highly dynamic commercial and military environments (Mason, 2012; Parasuraman & Riley, 1997; Russell & Norvig, 2009; Woods et al., 2004). They are highly reliable but still imperfect as they often draw from ambiguous data or noisy sensors. Therefore, human supervisors are tasked with monitoring these machines and intervening when necessary to correct plans, decisions, and actions (Sheridan & Parasuraman, 2005; Sheridan & Verplank, 1978). Increasingly, one operator is tasked with supervising multiple machines simultaneously. This imposes considerable cognitive demands on humans whose limited attentional resources require them to decide, at any given point in time, whether to rely or check on a machine, often based on their level of trust in the system (Lee & See, 2004; Parasuraman & Riley, 1997).

Lee and See (2004) define trust as the "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Trust calibration refers to the proper mapping of an operator's trust to a machine's capabilities. Since machine capabilities can vary across tasks and operational contexts, two important aspects of trust calibration are trust resolution and trust specificity. Trust resolution refers to the mapping between an operator's trust level and the range of a system's capabilities across contexts (e.g. an increasingly autonomous vehicle operating in an environment with good versus poor visibility;

Cohen, Parasuraman, & Freeman, 1998; Lee & See, 2004). Trust specificity, on the other hand, describes the tracing of trust to momentary fluctuations in capabilities (Lee & See, 2004).

This dissertation examined the close and important relationship between trust calibration and attention management. A multi-UAV supervisory simulation assessed candidate training methods to support mental model development, trust resolution, and top-down attention management. A similar testbed was used to evaluate candidate visual and auditory representations of a machine's estimated accuracy in its recommendations and their impact on trust specificity and bottom-up attentional guidance. Eye tracking complemented traditional trust measures such as behavioral data and subjective ratings.

Specifically, the goals of this line of research were:

- To investigate and compare the impact of active, experiential training with more traditional forms of training on a person's mental model development, trust resolution, and attention management.
- To assess how various visual and multimodal representations of a machine's confidence in its own abilities, and the framing of a machine's estimated accuracy as confidence or uncertainty, shape trust specificity and support attention management and joint system performance.

The first experiment in Chapter 2 compared active and passive training methods that instructed participants "how the system worked" with training that taught participants only "how to work the system." Active, experiential training facilitated the development of a better mental model of the system (i.e. fewer gaps in their understanding of the system), which supported contextually-driven trust resolution and top-down attention allocation. Participants receiving

experiential training appropriately lowered their trust and monitored a UAV's health more closely when its environment (e.g. turbulence) reduced the UAV's capabilities.

The next two studies in this line of research investigated how real-time feedback on system confidence (or uncertainty) in its own performance affect operators' trust calibration and monitoring behavior. The experiment in Chapter 3 focused on how best to present system confidence information during high tempo UAV operations involving numerous competing attention demands. A comparison of visual and auditory representations of confidence indicated that a natural mapping of tonal pitch to a machine's confidence is intuitive and leads to effective allocation of attention. Visual confidence representations appear to interfere with attention management unless designed to be processed in peripheral vision; in contrast, an auditory display of confidence information supported time sharing with the participants' visually demanding tasks as evidenced by faster task completion times. These results are in line with Multiple Resource Theory (Wickens, 2008), which predicts better time sharing when information is distributed across sensory channels due to reduced competition for attentional resources.

Since it was not clear whether the findings from this experiment were the result of the specific implementation of visual confidence information (i.e., an additive representation with high confidence corresponding to a highly saturated border), different types of visual representations as well as framing of confidence (confidence versus uncertainty) were investigated in the final study presented in Chapter 4. Both confidence and uncertainty supported trust specificity in this study. In other words, high estimates of system accuracy (i.e. high confidence or low uncertainty) were associated with high trust ratings and low estimates of accuracy were associated with low trust ratings. An uncertainty framing of a machine's estimated accuracy (as compared to a confidence framing) led to both increased monitoring (without

sacrificing performance) and also to trust more closely approximating a UAV's true accuracy. Representing a machine's estimated accuracy with a familiar hue-based palette (i.e. the colors red, yellow, and green) was considered intuitive by participants who interpreted this representation faster than an additive, salience-based version.

The second and third studies investigated not only the interpretability of representations but also the impact of mappings on attention management. Because additive representations were less salient in case of low confidence, they were not as effective at capturing attention when supervision and intervention were needed the most. This finding is aligned with the N-SEEV model, which identifies salience as one signal feature that influences noticing (Steelman-Allen, McCarley, Wickens, Sebok, & Bzostek, 2009; Wickens et al., 2009). According to the model, a signal's static salience (e.g. color contrast) or dynamic salience (e.g. a transient stimulus) will affect whether and how fast it captures attention. Since the value of providing confidence and uncertainty information in the context of supervisory control materializes only if the indication is noticed in the first place, designers need to use mappings that accomplish the goal of attracting attention even if that means the mapping would traditionally not be considered 'natural' (such as high salience associated with low confidence).

Building on previous research, eye tracking was used in all three studies to evaluate attention management, infer trust (Hergeth, Lorenz, Vilimek, & Krems, 2016; Lu & Sarter, 2019), and explain behavior not captured by performance outcome variables alone. Our findings highlight the need for careful use and interpretation of eyetracking data as trust may not be the only factor that determines gaze behavior. For example, in the first study, participants quickly found an expected target (i.e. the target recommended by the machine) when it was present in an image, whereas they spent longer reviewing the image when the expected target was not there.

This indicates that gaze behavior is influenced not only by trust, but also by an attentional phenomenon known as "contingent orienting" (Folk et al., 1992). The third study analyzed imagery that was classified correctly by the UAVs (i.e. the stimulus was the expected target), and found that participants gazed longer at the imagery during periods when other images were consistently classified incorrectly (i.e. periods with an *overall* accuracy of 30% instead of 95%). By comparing gaze behavior of correctly classified imagery during periods of lower and higher reliability, it can be inferred that longer gaze durations were a top-down effect caused by changes in a person's trust.

Another limitation of using eye tracking to infer trust was observed in the third study which found that machine capability differences of 15% did not result in significant changes in monitoring behavior; specifically, there was no statistically significant difference in gaze durations between high confidence recommendations with a 95% accuracy rate and medium confidence recommendations with an 80% accuracy rate. However, when the machine overestimated its accuracy and high confidence recommendations dropped from an accuracy rate of 95% to 30%, then a noticeable difference in monitoring behavior was observed, which also corresponded to changes in trust scores. This finding qualifies that eye tracking may be an effective method to infer trust when there are large changes in a machine's capabilities, but small changes in machine capabilities and a supervisor's trust may not be detected via eye tracking metrics.

Finally, this line of research also provides evidence that participants may not provide trust ratings in a consistent manner or in ways that are anticipated by researchers. The first study found that – despite being told the definition of trust that was to be used in this study – many participants adopted different definitions of trust. While the intent was for participants to rate

their general trust in a UAV (to encompass all its capabilities), participants indicated at the end of the study that they were rating their trust in a machine's classification abilities (rather than also incorporating the UAV's health management capabilities). Some of the participants in the second study, which also asked participants to rate their trust in each of the system's eight UAVs, reported in the debrief questionnaire that they had trouble remembering the performance of each of the vehicles. These observations suggest that researchers must not rely on subjective ratings alone but should complement them with additional measures (such as eye tracking or performance-based measures) to further investigate trust in human-machine systems.

In conclusion, this research makes important theoretical and applied contributions to supporting trust calibration, attention management, and joint performance in supervisory control. It adds to the knowledge base on training by demonstrating the efficacy of active, experiential training on supporting the development of a person's mental model of a system, their trust resolution, and attention allocation over traditional training methods. This research project also bridges past research in the design of visual and multimodal confidence representations and uncertainty representations, and it assesses the effect of employing a natural mapping of confidence or uncertainty on attention. In an environment with high attention demands, this research suggests that auditory and symbolic, hue-based representations of a machine's estimated accuracy may effectively support trust calibration and attention allocation better than visual, salience-based representations. Furthermore, findings support that framing a machine's estimated accuracy as uncertainty, rather than confidence, may lead to better trust calibration and closer supervision of the system. Finally, this research makes significant contributions to the advancement of research methods to measure trust in human supervisory control by identifying ways to mitigate confounds in eye tracking measures and subjective ratings. Cumulatively, this

line of research contributes to the safe adoption of machines in supervisory control and human-machine teams, including multi-agent systems for transportation, defense, search and rescue, surveying, and agriculture.

Future Work

There are some exciting ways in which future work can build on the findings from this research. First, future research might assess how the number of confidence or uncertainty levels influences trust specificity, attention management, reliance, and performance. For visual indications, for example, a multi-task control room simulation study conducted by Zirk et al. (2020) found that four-level likelihood alarms, such as an alarm that used one of four colors to encode the relative likelihood of an undesired event, reduced false alarms and misses, compared to three-level visual alarms. Future studies might consider whether there is an optimum number of levels that should be incorporated into visual and auditory confidence and uncertainty representations. It is likely that too few levels might hurt trust resolution and specificity, but as the number of levels increase beyond a certain point, operators will find it hard to distinguish between degrees of brightness or saturation and associate tones with respective accuracy levels, especially if required to make absolute judgments.

Additionally, the second study demonstrated the general benefits of auditory representations of a machine's estimated accuracy (i.e. intuitive mapping and supporting time sharing in visually demanding environments). However, there has been little research and specific design guidance on the sonification of confidence or uncertainty. The present research evaluated only whether a person could distinguish between and associate two tones – a high pitch tone and a low pitch tone – with the machine's confidence level. Future research might explore the auditory analogs of additive and substitutive visual representations.

Multimodal representations of confidence and uncertainty might also be extended into the tactile domain. In the study reported in Chapter 3, auditory representations of confidence informed the participant if a UAV had classified a target with high or low confidence. However, the binary, auditory representation was limited in that it did not indicate *which* UAV had classified a target and might need closer review. Rather, using a multimodal approach, when the auditory representation was displayed, the video feed for the corresponding UAV brightened to provide a visual cue and indicate that auditory confidence information was associated with that UAV. However, past studies (Ferris & Sarter, 2008; Riggs & Sarter, 2019) have shown that tactile cues may be an effective method to both encode information such as confidence or uncertainty while, at the same time, guide a supervisor's attention to the corresponding UAV, without relying on the overburdened visual channel.

While this line of research studied how a person might calibrate their trust, supervision, and reliance based on receiving information about a machine's confidence in its performance, future work needs to consider the opposite case also, i.e., how sharing information about *a person's confidence in their own performance* with a machine might enable that system to adapt to better support joint system performance. For example, a machine might change the display of information or provide additional attention guidance if a person appears to have trouble locating a target or interpreting imagery that is captured in a foggy environment. Ultimately, a better understanding of and support for mutual adaptation and coordination will be critical for the success of human-machine teams.

Finally, this line of research demonstrated that trust calibration in human-machine teams might benefit from better mental models – a top-down influence on trust resolution that was examined in the first study – and the presentation of moment-to-moment information on the

trustworthiness of a system – a bottom-up influence on trust specificity that was the focus of the second and third studies. Future research might examine the degree of synergy and *combined* impact of these two interventions, possibly combined with longer term operational experience with a system.

References

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. *In Proc. Command and Control …*, 1–37. https://doi.org/10.1.1.90.2591

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary Covert Orienting Is Contingent on Attetional Control Settings. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(4), 1030–1044.

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: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lu, Y., & Sarter, N. B. (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. https://doi.org/10.1109/THMS.2019.2930980

Mason, M. T. (2012). Creation Myths: The Beginnings of Robotics Research. *IEEE Robotics and Automation Magazine*, *19*(2), 72–77. https://doi.org/10.1109/MRA.2012.2191437

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/https://doi.org/10.1518/001872097778543886

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson.

Sheridan, T. B., & Parasuraman, R. (2005). Human-Automation Interaction. *Reviews of Human Factors and Ergonomics*, 89–129.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators*. Cambridge.

Steelman-Allen, K. S., McCarley, J. S., Wickens, C., Sebok, A., & Bzostek, J. (2009). N-SEEV: A Computational Model of Attention and Noticing. *Proceedings of the Human Factors and Ergonomics Society*, *2*, 774–778. https://doi.org/10.1518/107118109x12524442637381

Wickens, C. D. (2008). Multiple Resources and Mental Workload. *Human Factors*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wickens, C. D., McCarley, J., Steelman-Allen, K., Sebok, A., Bzostek, J., & Sart. (2009). NT-SEEV: A Model of Attention Capture and Noticing on the Flight Deck. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (pp. 769–773).

Woods, D. D., Tittle, J., Feil, M., & Roesler, A. (2004). Envisioning Human-Robot Coordination in Future Operations. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, *34*(2), 210–218. https://doi.org/10.1109/TSMCC.2004.826272