Spring 2020

Effects of Transparency and Haze on Trust and Performance During a Full Motion Video Analysis Task

Sarah C. Leibner
Old Dominion University, sarah.c.leibner@gmail.com

EFFECTS OF TRANSPARENCY AND HAZE ON TRUST AND PERFORMANCE DURING A FULL MOTION VIDEO ANALYSIS TASK

by

Sarah C. Leibner
B.S. Psychology, May 2017, Old Dominion University

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

PSYCHOLOGY

OLD DOMINION UNIVERSITY
May 2020

Approved by:

James P. Bliss (Director)

Mark Scerbo (Member)

Konstantin Cigularov (Member)

ABSTRACT

EFFECTS OF TRANSPARENCY AND HAZE ON TRUST AND PERFORMANCE DURING A FULL MOTION VIDEO ANALYSIS TASK

Sarah C. Leibner
Old Dominion University, 2020
Director: Dr. James P. Bliss

Automation is pervasive across all task domains, but its adoption poses unique challenges within the intelligence, surveillance, and reconnaissance (ISR) domain. When users are unable to establish optimal levels of trust in the automation, task accuracy, speed, and automation usage suffer (Chung & Wark, 2016). Degraded visual environments (DVEs) are a particular problem in ISR; however, their specific effects on trust and task performance are still open to investigation (Narayanaswami, Gandhe, & Mehra, 2010). Research suggests that transparency of automation is necessary for users to accurately calibrate trust levels (Lyons et al., 2017). Chen et al. (2014) proposed three levels of transparency, with varying amounts of information provided to the user at each level. Transparency may reduce the negative effects of DVEs on trust and performance, but the optimal level of transparency has not been established (Nicolau & McKnight, 2006). The current study investigated the effects of varying levels of transparency and image haze on task performance and user trust in automation. A new model predicting trust from attention was also proposed. A secondary aim was to investigate the usefulness of task shedding and accuracy as measures of trust. A group of 48 undergraduates attempted to identify explosive emplacement activity within a series of full motion video (FMV) clips, aided by an automated analyst. The experimental setup was intended to replicate Level 5 automation (Sheridan & Verplank, 1978). Reliability of the automated analyst was primed to participants as 78% historical accuracy. For each clip, participants could shed their decision to an automated analyst. Higher transparency of

automation predicted significantly higher accuracy, whereas hazy visual stimuli predicted significantly lower accuracy and 2.24 times greater likelihood of task shedding. Trust significantly predicted accuracy, but not task shedding. Participants were fastest in the medium transparency condition. The proposed model of attention was not supported; however, participants' scanning behavior differed significantly between hazy and zero haze conditions. The study was limited by task complexity due to efforts to replicate real-world conditions, leading to confusion on the part of some participants. Results suggested that transparency of automation is critical, and should include purpose, process, performance, reason, algorithm, and environment information. Additional research is needed to explain task shedding behavior and to investigate the relationship between degrade visual environments, transparency of automation, and trust in automation.

ACKNOWLEDGEMENTS

I would like to acknowledge my advisor and committee chair, Dr. James Bliss, for his patient guidance throughout this process. I would also like to acknowledge Dr. Anne Michalek for her kind assistance and extended loan of eye tracking equipment during data collection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Military operations rely increasingly on intelligence, surveillance, and reconnaissance (ISR). Operations in degraded visual environments (DVEs) are particularly dependent on ISR assets that offer the capability to "see" through fog, dust, or darkness. Advances in technological capabilities have led to an unprecedented increase in the amount of information available from ISR assets. This information increase far outpaces the number of available human analysts (McDermott et al., 2015; Yarovinskiy, 2017). Full motion video (FMV) is motion imagery transmitted at 31-60 Hz (Plott, McDermott, & Barnes, 2017), similar to television frame rates. FMV is a key real-time ISR asset, but also one of the most data- and time-intensive, making it difficult for field systems to transmit at high quality and for humans to comprehensively analyze trends. Solutions include reducing the amount of data transmitted from the sensor and automating target identification and tracking (Kreitmair & Coman, 2014; Poostchi, 2017). As automated FMV analysis increases in sophistication, human analysts' ability to work with these and other automated ISR systems becomes increasingly important.

Currently, FMV analysts must monitor continuously streaming, near-real-time video steams to identify anomalous objects or activities (Cordova et al., 2013). This task involves building a detailed picture of the pattern of life of an area, so that suspicious activity can be differentiated from benign activity. Other common tasks include identifying and tracking specific targets, monitoring targeted areas for activity, providing overwatch for patrols, and coordinating fire support for ground troops. These tasks vary from simple and repetitive to high-speed, dynamic, and complex. In degraded visual environments, FMV can provide critical information for operations, but the analytic burden on human analysts increases dramatically.

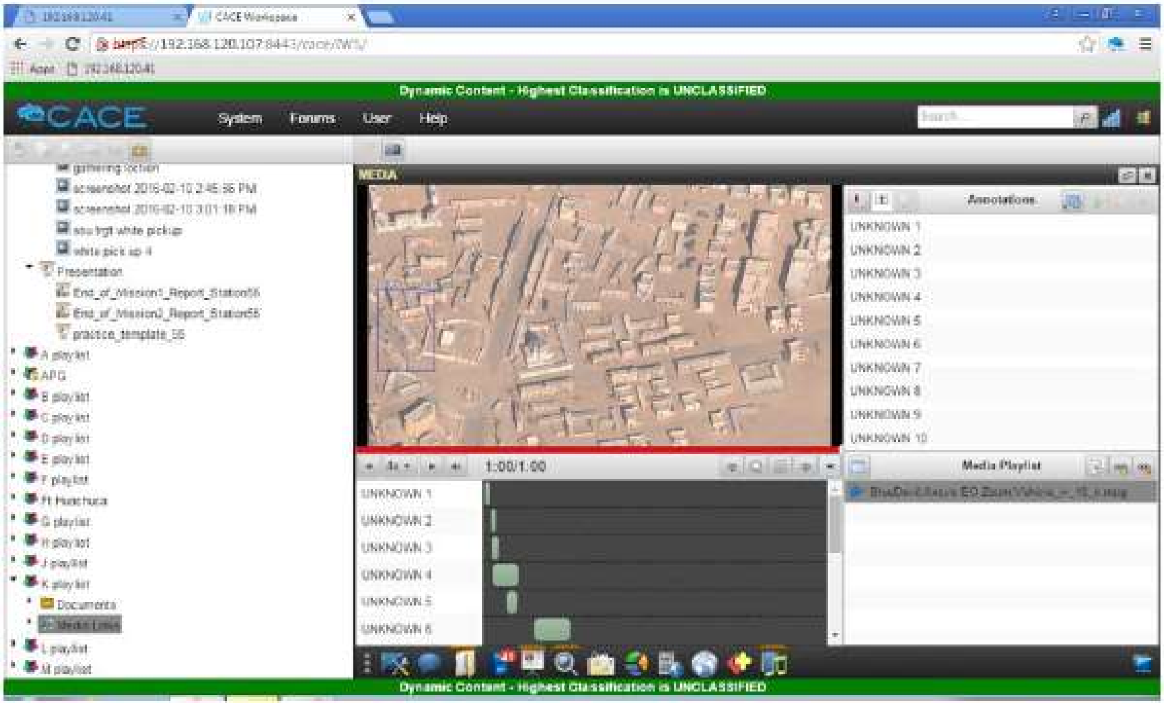
Transparency of automation is key to improving user performance (Chen et al., 2014; Itoh, 2010). Because automation should ideally reduce the analyst's task load, systems must be able to adjust the amount of information provided in real time to enable optimal performance. Transparency affects use of automation both directly and indirectly through trust (Nilsson, Laere, Susi, & Ziemke, 2012). In turn, appropriate trust in automation improves situation awareness and decision making (Endsley, 1996; Nilsson, Laere, Susi, & Ziemke, 2012).

Automated FMV Analysis

Automation of FMV analysis could reduce the perceptual and cognitive burden of FMV exploitation, particularly over widely dispersed areas or times or in DVEs. Other potential benefits include increasing the amount of FMV that can be exploited, improving the rate at which actionable intelligence is produced, and facilitating the use of FMV in multi-source intelligence (Thissell et al., 2015). Success of these goals can be assessed by considering four activities: instance recognition, category recognition, activity recognition, and target tracking (Cordova et al., 2013). Cordova et al. define instance and category recognition, respectively, as object recognition and identification (e.g., a T-54/55 Main Battle Tank), and determining whether an object belongs to a general category (a tank). Activity recognition seeks to identify specific human actions (e.g., Improvised Explosive Device emplacement activity), whereas tracking follows the movement of an object through a sequence of images (e.g., tracking a suspicious vehicle). These functions rely on computer vision algorithms, and current implementations must exchange accuracy for ease of implementation and computing efficiency, due to the limitations of available platforms.

Figure 1

Advanced Video Activity Analytics Interface



Note. Adapted from Schweitzer, K. M., Ries, A. J., McDermott, P. L., Plott, B. M., Wilson, E. A., & Morrow, G. P. (2018). *Human Factors Evaluation of Advanced Video Activity Analytics (AVAA) Functionality* (No. ARL-TR-8301). US Army Research Laboratory, San Antonio.

Advanced Video Activity Analytics (AVAA) is a system currently under development by the Army Research Laboratory. Figure 1 depicts the AVAA user interface. It is anticipated to become the Army's FMV exploitation system of record (McDermott et al., 2015). AVAA has the capability to detect, classify, track, and annotate persons, vehicles, and objects, and to filter video segments according to the Video National Imagery Interpretation Rating Scale (V-NIIRS) rating or by annotations on the video (Table 1 lists the VNIIRS rating scale with examples of each rating). The V-NIIRS scale is commonly used by military and government agencies to rate the quality of motion imagery. Each frame in a video has its own rating, so a video will have a range of V-NIIRS ratings.

Table 1

Video National Imagery Interpretation Scale

V-NIIRS rating	Identifiable targets
0	Interpretability of the imagery is precluded by obscuration, degradation, or very poor resolution.
1 [> 9.0 m (GRD*)]	Detect a medium-sized port facility and/or distinguish between taxi-ways and runways at a large airfield.
2 [4.5–9.0 m GRD]	Detect large static radars. Detect large buildings (e.g., hospitals, factories).
3 [2.5–4.5 m GRD]	Detect the presence / absence of support vehicles at a mobile missile base. Detect trains or strings of standard rolling stock on railroad tracks (not individual cars).
4 [1.2–2.5 m GRD]	Detect the presence of large individual radar antennas. Identify individual tracks, rail pairs, control towers.
5 [0.75–1.2 m GRD]	Identify radar as vehicle-mounted or trailer-mounted. Distinguish between SS-25 mobile missile TEL and Missile Support Vans in a known support base, when not covered by camouflage.
6 [0.40–0.75 m GRD]	Distinguish between models of small/medium helicopters. Identify the spare tire on a medium-sized truck.
7 [0.20–0.40 m GRD]	Identify ports, ladders, vents on electronics vans. Detect the mount for antitank guided missiles (e.g., SAGGER on BMP-1).
8 [0.10–0.20 m GRD]	Identify a handheld small-arms munition (e.g., SA-7/14, REDEYE, STINGER). Identify windshield wipers on a vehicle.
9 [< 0.10 m GRD]	Identify vehicle registration numbers on trucks. Identify screws and bolts on missile components.

**Note*: GRD = Ground Resolved Distance. Adapted from Plott, B. M., McDermott, P. L., & Barnes, M. (2017). *Advanced Video Activity Analytics (AVAA): Human Performance Model Report* (No. ARL-TR-8255). US Army Research Laboratory Aberdeen Proving Ground United States.

The AVAA system is currently in Phase II of development, so no accuracy data are yet available, although user testing has occurred. A study of the V-NIIRS filtering function found that analysts viewed 30% less video than when filtering was unavailable, and located 40% more targets of interest (McDermott et al., 2015). Discrete event simulation modeling based on these results suggested that AVAA could reduce historical video analysis time by 70%; that is, reviewing FMV feed later for non-time-critical tasks. However, real-time analyses, that is, viewing current FMV feed for time-critical tasks, were not significantly faster when using AVAA (Plott et al., 2017). A follow-on investigation comparing conditions with and without the automated annotation function found no significant difference for measures of task completion time or accuracy (Schweitzer et al., 2018). This suggests that task type and time availability affect the ability of automated systems to improve the performance of human analysts.

Although participants in Schweitzer et al.'s experiment were able to change automated annotation status from the default of "suggestion" to "reject" or "agree", the authors did not collect data that reflected users' level of agreement with the automation. Additionally, the automated annotations were restricted to yellow borders around an object of interest; the automation's reasoning process was not communicated. Poostchi, Palaniappan, and Seetharaman (2017) developed a computer vision algorithm that accurately detects and tracks objects. Color, intensity, gradient, and edge features are used in conjunction with background filters and path prediction. Poostchi et al. compared performance results from FMV testing with 61 contemporary object tracking algorithms. The new algorithm ranked 11 out of 62, losing the target 1.3 times on average. The best object recognition and tracking algorithms tested using the 2016 Visual Object Tracking Challenge dataset lost targets about .8 times per video sequence on average. As an example of the complexity of the tracking task, parallax of tall buildings can induce apparent motion (wobble) in FMV, resulting in motion detection false positives.

Poostchi's context-aware algorithm was able to reduce false positives to 16% by filtering out buildings.

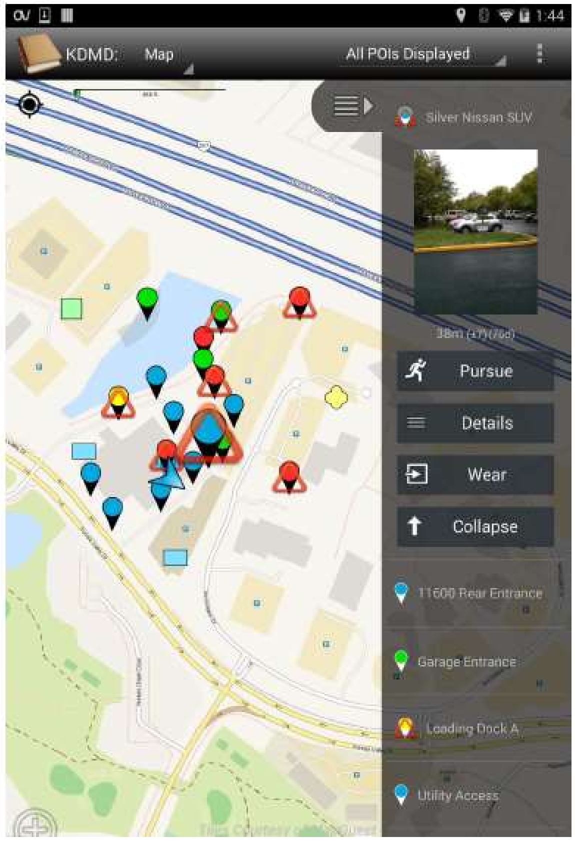
Yarovinskiy (2017) discussed contemporary algorithms designed to automatically detect suspicious activity within human crowd footage. Traditionally, suspicious activity has been identifiable only after an analyst acquires knowledge of the region's pattern of life from hours of monitoring an area (Kuwertz, Sander, Pfirrmann, & Dyck, 2017). Automated detection can reduce analysts' cognitive workload levels by identifying video segments containing anomalous behavior. Automated detection algorithms are either supervised or unsupervised by human operators. Supervised algorithms detect anomalies based on preloaded rules, whereas unsupervised algorithms detect anomalies based on statistical analysis of detected activities. Eight contemporary algorithms achieved 75-90% accuracy while analyzing camera footage of a pedestrian walkway at the University of California, San Diego (Yarovinskiy, 2017).

Finally, another approach to automated FMV analysis reduces transmission of video from an unmanned aerial system (UAS) by selecting regions of interest based on content. This approach has the additional advantage of reducing transmission-related quality loss. The main disadvantage is that any footage that may contain unidentified regions of interest will be unavailable. Kölsch and Zaborowski (2014) developed an algorithm that achieved 77.2% accuracy identifying vehicles at Camp Roberts, California, with 0.2 false positives on average per image. However, only still imagery was used.

FMV video feed is available to analysts on several technological platforms, including stationary computer terminals, displays in mobile tactical ground stations, and even smartphones (Gilmore, 2016; Madden et al., 2014; see Figure 2 for an example of a mobile FMV application interface). Mobile FMV analysis exacerbates the challenges of limited bandwidth and

Figure 1

Mobile FMV System



Note. Adapted from *Madden, D., Choe, T., Deng, H., Gunda, K., Gupta, H., Ramanathan, N., … Hakeem, A. (2014)*. Mobile ISR: Intelligent ISR management and exploitation for the expeditionary warfighter. *In 2014 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1–11. Retrieved from https://doi.org/10.1109/AIPR*

resolution, necessitating advanced approaches discussed above such as automated region of interest selection and prioritization (Kreitmair & Coman, 2014). Other challenges include narrow fields of view (FOVs), degraded visual environments (DVEs), and problematic environments with excessive clutter, terrain, shadows, and motion (Olson, Gaudiosi, Beard, & Gueler, 208; Parker, 2015; Poostchi, 2017). Operators must be aware of these limitations to make informed decisions about automation usage.

Levels of Automation

Automation has been defined as full or partial replacement of a function previously carried out by the human operator (Parasuraman, Sheridan, & Wickens, 2000). Although researchers have developed numerous automation models, Sheridan and Verplank's Ten Levels of Automation (1978) remains the most commonly cited model. The scale ranges from low (1: no computer assistance) automation to high (10: human-out-of-the-loop system) automation (Sheridan & Verplank, 1978; Table 2). Automation handles mundane or repetitive tasks, freeing humans to engage in more critical work. It also sorts and integrates large quantities of information and manages information visualization (Cordova et al., 2013; Parasuraman & Riley, 1997). However, the capabilities of computer vision still trail human abilities in object and

Table 2

Levels of Automation

LOW	
	1. The computer offers no assistance: Human must take all decisions and actions
	2. The computer offers a complete set of decision/action alternatives, or
	3. Narrows the selection down to a few, or
	4. Suggests one alternative;
	5. Executes that suggestion if the human approves, or
	6. Allows the human a restricted time to veto before automatic execution, or
	7. Executes automatically, then necessarily informs the human, and
	8. Informs the human only if asked, or
	9. Informs the human only if it, the computer decides to
	10. The computer decides everything and acts autonomously, ignoring the human
HIGH	

Note. Adapted from Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Cambridge, MA: Massachusetts Inst Of Tech Man-Machine Systems Lab.

activity recognition (Khosla, Uhlenbrock, & Chen, 2017). Therefore, automation reliability is imperfect, which in turn may impact user trust and performance.

Reliability

Decision-making research approaches such as decision ecology and Bayesian inference have been applied to predict users' behavior when interacting with unreliable automation. Decision ecology considers factors in the decision environment, such as the source, value, reliability, context, and cost of accessing information to aid choice making. Bayesian inference principles suggest that a rational agent assesses the current probability of an event based on the prior probability, updated with the outcome of each new trial. These models agree with previous studies that automation will result in reduced performance when reliability falls below 70%, at least when the reliability level is consistent (Acharya et al., 2018; Wang, Zhang, & Yang, 2018; Wickens & Dixon, 2007). Automated video analysis systems currently achieve between 70% and 91% object recognition accuracy, with even higher object detection rates (Gaszczak, Breckon, & Han, 2011; Kölsch & Zaborowski, 2014; Muncaster et al., 2015; Poostchi, 2017; Yarovinskiy, 2017), and are well above the 70% threshold on average (M = .82, SD = .06). Although Mishler et al. (2017) suggest that feedback may result in better reliability estimates than priming (providing users an estimate of system reliability prior to use), priming reliability reduces session time as well as the possibility that environmental factors such as DVEs may affect reliability estimates (Schaefer et al., 2016).

Human-Automation Trust

Trust of automation impacts SA, decision making, complacency, and detection of deception (Biros, Daly, & Gunsch, 2004; Chung & Wark, 2016; Parasuraman & Riley, 1997).

Overtrust is the most common cause of misuse and disuse of automation (Lee, 2008). This leads to complacency, lack of attention, and overreliance on heuristic decision making. Overtrust also compounds the effects of errors in automation (Rovira, McGarry, & Parasuraman, 2007; Wickens et al., 2007). Though lower trust can improve performance by increasing operator attention, trust that is excessively low can result in automation disuse and decreased task performance. Therefore, levels of trust calibrated to an automated system's reliability are necessary for optimal performance (Chen & Terrence, 2009). Low trust is often referred to as distrust; however, some argue that distrust is negative trust (regarding a system as harmful or nefarious), distinct from negation of trust (uncertainty about the trustworthiness of the system; Dimoka, 2010; Itoh, 2010; Marsh & Dibben, 2005). Mistrust is defined as misplaced trust. Mistrust can occur when a user either trusts unreliable automation, or distrusts reliable automation (Muir, 1994). Both distrusting and mistrusting automation lead to disuse, which often manifests as missed signals or slower reactions (Parasuraman & Riley, 1997). Current self-report instruments, however, do not distinguish between the subtypes of trust.

The current study taps into negation of trust, because the primary effect of degraded visual environments (DVEs) and inadequate transparency is increased uncertainty about the accuracy of the automation. Accuracy, speed, and workload are all critical components of performance (Wickens, Hollands, Banbury, & Parasuraman, 2015). Misses are especially problematic in FMV analysis, because a missed target could result in casualties. Additionally, speed can be critical when missions occur in real time (Cordova et al., 2013). Other factors that affect trust include automation reliability, as previously discussed, and information quality (i.e., DVEs) and transparency of automation, which are the variables of interest in this study (Bailey & Scerbo, 2007; Chen et al., 2018; Parasuraman, Mouloua, & Molloy, 1997).

Situation Awareness and Decision Making

The purpose of ISR is to enhance situation awareness (SA) and decision making (Cook, Angus, & Adams, 2004). Any discussion of automation or design, therefore, must consider potential impacts on these outcomes. SA includes perceiving and understanding elements in the environment, conceptualizing their interaction, and predicting the status of such elements in the future (Endsley, 1995). These are also critical elements of transparency of automation, as discussed later. SA affects performance and decision making by enhancing attention, perception, and memory (Sohn & Doane, 2004). Decision making results from intuitive processes that match current challenges with previously successful strategies, as well as deliberate processes that analyze challenges according to set rules (Klein, 2008). Automation has the potential to improve these processes, but when poorly implemented, decision quality is reduced.

Poorly implemented automation reduces SA by inducing operator complacency, fostering passive interactions with the system, and limiting task process feedback to the operator (Endsley, 1996; Nilsson, Laere, Susi, & Ziemke, 2012; Ruff et al., 2004). Complacency occurs when humans over-trust automation. They reduce their attention to automated tasks, and therefore are unaware of errors made by the automated system. Complacent operators show reduced comprehension of information and poor manual operating skills (Parasuraman et al., 2000). When automated tasks provide minimal feedback, operators lose the ability to determine information source, reliability, and analytic rules. Both reliable and unreliable automation can negatively impact SA when attention is reallocated to other tasks (Onnasch, Wickens, Li, & Manzey, 2014).

Overreliance on automation negatively affects decision making when the automation recommends decisions (manifested within Levels 5-10 of Sheridan & Verplanck's model; Rovira et al., 2007). Heuristic decision making may be troublesome if false alarms and misses are treated equally. In ISR, false alarms are preferable to misses because the implications for misses are greater (Cordova et al., 2013). When imperfect automation necessitates additional human decision-making, performance may suffer. Decision making may be even worse when automation that is considered to be reliable behaves in an unreliable manner (Endsley & Kaber, 1999). To mitigate these effects, most researchers recommend that operators be kept in the loop (Endsley, 1996; Onnasch et al., 2014). Current FMV analysis systems represent automation Level 5, executing an automated suggestion only when the operator approves (Sheridan & Verplank, 1978). Therefore, operators should be kept in the loop and potential decrements due to over-reliance on automation must be mitigated.

Degraded Visual Environments

Degraded visual environments (DVEs) are of particular concern to military operators; enemy forces often gain an advantage during low-visibility conditions (Knights & Mello, 2017). Within FMV, DVEs can arise from variable atmospheric conditions, illumination, transmission errors, resolution, noise, background motion, clutter, or overloaded or weak networks (Harguess, Shafer, & Marez, 2017; Hollock, 2017; Kölsch & Zaborowski, 2014; Kreitmair & Coman, 2014; Parker, 2015). Atmospheric conditions can be corrected by post-processing algorithms (Zhang, Li, Qi, Leow, & Ng, 2006); however, the most reliable corrective approaches are too slow for real-time FMV applications, whereas faster approaches are unable to achieve high image quality (Kumari & Sahoo, 2016). Currently deployed aerial ISR platforms must still fly lower and slower in foggy conditions to collect sufficient quality video (Menthe, Hura, & Rhodes, 2014). Illumination, clutter, and camouflage issues are often addressed by overlaying electro-optical and thermal feed visualizations (Gaszczak et al., 2011; Parker, 2015), but not all FMV sensors have this capability (Myhr et al., 2018). Reducing bandwidth demands by identifying regions of interest to transmit, rather than transmitting the full video, can reduce transmission errors and degradation; however, this approach risks omitting important areas of interest (Kölsch & Zaborowski, 2014). Resolution is often exchanged for bandwidth. The average military FMV resolution is 480x720 pixels, which is adequate for object recognition at around 90%, but results in very small fields of view (Ross & Coman, 2014). Therefore, context is often lacking. The implication for automated video analysis users is that FMV will either be degraded or lack context, making it difficult to analyze, and that automated analysis systems will be unreliable.

Atmospheric haze is a type of DVE that negatively affects the sharpness, contrast, and brightness of images (Guevara et al., 2017; Kahraman & De Charette, 2017). Haze consisting of 1-10 μm aerosol particles scatters ambient light and reduces visibility. Haze particles take the form of smoke, fog, humidity and air pollutants (Kahraman & De Charette, 2017). Haze varies greatly across geographic environments; for example, the visual range on an average day in the eastern U.S. is 17 km, while the visual range on an average day in the western U.S. is 155 km (Molenar, Malm, & Johnson, 1994). Even in the absence of visible fog, haze can noticeably blur imagery taken at optical path distances of over 1.5 km (Kopeika et al., 1998). Because UAVs collect FMV from altitudes of 500 m (smaller crafts, night-time) to 5.8 km (larger crafts, daytime), haze represents a common source of degradation (Menthe et al., 2014). Haze is measured by calculating a scattering coefficient; a coefficient of 10 mm^{-1} corresponds to approximately 50 km visibility on average, whereas 50 mm^{-1} corresponds to approximately 6 km visibility (National Research Council, 1993). For example, a one-year collection of scatter values around Kanpur, India found values ranging from 58 to 584 mm^{-1} (Ram et al., 2016).

Current solutions are unable to address all types of DVE. Degradation of FMV, therefore, presents a challenge to automated video analysis implementation. For example, image distortion has been found to reduce trust in automated target recognition (MacMillan, Entin, & Serfaty, 1994). Analyst performance is lowered in the presence of reduced resolution and brightness, even when automated target recognition aids are utilized (Narayanaswami, Gandhe, & Mehra, 2010). In these circumstances, analysts may not rely fully on the automated aid. Perceived information quality predicts trust as well as perceived risk (Nicolaou & McKnight, 2006), which in turn may influence decision making. When analysts can filter FMV by quality, they view 55% less video feed (Plott et al., 2017), which may have implications for missing targets of interest. Although Hancock et al.'s (2011) meta-analysis of trust in automation research suggested that environmental factors are only moderately associated with trust, physical environment was not included in their analysis, and the interaction between trust and visual degradation was not considered. Visual degradation and transparency may interact to affect trust (Nicolaou & McKnight, 2006; Yeh & Wickens, 2001). The purpose of the current study was to examine the effects of haze degradation on analyst trust and performance using sharpness and contrast variations.

Transparency of Automation

Information transparency is critical for establishing appropriate levels of trust in automation (Chen et al., 2014; Itoh, 2010; Lyons, 2013). Generally, greater transparency leads to quicker calibration of trust in automation (Chen et al., 2014; Lyons, 2013; Yang, Unhelkar, Li, & Shah, 2017), although some studies found increases in latency and workload (Chen et al., 2014). Two convergent models of transparency have been proposed. Lyons (2013) defined transparency as a system's capacity for relaying information that the human operator needs to know about the automated (robotic) system, such as system capabilities, processes, and limitations, and any information that the automated system needs to know about a human, such as the human's goals or task criteria. Transparency of the human is important to human-robot teaming; for example, Lyons suggested that a robot could recommend increasing its level of autonomy if it were aware that the human was overloaded. However, level of automation does not typically vary within an automated video analysis system, nor are such systems currently constructed according to a human-robot teaming model. Therefore, this component of transparency is not clearly related to the current problem of interest. Lyons recommended providing the user with an intentional model (intent or purpose of the automation), a task model (progress, capabilities, and errors), an analytical model (analytical principles followed by the automation), and an environmental model (current conditions and any automation limitations due to environment).

Chen et al. (2014) proposed a model similar to Lyons' but added the component of projection to future state. Chen et al.'s Situation Awareness-based Agent Transparency model describes the components of transparency within Endsley's Situation Awareness framework. Purpose, process, and performance support Level 1 transparency; reasoning process and environment support Level 2 transparency; and projection to future states supports Level 3 transparency. There are few differences in transparency recommendations from these two models, but Lyons' model focuses on aspects of human-robot teaming that are less relevant to automated video analysis, so the current paper will rely on Chen et al.'s (2014) model to inform variations in level of transparency.

As Lyons' (2013) and Chen's (2014) models suggest, the content that defines transparency varies across tasks and environments. Manipulations of transparency in research have included automation reliability (Yang et al., 2017), location, resources, goals, predicted outcomes, reasoning (e.g., primary reason for a decision, such as rerouting a convoy due to traffic), time of last information (Chen et al., 2018), weather, equipment status, and menu options for additional information (Lyons et al., 2017). Transparency literature stresses that reliability must be communicated to the user (Chen et al., 2014; Itoh, 2010; Lyons et al., 2017). However, perceptions of reliability may be reduced by degraded images (Yeh & Wickens, 2001), or trust may increase if degraded imagery is perceived as difficult to analyze (Schaefer, Chen, Szalma, & Hancock, 2016). Transparency has also been found to mitigate the effects of perceived information quality on perceived risk and trust (Nicolau & McKnight, 2006). These findings have differing implications for the use of transparency in automation. Too much transparency information confuses users, whereas too little reduces trust; but in the presence of degraded visual conditions, the optimal transparency level may be different. Therefore, research suggests that it is important to explore these two constructs (transparency and haze) together.

Convergent Measures

Synthesizing subjective, performance-based, and physiological measures is desirable due to the increased levels of detail and of validity that can be achieved (Neupane, Saxena, & Hirshfield, 2017; Wierwille & Eggemeier, 1993). Decision speed, frequency of task shedding, and eye-tracking fixations have effectively reflected trust, but the results of these studies demand replication. The proposed study will add to previous literature, cross validating these different measurements with each other and with Madsen and Gregor's (2000) self-report measure of trust in automation.

Performance in the context of an ISR task means quickly and accurately identifying suspicious activity, as well as freeing operator resources to focus on other tasks such as communication and reporting. However, decision time is sometimes neglected in automation research. As well as a desired outcome of automation, decision speed is affected by automation reliability, task criticality, and task complexity (Rovira et al., 2007). For example, decision speed was not significantly different when using the AVAA system (Plott et al., 2017), suggesting that any decision support provided may have been nullified by reliability concerns or increased complexity. Although shorter decision times might be expected to reflect heuristic decision-making, Getty, Swets, Pickett, and Gonthier (1995) found that participants' decision-making strategies reflected sensitivity to the system's overall accuracy and false alarm rates. Unreliable automation resulted in slower decision times when there was no risk associated with delay, but a benefit for accuracy, suggesting increased analysis due to lower trust. These results suggest that, after controlling for task complexity and operator engagement, decision speed is a promising behavioral measure of trust in automation.

Trust of automation often translates to usage (Parasuraman & Riley, 1997; Schaefer et al., 2016). Usage takes the form of reliance (allocating attention to other tasks when automation is not signaling), and compliance (switching attention to an automated alarm) (Dixon et al., 2007). Adaptive task allocation to the machine (Parasuraman & Hancock, 2001) is another type of usage that occurs when operator reliance on automation varies with workload. However, research suggests that user trust interacts with workload to predict task shedding, so this is also a behavioral measure of trust (Bliss, Harden, & Dischinger, 2013).

Optimal use of automation involves appropriate allocation of attention (Parasuraman & Manzey, 2010). Attention can also be an acceptable proxy for trust. For example, attention to automated signals and attention to other tasks when the automated system is not signaling indicate optimal trust in automation, whereas reduced attention to the automation may indicate over-trust (Dixon et al., 2007). Conversely, paying more attention than necessary to the automation may indicate distrust. Eye movement is an acceptable measure of attention (Parasuraman & Manzey, 2010), and eye tracking behaviors (e.g., scanning frequency, gaze dispersion) have been used successfully as a convergent measure of trust in automation (Karpinsky, Chancey, Palmer, & Yamani, 2018; Louw & Merat, 2017). For example, Karpinsky et al. measured participants' allocation of visual attention in terms of percent dwell time and found that higher workload led to lower trust which in turn led to lower percent dwell time on the automated system display. Similarly, Metzger and Parasuraman (2005) found that complacent operators made significantly fewer fixations on an automated radar display. In a DVE driving task, Louw and Merat (2017) found that degraded conditions resulted in greater attention to the road compared to high visibility and manual conditions, suggesting reduced trust in automation; however, attention switching to alerts was not affected. Overall, performance was best in the degraded visibility condition, likely due to increased attention. Taken as a whole, past research findings suggest a complex relationship among task, task environment, trust, and visual attention.

Current Study

The current research involved an investigation of the effects of haze and automation transparency on users' trust in automation and performance on an FMV analysis task. Researchers have studied the effects of transparency on trust (Chen & Terrence, 2009; Itoh, 2010; Lyons, 2013; Wright et al., 2017), but perceptual conditions in such experiments have usually been ideal (MacMillan et al., 1994; Narayanaswami et al., 2010). It is important to consider these together due to the prevalence of degraded conditions and their impact on transparency-mediated trust (Yeh & Wickens, 2001). The current research investigated performance (accuracy and decision time), as well as behavioral (task shedding and gaze dispersion) and qualitative (well-established self-report surveys) measures of trust, as a function of information transparency and scene degradation.

The experimental task consisted of visually searching for improvised explosive device emplacement activity within short FMV clips, half of which had reduced contrast and brightness (haze). The task environment simulated an automated analyst's recommendation in the top left corner of each FMV clip. Participants were asked to decide whether to agree or disagree with the automated feature, request more information, or delegate decision making to the automated system. Reliability was fixed, but the amount of information available about the automated analyst varied. This study used a combination of traditional and novel approaches to collect convergent measurement data reflecting participants' trust in automation. The expectation was that findings could be integrated with current theory regarding the effects of degraded visual environments and transparency as well as the appropriateness of using decision speed and task shedding as indices of trust in automation.

Hypotheses

H1: During a convoy route-planning task, Wright et al. (2017) found that providing an automated agent's reason for recommending a route change resulted in fewer incorrect agreements. Wright et al. manipulated three levels of transparency while asking participants to accept or reject automated route-change recommendations. Participants displayed complacent behavior in the non-transparent condition but appeared to be overwhelmed by the amount of information in the highest transparency condition, suggesting the importance of determining the optimal transparency level. Therefore, it was expected that higher transparency would predict higher proportions of participants accuracy measured in terms of proportion of correct agreements/disagreements with the automated system's recommendations. However, accuracy in the highest transparency condition was expected to be reduced.

H2: Low trust has been associated with longer decision times (Chen & Terrence, 2009), whereas image distortion has been found to reduce trust in automated systems (Macmillan et al., 1994). Therefore, it was expected that (**H2a**) haze level would positively correlate with decision times for agreement with the automated system's recommendation and that (**H2b**) haze would negatively correlate with trust measured by Human Computer Trust Scale (HCTS) score (Madsen & Gregor, 2000).

H3: Task shedding has been found to increase under conditions of greater task complexity and uncertainty (Bliss et al., 2013; Parasuraman & Hancock, 2001), such as can result from poor imagery quality (Hooey et al., 2018). Therefore, it was expected that transparency and degradation would interact to predict task shedding, measured as number of choices to delegate the task to the automation. For the main effects, low transparency and high degradation were expected to predict higher levels of task shedding.

H4**:** Degraded environmental conditions have been shown to result in lower gaze dispersion in a driving task, with attention focused on the primary region of interest (ROI; Louw & Merat, 2017), whereas bandwidth (amount of information found in a given ROI) has been shown to be a strong predictor of scanning behavior (Horrey, Wickens, & Consalus, 2006; Wickens et al., 2003). Therefore, it was anticipated that greater degradation would be positively related to fixations on the high-bandwidth ROI; that is, the region of interest within the display that provides the highest amount of task-relevant information.

H4a**:** Reduced attention to the automation may indicate over-trust, whereas greater than optimal attention may indicate distrust (Dixon et al., 2007). Therefore, attention to the primary ROI was expected to negatively relate to self-reported trust level (dependent variable), and vice versa.

H4b: Distribution of attention to the identified ROIs was expected to differ significantly from the predicted optimal proportion only when trust was very high or very low.

CHAPTER 2

METHOD

Design

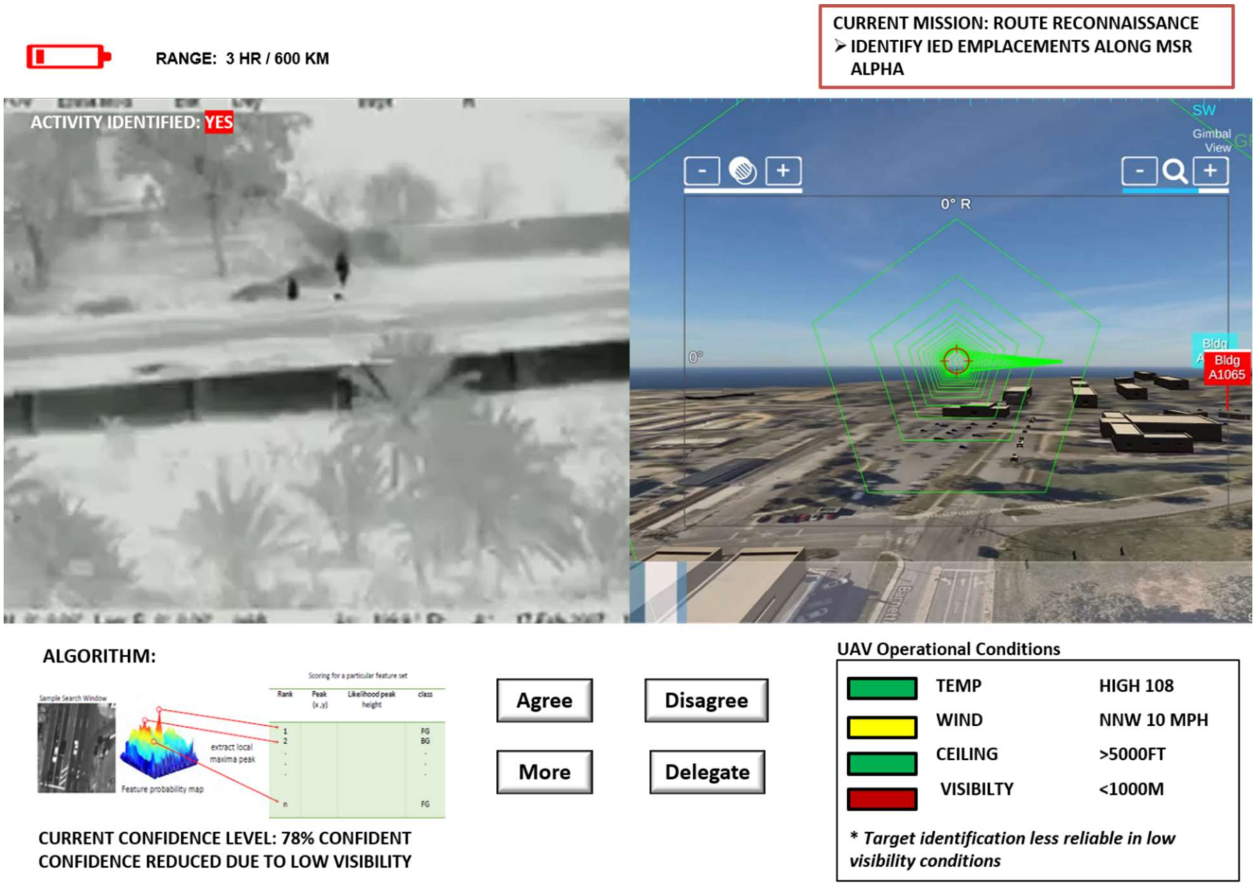
The current study employed a 2 (haze: 0, +30% light scattered) x 3 (transparency: low, medium, high) within-group design. The experimental task consisted of an information screening task that required participants to analyze 18 10-second FMV clips with the aid of an automated analyst (Figure 3). The automated analyst textually noted the presence or absence of suspected improvised explosive device (IED) activity within the FMV clips, and the participants' task was to choose whether to "agree" that there was IED activity present; "disagree" with the automated analyst; ask for additional information by choosing the "More" button; or "delegate" the decision to the automated system.

Independent variables

The first independent variable was imagery degradation (within groups), which was set at either zero or high haze (+30% light scattering). Atmospheric haze is measured by the scattering coefficient of incident light. Due to the difficulty involved in computing the scattering coefficient of an image, the current study simulated haze using sharpness and contrast settings derived from previous research (Liu et al., 2017). To simulate a 30% increase in light scattering, image saturation was reduced by 100% and brightness was increased by 35%. Half of the FMV clips were zero haze (control) and half were high haze. Zero and high haze clips were randomly presented throughout the experiment; the order was determined using Excel's random number generator function.

Figure 3

Task Environment Interface, Transparency Level 2, Zero Degradation



The second independent variable was transparency (within groups), which was manipulated incrementally (additional levels added on to the first). Low transparency included providing information about the system's purpose (identify IED emplacement activity), process (the current range of the UAV whose feed the system is analyzing), and performance (automation reliability) information, equating to Level 1 of Chen et al.'s (2014) Situation Awareness-based Agent Transparency model (see Table 3). Medium transparency included

Table 3

Transparency Levels

Level	Information Displayed
1	Purpose: Identify IED emplacements along a route Performance: Reliability Reason: Limits of sensor-reduced confidence in haze Process: Current range of UAV
2	Purpose: Identify IED emplacements along a route Performance: Reliability Reason: Limits of sensor-reduced confidence in haze Process: Current range of UAV Reasoning: Computer vision algorithm used to identify activity Environment: Current weather with limits of UAV/Sensor
3	Purpose: Identify IED emplacements along a route Performance: Reliability Reason: Limits of sensor-reduced confidence in haze Process: Current range of UAV Reasoning: Computer vision algorithm used to identify activity Environment: Current weather with limits of UAV/Sensor Future states: Weather forecast with limits of UAV due to wind, cloud ceiling

information about the system's reasoning process (object recognition algorithm) and environment (current weather and its limiting effects on the UAV), in addition to Level 1 information. High transparency included projections to future states (weather forecast and weather effects on the UAV), in addition to Levels 1 and 2 information.

Dependent variables

Dependent measures included the accuracy of participants' agree/disagree choices, their overall decision speed in seconds, frequency of task shedding, and attention (pattern of gaze fixations on the stimulus). Each dependent variable is fully described below.

Accuracy. Accuracy was reflected by a combination of two measures derived from signal detection theory, hits and correct rejections. This measure reflects the *a-b* Signal Detection Theory model approach to measuring accuracy (Bustamante, 2014). Accuracy was defined as the participants' combined proportion of hits (correct agreements with the automated analyst) and correct rejections (correct disagreements). Because the automated analyst was 78% accurate overall, with a 6% miss, and a 16% false alarm rate, a highly accurate user would be expected to accept 78% of automated annotations and reject 22% in a manner that matched the system's accuracy rate.

Decision Speed. Decision time was measured by the number of seconds between stimulus onset and participant choice. The automated analyst's recommendation appeared concurrently with the stimulus onset. This time was recorded during eye tracking by Tobii Studio version 3.2.

Task shedding. Task shedding was measured by the proportion of "delegate" decisions the participants made within each condition. Participants were informed that they could choose to delegate any analysis decision, and that the automation was highly reliable but not perfect.

Participants did not receive feedback about the outcomes of their delegate choices, to avoid the "first failure effect." That effect occurs when participants are first alerted that the automation is in error, as trust is reduced and then slowly recovers (Parasuraman & Manzey, 2010). The first failure effect could have obscured the effects of varying haze and transparency, which were the focus of this study. Additionally, some research suggests that primed reliability may result in more appropriate trust levels compared to feedback (Koo et al., 2014; Schaefer et al., 2016).

Attention. Gaze dispersion was measured by the percentage of fixations on the high-bandwidth region of interest (ROI 1, Figure 4), using a Tobii X2-60 portable eye tracking device. Although raw eye-tracking metrics can be analyzed to reveal interesting patterns, fixation data should be compared to optimal scanning levels to assess trust. Optimal scanning behavior suggests well-calibrated trust. However, optimal scanning frequency is extremely difficult to calculate in many applied tasks (Parasuraman & Manzey, 2010). Instead, most researchers use Wickens et al.'s (2003) adaptation of the Saliency, Expectancy, Effort, and Value (SEEV) model (see Figure 5a). This model attempts to describe visual attention as a function of the physical ability of a display area to capture attention (salience), user's expectation of gaining information from a visual area (expectancy), physical or time cost of accessing information (effort), and relevance of a visual region to the task (value). It could be argued that Johnson, Duda, Sheridan, and Oman's (2017) model (Figure 5b) is closer to a prescriptive model of optimal scanning, because uncertainty is a component of sampling theorems, but it does not predict users'

Figure 4

Regions of Interest (ROIs) for gaze data. ROI 1 is the high-bandwidth ROI

Figure 5

Models of Attention

a) $$Visual\ Attention\ to\ AOI = \sum_{task=1}^{n} \left[BW\ x\ R_{xy}\ x\ V\right]$$

b) $$Visual\ Attention\ to\ AOI = \sum_{task=1}^{n} [V + 2U - E\ ]$$

c) $$Visual\ Attention\ to\ AOI\ = \textstyle\sum_{task=1}^{n}\left[(BW\ x\ R_{xy}\ x\ V) + 2U\right]$$

Note. *(a)* Model of optimal attention adapted from Wickens et al. (2003). AOI = area of interest, BW = bandwidth, R_{xy} = relevance to task, and V = value of task (priority). *(b)* Model of optimal attention adapted from Johnson et al. (2017). V = value of task (priority), U = uncertainty, and E = effort. *(c)* Proposed model of optimal attention, adapted from Wickens et al. (2003), Johnson et al. (2017), and Horrey et al. (2006).

behaviors well, especially under high workload conditions (Parasuraman & Manzey, 2010; Senders, 1983). However, Horrey et al. (2006) were able to account for 97% of variance in scanning behavior using only bandwidth and value, so the importance of bandwidth to optimal scanning is clear. A high bandwidth area of interest is an area in which relevant information appears more frequently compared to other areas. Similarly, Wickens et al. (2003) were able to explain over 90% of variance in gaze fixations by loading only bandwidth, relevance, and value into their model. Thus, an optimal scanning model should include bandwidth, relevance, value (highly explanatory variables) and uncertainty (prescriptive variables), at a minimum. For each stimulus, the optimal scanning frequency was calculated according to the proposed visual attention model in Figure 5(c).

Participants

The number of participants needed was determined through a review of relevant literature and an *a priori* power analysis conducted using PASS 16.0.1. Because several different factors were being measured, the highest effect size, which was medium, was used. For eye tracking model fit, a X^2 power analysis indicated that 46 participants should be required to achieve a power of .80 and an effect size of $\omega = 0.50$. The current research proposes six hypotheses, which Tseng and Shao (2012) found would not greatly affect the sample size needed to maintain good familywise power. A significant hypothesis testing threshold of $p = .05$ was used for all tests; to balance the risks of making a Type I and Type II error.

Participants were 48 undergraduate students (35 female) from Old Dominion University, recruited using the Sona Research Participation System. The average age of the participants was 20.8 years ($SD = 4.97$), and 8% of the sample reported current or prior military service. Participants reported playing video games 2.5 hours per week on average ($SD = 3.7$) and using a computer 18.6 hours per week on average (*SD* = 13.9). Participants received 1 research credit for participation. The study took approximately 30 minutes per participant to complete. The study was approved by Old Dominion University's Institutional Review Board and signed informed consent forms were obtained from each participant prior to participation (Appendix A).

Materials

Information screening task

The FMV task interface (Figure 3) was created using PowerPoint, YouTube Movie Maker video editor, and Tobii Pro Studio (version 3.2). Real military FMV videos were downloaded from military.com and clipped into eighteen 10-second segments. Half of the clips were degraded by reducing saturation by 100% and increasing brightness by 35%. This created a haze effect over the imagery according to the noise modeling equation $M_n = \rho(1 - B) + (1 - \rho)S$, where M_n is inversely proportional to the amount of haze effect (Liu et al., 2017). In this equation, B represents brightness, S represents saturation, and ρ was set to .85, based on Liu et al.'s analysis of an image with approximately the desired amount of haze. Each clip indicated whether the automated analyst had detected activity via a red highlighted YES or green highlighted NO in the upper left corner. Fourteen of the videos showed IED emplacement activity and indicated that the automation detected the activity (hits), three videos did not show emplacement activity but indicated that the automation detected the activity (false alarm), and one video showed emplacement activity but indicated that the automation did not detect the activity (miss). Participants were shown the 10-second FMV clips and instructed to look for IED emplacement activity with the assistance of the automated analyst. If there was no apparent IED emplacement but the automated analyst indicated there was, the participant should "disagree" with the automation. If there was IED emplacement activity, the participant should "agree" with the automation. The participants' choices, including "delegate" and "more information" choices, were visually captured on experiment screen recordings by the Tobii Studio software. The time in seconds required to make a choice was also recorded by Tobii Studio. Participants were told that selecting "more information" would alert the system to search for more information, but it would not be available during the experiment. This option was provided so that participants did not feel forced to make a choice. If no choice was made within 10 seconds, the next stimulus was presented and no decision time data was recorded.

Secondary task

A secondary monitoring task was provided side-by-side with the information screening task. The monitoring task consisted of eighteen 10-second simulated drone footage clips of a military convoy traveling through a Middle Eastern city. The videos were created using Virtual Reality Rehab's Fused Augmented Reality User Interface version 1.4.6 system, with permission. Participants were told that they needed to monitor the convoy as well as search for IED emplacement activity in the FMV feed.

Eye tracking

A Tobii X2-60 portable eye tracking device was used to collect gaze data. This device was mounted beneath the monitor and had a sampling rate of 60 Hz with an accuracy of .2 degrees of visual angle. No chin rest was used to emulate real-world viewing conditions. The Tobii allows 44cm x 32cm of head movement. Fixations were extracted from the raw gaze data using the recommended Tobii Pro I-VT filter. This filter classifies fixations as gaze samples with a velocity below 30° per second. An unweighted moving average filter is used to reduce noise. This filter has been found to be one of the most accurate for categorizing fixations and eliminating noise (Hild, Voit, Kühnle, & Beyerer, 2018; McChesney & Bond, 2017).

Measures

Demographics Questionnaire

Each participant completed a questionnaire indicating his or her age, biological sex, military service, visual deficiencies, average hours per week spent gaming, and average hours per week of computer use (Appendix B).

Online Trust Questionnaire

Participants completed a six-question trust survey (Madsen & Gregor, 2000; Appendix C) after each stimulus presentation for each unique combination of IVs; a total of six surveys. The trust questionnaire included five items relating to trust and one item relating to confidence in their decision. This questionnaire was derived from Madsen and Gregor's Human-Computer Trust Scale, a 25-item questionnaire based on five trust-related constructs: perceived reliability, perceived technical competence, perceived understandability, faith, and personal attachment. The current study chose the highest-loading item (Madsen & Gregor, 2000) with good discriminant validity from each of these five constructs. From perceived reliability, the item "I can rely on the system to function properly" was chosen. From perceived technical competence, the item "The system has sound knowledge about [the key identification features of IED emplacers]" was chosen. From perceived understandability, the item "Although I may not know exactly how the system works, I know how to use it to perform well" was used. From faith, the item "Even if I have no reason to expect the system will be able to identify IED emplacement activity, I still feel certain that it will" was used. From personal attachment, the item "I feel a sense of attachment to using the system" was chosen. The items were scored on a 10-point Likert scale (0 = does not describe participant, 12 = very descriptive of participant). The full-length instrument was validated by Dolgov and Kaltenbach (2017) using principal components analyses and demonstrated high reliability (Cronbach's alpha = .94) and high inter-rater reliability (Cronbach's kappa = .83). The abbreviated instrument was analyzed following data collection (n = 282), and demonstrated high internal reliability, Cronbach's alpha = .88. The confidence item that was added to Madsen and Gregor's subconstructs appeared to contribute to the reliability of the instrument, since reliability analyses showed that removal of the item would lower Cronbach's alpha from .88 to .87.

Offline Experience Questionnaire

At the end of the study, participants completed a post-task opinion questionnaire to provide impressions of the task (Appendix D; derived from Long, 2019). The questionnaire gathered participants' impressions of their performance, comfortability with the experiment, motivation, and enjoyment. Participants' average ratings of task difficulty, confidence, and adequacy of task training were calculated, and comments were examined for common themes.

Procedure

After arriving at the laboratory, participants read and signed an Informed Consent Form (Appendix A) specifying the risks and benefits of participating in the study and completed the Demographics Questionnaire (Appendix B). Participants then proceeded through eye tracking calibration. The Tobii system automatically calibrates while participants track a large red dot moving across the screen. After calibration, participants read on-screen instructions explaining the nature of the task, including reliability information. They were then shown an example FMV clip, with a summary of the instructions. Following this, they completed a familiarization trial with a sample task created for training. The investigator explained the task while participants viewed the familiarization trial and ensured that participants felt comfortable with the task before proceeding.

Participants played the role of an ISR analyst tasked with identifying possible IED placement activity in preparation for a presence patrol. Participants were informed in the initial instructions that the automated system used in the task identifies probable IED emplacement activity along future convoy routes and indicates whether it has identified IED emplacement activity (see Appendix E). The instructions described IED emplacement activity as small groups stopping in or near a road, digging and/or unloading objects on the road, and then hastily departing. Participants were instructed to ensure that no IED emplacements are missed, and that missed emplacements could result in friendly and civilian casualties. Participants were told that they could delegate their analysis to the automation, but that the automation was not perfect. Reliability of the automation was described as able to detect objects 95% of the time, and to correctly classify activity 78% of the time. Participants were instructed to respond to each stimulus by clicking on an option ("agree", "disagree", "more information", or "delegate"). The reliability level of the current study was based on the average reliability of several current automated video analysis systems (Table 4). Correct and incorrect trials were randomly interspersed. The automated system was designed to be high in false alarms and low in misses, with a 16% false alarm rate and a 6% miss rate. Some findings suggest that false alarm-prone automation reduces trust more than miss-prone automation (Chen & Terrence, 2009; Dixon et al., 2007), which could constitute a conflict for ISR, where false alarms are preferable to misses. However, differences were found at only the 60% reliability level, which is well below that of the current research design. Therefore, the reliability level of the current study was not expected to significantly influence participants' trust levels.

After the participants indicated that they were comfortable with the task, they began the experiment. Each task stimulus was present for approximately 10 seconds, after which the

Table 4

Reliability Levels of Several Automated Video Analysis Systems

Author	**Year**	**System**	**Reliability (%)**
Poostchi	2017	Spatial Pyramid Context-Aware Tracking	84
Sabokrou et al	2018	Fully Convolutional Neural Network	89
Gunduz, Ongun, Temizel, & Temizel	2016	Density Aware Anomaly Detection	85
Li, Mahadevan, & Vasconcelos	2014	Anomaly Detection and Localization	65
Lu, Shi, & Jia	2013	Spatial Abnormality Detection	82
Reddy, Sanderson, & Lovell	2011	Cell-based Anomaly Detection	68
Roshtkhari & Levine	2013	Spatio-Temporal Compositions	91
Xiao, Zhang, & Zha	2015	Sparse Semi-nonnegative Matrix Factorization	84
Kolsch & Zaborowski	2014	Small Unmanned Aircraft Vehicle Detection	64
Gaszczak, Breckon, & Han	2011	Real-time People and Vehicle Detection	70
Muncaster, Collins, & Waltman	2015	VideoPlus-Aware	82
Average			**78.5**

participant chose to either agree or disagree with the automated analyst, request more information, or delegate the choice to the automation. If the participant chose to request more information, the investigator reminded them that the information would not be immediately available. If the participant did not choose within the 10 seconds, the next stimulus or questionnaire was presented. If this happened, participants were reminded to make a choice as soon as they felt confident to do so. Following each unique stimulus presentation, participants responded to the 6-question online trust questionnaire (Appendix C). Participants had unlimited time to complete the questionnaire. Upon completing the experiment, participants completed the offline experience questionnaire (Appendix D). They were then debriefed concerning the purpose of the experiment and thanked for participating.

CHAPTER 3

RESULTS

Data Cleaning

Accuracy, decision time, and task shedding data were analyzed using the IBM Statistical Package for the Social Sciences (SPSS) Version 26 software. Descriptive statistics for all the dependent variables are listed Table 5. Outliers were identified using box plots. Accuracy only had one outlier, whose accuracy was less than 20%; task shedding had six outliers, who delegated more than 15% of the time; and decision time had zero outliers. None of the outliers

Table 5

Descriptive Statistics for Dependent Variables

Variable	*N*	*M*	Median	Mode	Min	Max	Range	*SD*
Accuracy	48	.73	.72	.67	.4	1	0-1	.14
Decision Time	864	6.26	6.1	10	.11	10.53	0-10	2.66
Task Shedding	48	.04	0	0	0	.22	0-1	.07
Attention ROI1	864	.29	.25	.33	.03	1	0-1	.20
Trust	283	7.17	7	7	1.5	11.83	1-12	2.11

Note: Range indicates the possible, rather than the actual, range of values were extreme outliers, defined as data points more than 3 standard deviations from the mean. Since the outliers were not extreme, and clearly reflected individual differences in performance, these were retained for further analyses.

Normality of decision time, trust scores, and attention (fixation data) was tested using the Shapiro-Wilk test (Maxwell & Delaney, 2004). Trust and attention were significantly non-normal ($p < .05$). Because ANOVA is robust to deviations from normality, it was used to test hypotheses involving those two dependent variables. Skewness and kurtosis were also tested. Only attention was significantly skewed (> +2). Levene's test for homogeneity of variance was also significant for attention, so the Greenhouse-Geisser adjustment was reported when using ANOVA to compare attention across transparency levels. Gaze fixation data for each ROI was extracted using Tobii Pro Studio 3.2. The recommended Tobii Pro I-VT fixation filter was used to distinguish between fixations and noise.

Data Coding

The accuracy of participants' choices to agree or disagree with the automated system's

Table 6

Participant-Automation Accuracy Matrix

Accuracy	Hit	FA	Miss	Correct Rejection
Agree	1	0	0	1
Disagree	0	1	1	0
More Information	N/A	N/A	N/A	N/A
Delegate	N/A	N/A	N/A	N/A

Note: 1 = accurate, 0 = inaccurate. Overall accuracy = sum (participant choices/correct choices).

recommendations was coded in a binary fashion for each trial, as depicted in Table 6. These categories correspond to the four possible decision states according to signal detection theory (Dixon et al., 2007). Task shedding was coded as the proportion of "delegate" choices within each condition. For eye tracking data, fixation counts on ROI 1 were extracted from the raw gaze data using the Tobii Studio I-VT and calculated as a percentage of the overall fixation count.

Accuracy

To investigate differences in the proportion of accurate decisions across transparency levels, a repeated measure analysis of variance (ANOVA) was conducted. Sphericity was violated, $X^2 = 9.19$, $p = .01$, so the Greenhouse-Geisser correction for degrees of freedom was used. Participants were significantly less accurate ($M = .58$, $SD = .49$) in the low-transparency condition compared to the medium ($M = .82$, $SD = .39$) and high ($M = .79$, $SD = .40$) transparency conditions, $F(1.91, 374.75) = 17.94$, partial $\eta^2 = .08$, $p < .001$ (Table 7). There was no significant difference between the medium and high levels, $p > .05$. These results supported Hypothesis 1 (Figure 6).

Haze also significantly affected accuracy in the expected direction, with higher accuracy in the zero-haze condition ($M = .80$, $SD = .40$) than in the hazy condition ($M = .67$, $SD = .47$),

Table 7

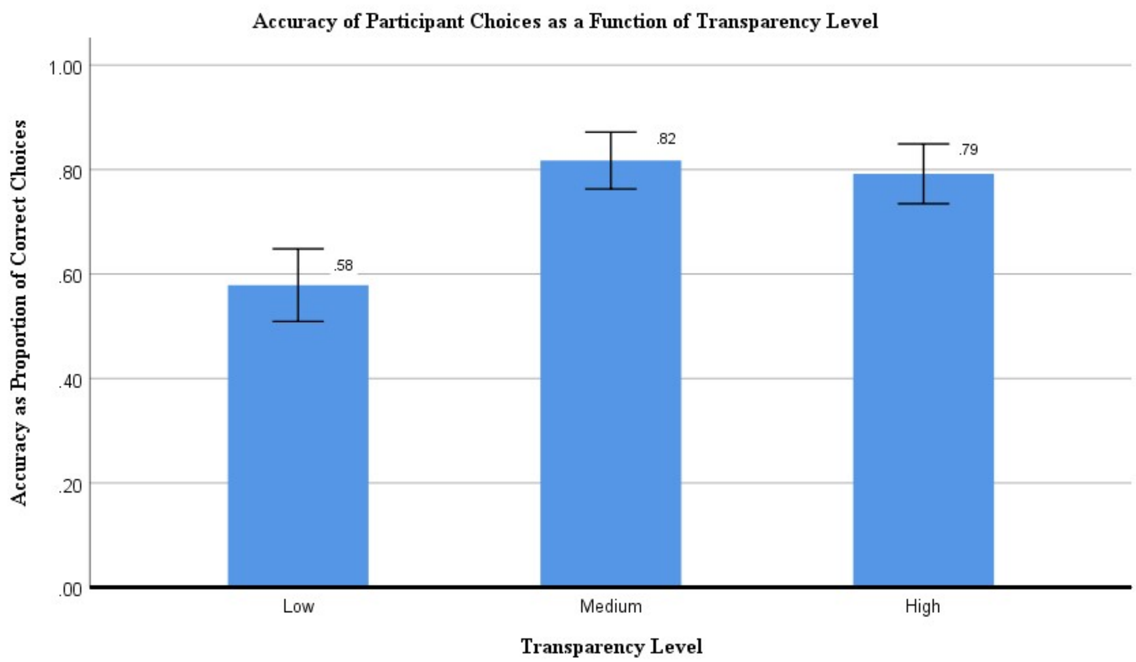
Repeated Measures ANOVA results for accuracy as a function of transparency

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2
Transparency	6.77	1.91	3.54	17.94	.000***	.08
Error	73.90	374.75	.20			

***$p < .001$

Figure 6

Proportion of Participants' Accurate Agree/Disagree Choices Across Transparency Levels



Note. Error bars represent standard error.

although with a small effect size, Cohen's d = .30, $t(294) = 3.51$, p = .001 (Figure 7).

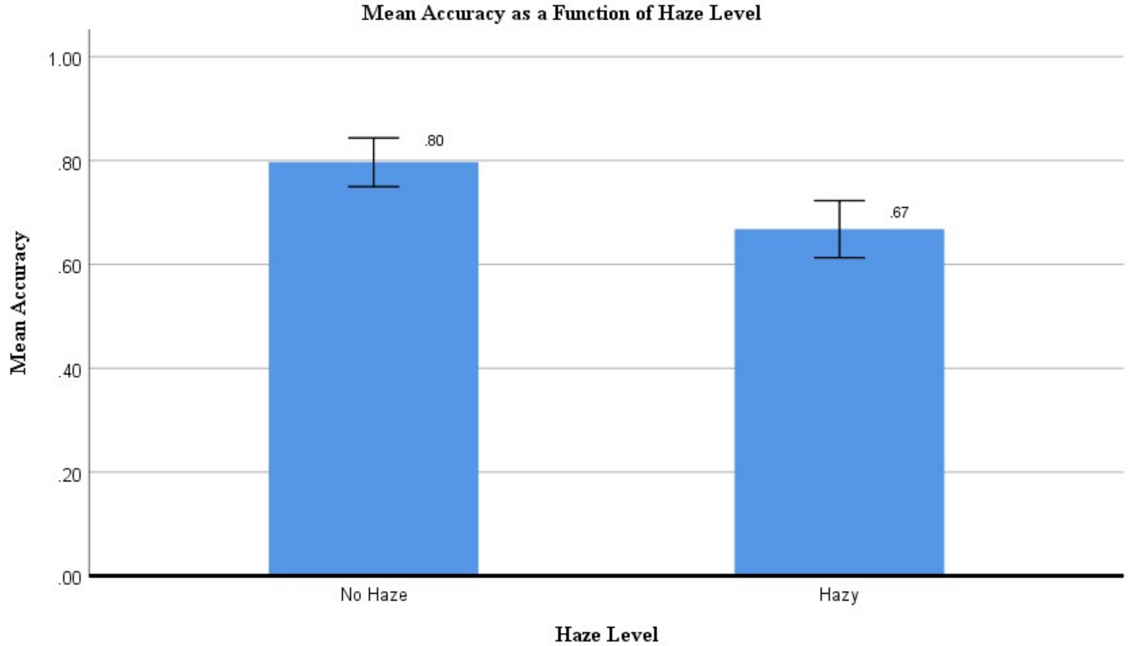
Decision Speed

A paired samples t-test was conducted comparing decision time in seconds of participants' choices across the no haze and hazy conditions (Figure 8). There was no significant difference in decision times between no haze (M = 6.23, SD = 2.74) and hazy (M = 6.29, SD = 2.59) conditions, $t(431) = -.368$, Cohen's d = .02, p = .713. Decision time and trust were also not significantly correlated, $r(282) = -.048$, p = .212. Therefore, Hypothesis 2 was not supported. A limitation of this test was a probable ceiling effect due to the 10 second limit on participants' choices, as 15% of all stimuli presented timed out before the participants made a choice.

In these cases, the decision time was recorded as 10 seconds. Although pilot testing of 10

Figure 7

Participants' Average Accuracy Across Haze Levels



Note. Error bars represent standard error.

Figure 8

Decision Time in Seconds of Participants' Choices Across Haze Levels

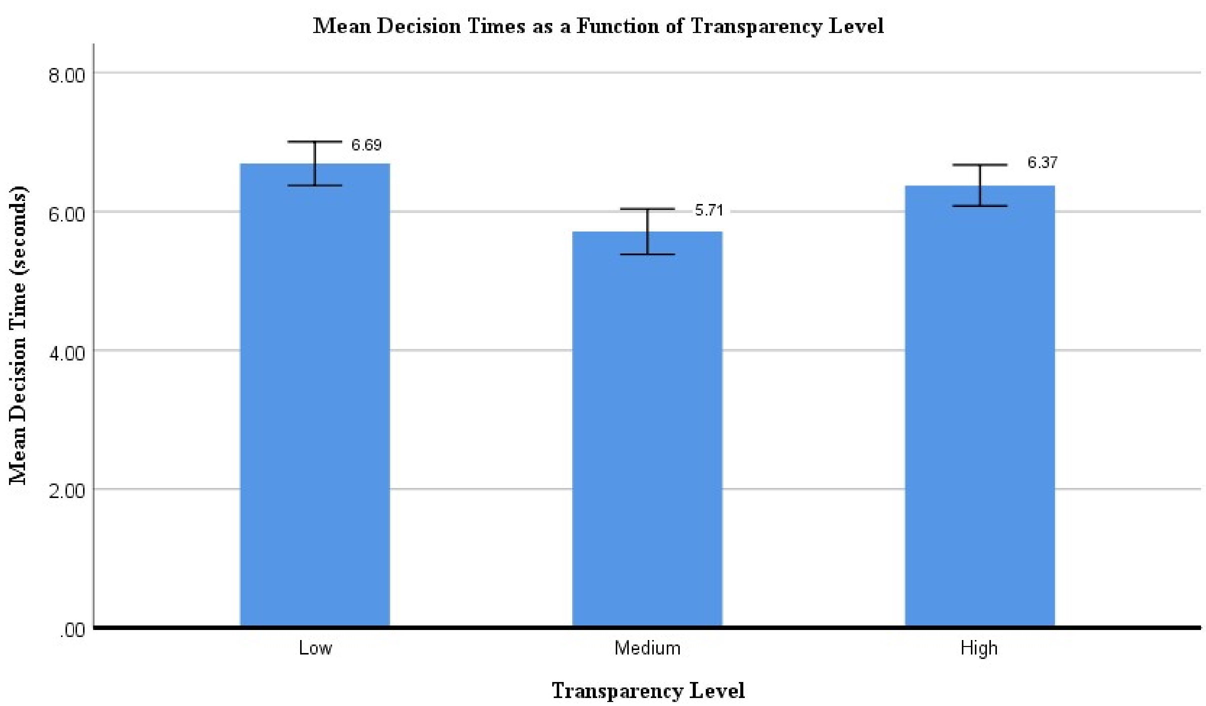
Note. Error bars represent standard error.

subjects suggested that the 10-second time limit was acceptable, some later participants struggled with the task. Additionally, missing data was unequally distributed across conditions. Participants missed (failed to respond within 10 seconds) 12% more stimuli in the hazy condition than in the zero-haze condition, and 41% more in the low transparency than in the medium transparency conditions. Re-comparing decision times and trust without these two conditions still failed to achieve significance but changed the sign of the correlation from negative to positive, $r(141) = .04$, $p = .642$; and the data still showed a strong ceiling effect.

A repeated measures ANOVA was also conducted to compare decision times across transparency levels (Figure 9). Mauchly's w was not violated, $\chi^2(2, n = 282) = 3.76$. Decision

Figure 9

Decision Time in Seconds of Participants' Choices Across Transparency Levels



Note. Error bars represent standard error.

Table 8

Repeated measures ANOVA results for decision time as a function of transparency

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2
Transparency	141.00	2	70.49	12.46	.000***	.04
Error	3181.13	562	5.67			

***$p < .001$

times were significantly lower (faster) in the medium transparency condition (M = 5.71, SD = 2.75) compared to the low (M = 6.69, SD = 2.66) and high (M = 6.37, SD = 2.49) transparency conditions, $F(2, 562) = 12.46$, partial $\eta^2 = .042$, $p < .001$ (Table 8).

Task Shedding

A 2x3 repeated-measures ANOVA was conducted to investigate the effects of transparency and haze on task shedding (as a proportion of each participant's choices). The interaction of haze and transparency to predict task shedding was not significant, $F(2, 74.21)$ = .33, partial $\eta^2 = .007$, $p = .673$ (Table 9). The main effect of haze was significant, $F(1, 46)$ =

Table 9

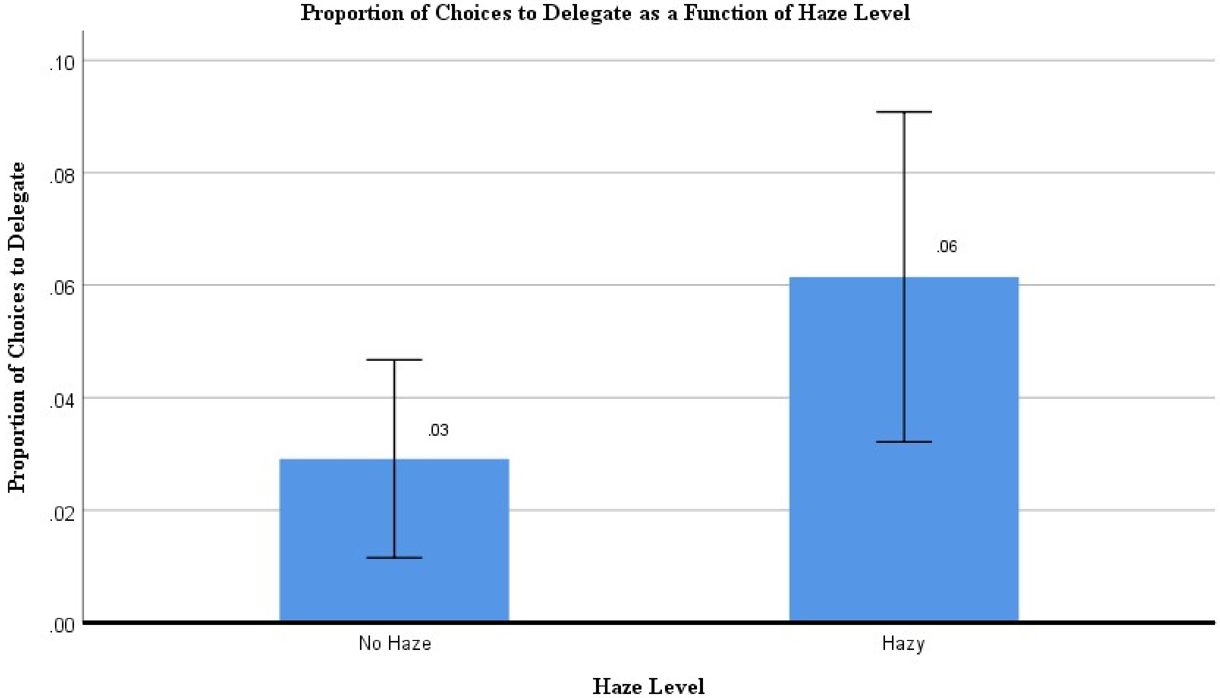
Repeated measures ANOVA results for task shedding as a function of haze and transparency

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2
Haze	.07	1	.07	4.55	.038*	.09
Transparency	.04	2	.02	1.73	.183	.04
Haze x Transparency	.01	1.61	.01	.33	.673	.01
Error	1.51	74.21	.02			

*$p < .05$

Figure 10

Proportion of Task Shedding (Participants' Delegate Choices) Across Haze Levels

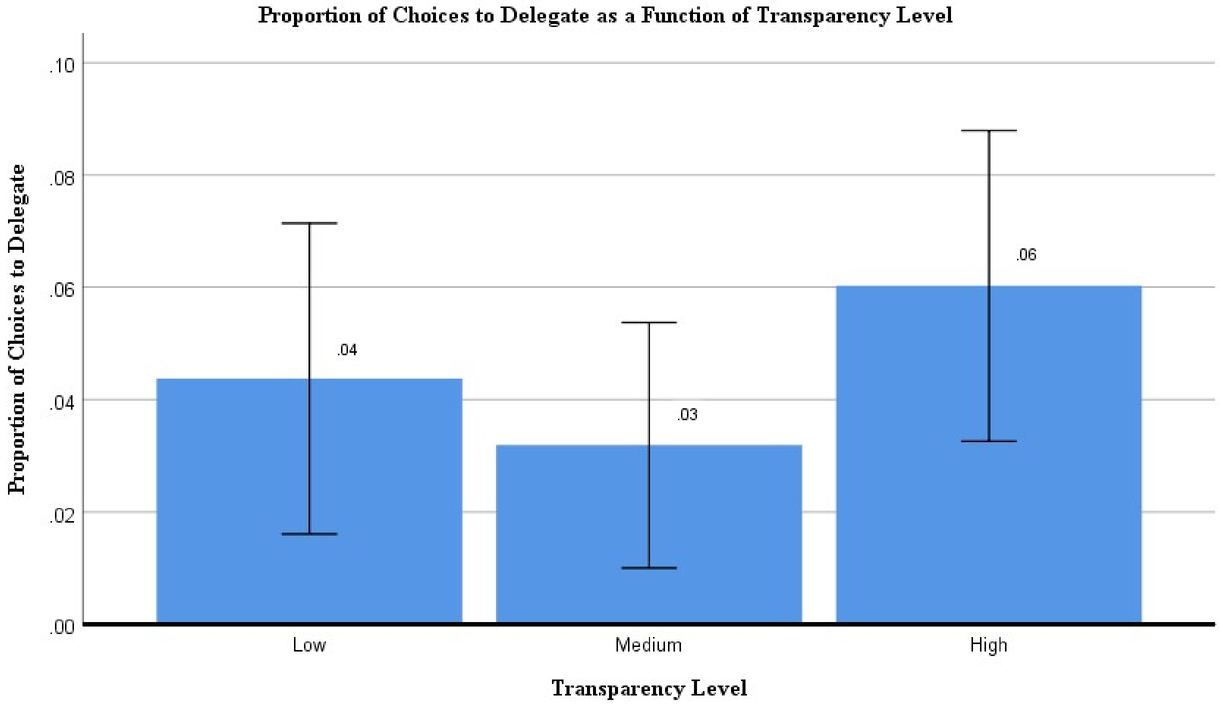


Note. Error bars represent standard error.

4.55, partial η^2 = .09, p = .038. Participants were significantly more likely to task shed in the hazy (M = .06, SE = .015) than in the non-hazy (M = .03, SE = .009) condition (Figure 10). The effect of transparency was not significant, $F(2, 92)$ = 1.73, partial η^2 = .036, p = .183. Although participants were more likely to task shed in the high transparency (M = .06, SE = .014) than in the low transparency (M = .04, SE = .014) condition, this effect did not reach significance (Figure 11). Therefore, Hypothesis 3 was only partly supported. A limitation of this result was the low overall proportion of delegate choices, resulting in a high standard error relative to task shedding proportion.

Figure 11

Proportion of Participants' Delegate Choices Across Transparency Levels



Note. Error bars represent standard error.

Attention

To test Hypothesis 4, the scanning behavior in terms of proportion of fixation duration (also called percent dwell time) within each ROI was compared to the optimal scanning behavior predicted by the SEEV-derived equation in Figure 5c (presented below), and to self-reported trust scores. To determine the optimal scanning proportion, model parameters were rank-ordered across conditions by ROI (Table 10; see Fig. 3 for ROIs). The parameters were assigned scores between one and five, following Wickens et al.'s (2003) methodology. For example, ROI 1 provides the most task-relevant information, but imagery degradation reduces information available. Therefore, the highest parameter value was assigned to ROI 1 when degradation was

Figure 5c

Visual Attention Model, repeated from Fig. 5

$$Visual\ Attention\ to\ AOI = \sum_{task=}^{n} \left[(BW\ x\ R_{xy}\ x\ V) + 2U\right]$$

Table 10

Visual Attention Model Parameter Values

	Bandwidth		Relevance		Value		Uncertainty	
Haze	None	Hazy	None	Hazy	None	Hazy	None	Hazy
ROI 1	5	4	5	4	5	4	4	5
ROI 2	3	3	3	2	3	3	3	3
ROI 3	2	2	4	3	2	2	2	2
ROI 4	1	1	2	1	1	1	1	1

Note. DVE = Degraded Visual Environment; low= low haze, high = high haze.

low. The FMV feed had the highest uncertainty compared to the other sources of information provided, and degradation also increased uncertainty. Therefore, the highest uncertainty value was assigned to ROI 1 when degradation was high. These scores, expressed as a proportion of the sum of all parameter values, represented optimal scanning behavior as percent of fixation durations.

Two predicted values were generated from each of the three models, the proportion of attention that should be paid to ROI 1 in the hazy condition and in the no haze condition. The difference between optimal scanning behavior, derived from the proposed visual attention model, and observed scanning behavior was calculated as the absolute value of the *predicted* proportion of fixation duration on ROI1 subtracted from the *observed* proportion of fixation duration on ROI1. This difference was also calculated using the predicted values from Wickens' and

Johnson's mathematical models. To test whether the proposed model approximated optimal attention, predicted-observed attention differences were regressed onto accuracy. Neither the proposed model, $\chi^2(1, n = 872) = .17$, $p = .680$, Wickens' model, $\chi^2(1, n = 872) = .234$, $p = .629$, nor Johnson's model, $\chi^2(1, n = 872) = .01$, $p = .918$, were significant.

Bivariate correlations were conducted to determine whether trust correlated with observed attention allocation to ROI1, and whether trust varied predictably with the difference between the predicted, optimal and the observed proportion of attention to ROI1. Attention was measured as the proportion of time (in seconds) that gaze fixations were directed to ROI1. To correlate attention with trust, gaze data was extracted from the six stimuli preceding the six HCTS trust questionnaires. This data was significantly, negatively skewed, so the data was transformed by exponentiating attention, reducing skew to an acceptable .06. Trust did not significantly correlate with observed attention, $r(264) = -.007$, $p = .913$. Correlations were also performed between the difference in predicted-observed attention and trust for each of the three models of attention. None of the models were able to significantly predict trust from the difference in predicted and observed scanning behavior, although the proposed model had a slightly higher r value ($r[264] = .05$) than Wickens' ($r[264] = .04$) or Johnson's ($r[264] = .02$) models.

A Chi-square test was performed to compare observed and expected proportions of attention by region of interest. The observed distribution of attention differed significantly from equal, $\chi^2(1, n = 48) = 24.24$ (i.e., participants paid more attention to ROI1 than to the other regions of interest). Neither the proposed model nor Wickens' model adequately predicted attention; predicted attention from the proposed model ($\chi^2(1, n = 48) = 29.45$) and Wickens' model ($\chi^2(1, n = 48) = 74.27$) significantly differed from observed attention. Predicted values from Johnson's model did not differ significantly from observed attention, $\chi^2(1, n = 48) = 4.25$, $p < .05$. Although these results suggest that Johnson's model of attention adequately *describes* scanning behavior, the question of a good *prescriptive* model of attention remains open.

A one-way, repeated measures ANOVA was conducted to compare observed attention in seconds to ROI1 (after square root transformation to reduce skew and kurtosis) across visual conditions (no haze, hazy). Participants paid significantly more attention to ROI1 in the hazy (M = 3.09 seconds, SD = 1.03) than in the no haze condition (M = 2.94, SD = .90), $F(1, 377) = 6.09$, partial $\eta^2 = .016$, $p = .014$ (Table 11; Figure 12). Therefore, Hypothesis 4, which predicted that participants would pay more attention to the most task-relevant area in the degraded condition, was supported, albeit with a low effect size. However, the ten-second limit applied to this experiment should be considered when considering the effect size.

Trust

Validation of trust measures

Because task shedding and accuracy were both dichotomous and trust was not normally distributed, binary logistic regression was used to test whether trust predicted task shedding or accuracy (Table 12). The logistic regression model with trust predicting task shedding was not

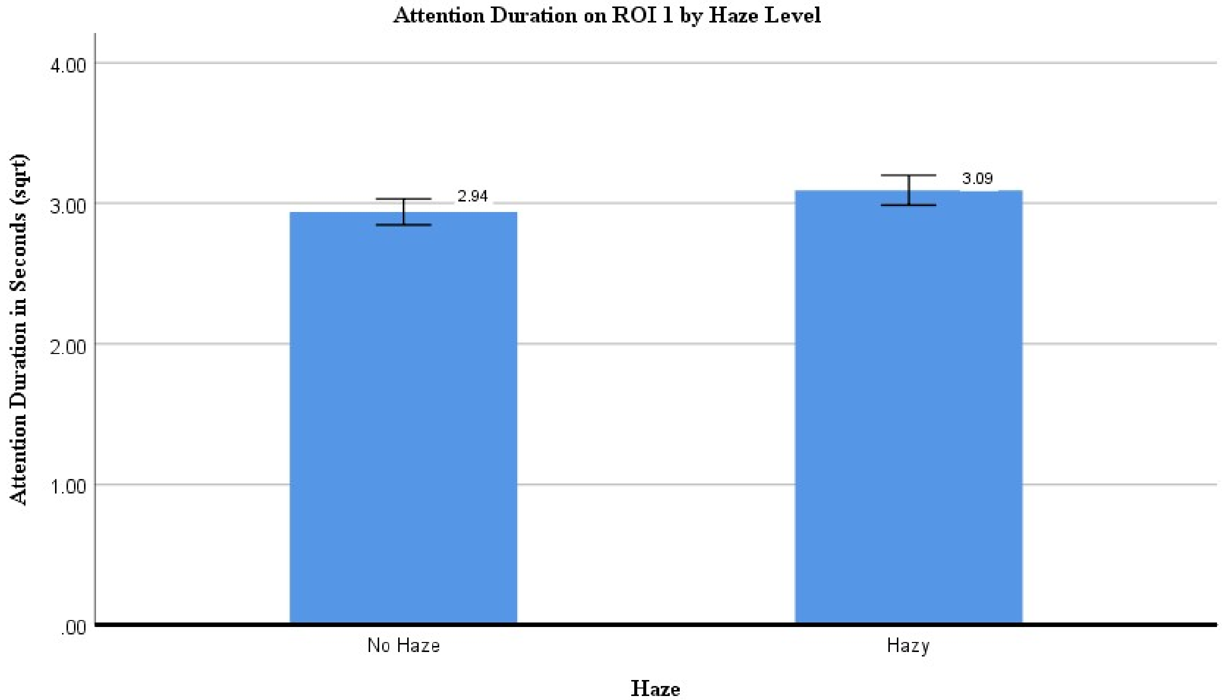
Table 11

Repeated measures ANOVA results for attention as a function of haze

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial η^2
Haze	4.52	1	4.52	6.09	.014	.016
Error	279.86	377	.74			

Figure 12

Average Fixation Duration in Seconds on ROI 1 Across Haze Levels



Note. Error bars represent standard error.

Table 12

Logistic regression results predicting task shedding and accuracy from trust

DV	B	SE	Wald	*p*	Exp(B)
Task Shedding	-.23	.19	1.43	.231	.80
Accuracy	.19	.09	4.80	.028*	1.21

*$p < .05$

significant, $\chi^2(1, n = 282) = 1.49$, Cohen's $d = .15$, $p = .222$. The logistic regression model with trust predicting accuracy was significant, $\chi^2(1, n = 220) = 4.99$, Cohen's $d = .30$, $p = .025$. Trust explained 4% (Nagelkerke R^2) of the variance in accuracy, and 82.7% of cases were correctly

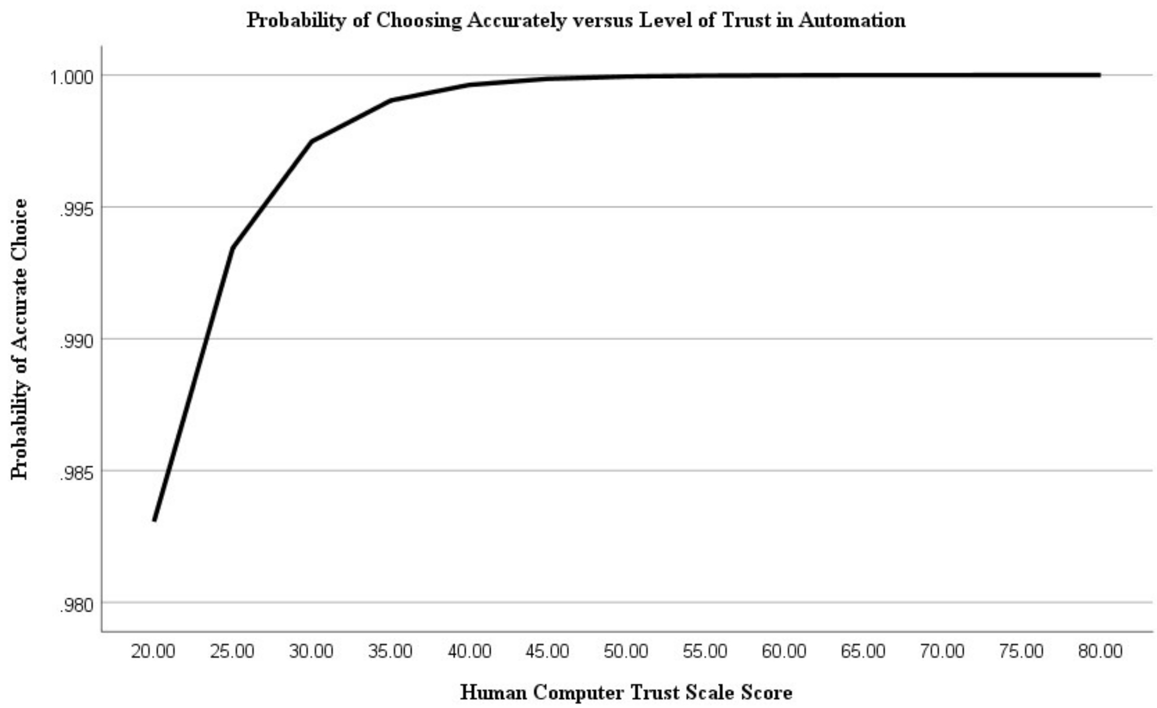


Figure 13. Probability of choosing accurately as a function of trust score.

classified. For each one-point increase in trust score, the odds of making an accurate decision increased by a factor of 1.21 (Figure 13). The low explanatory power of the model, however, suggests that other, unexplored factors contributed significantly to accuracy. These results suggest that neither accuracy nor task shedding are good indirect measures of trust, but that optimal levels of trust likely improve task accuracy.

Effects of degraded visual environment and transparency on trust

Issues with the Tobii software used to present the experiment resulted in an unequal distribution of trust questionnaires across conditions. Five of the six questionnaires were presented following a zero-haze stimulus, two following a low transparency stimulus, three

Table 13

Mixed effects model results for trust as a function of haze and transparency

Predictor	*df* Num	*df* Denom	*F*	*p*
Haze	1	235.06	.004	.949
Transparency	2	223.56	.64	.529

following a medium-transparency stimulus, and only one following a high-transparency stimulus. Therefore, a mixed effects model, which is more robust to unequal sample sizes, was conducted to determine the effects of haze and transparency on trust. Neither haze, $F(1, 278) = .002$, $p = .966$, nor transparency, $F(1, 278) = 1.43$, $p = .242$, significantly predicted trust (Table 13). However, due to the severe imbalance of data across conditions, these results be flawed.

Participant feedback

Participants' ratings of task confusion, perceived performance, task understandability, and motivation to perform the task were neutral (scored 3, on a 5-point Likert-type scale, on average). However, a frequency-based content analysis of participants' comments revealed a strong emphasis on confusion, particularly at the beginning of the experiment. Participants reported that task familiarization time was inadequate ($M = 2.5$, $SD = 1.13$). Participants also reported high effort on the task ($M = 4.4$, $SD = .53$) on average. Other themes that emerged from the comments were that the videos did not provide sufficient time to decide, and that several participants struggled with the degraded imagery.

Task confusion and difficulty with the short stimulus time was not evident during the pilot testing of the experiment, which was conducted with 10 participants. After these issues began to emerge, the experiment was adapted with the addition of an untimed example task and an additional experimenter script explaining the task and task interface components.

CHAPTER 4

DISCUSSION

Haze and Transparency

The current study investigated the effects of degraded visual environments (operationalized as haze) and transparency of automation on task accuracy, speed, and trust in automation. These variables were investigating using an FMV analysis task with a simulated automated analyst with a medium level of automation (Level 5; Sheridan & Verplanck, 1978), which offered participants a choice and executed the user's decision. The results reinforced previous findings (e.g. Lyons et al., 2014), suggesting that when the user has too little information about how the automation is operating, task accuracy is negatively affected. However, the highest transparency condition reduced participants' speed but not their accuracy, suggesting that the type of information and how it is presented may be more important than the amount of information presented. The middle level of transparency may have improved situational awareness compared to low transparency, and reduced cognitive load compared to high transparency. As the amount of information increases, situation awareness increases, facilitating decision-making (Chen et al., 2014). However, the cognitive load also increases, reducing task speed (Wright et al., 2017). At the optimal level of information, the opposing effects of situation awareness and cognitive load may intersect to predict the highest level of efficiency.

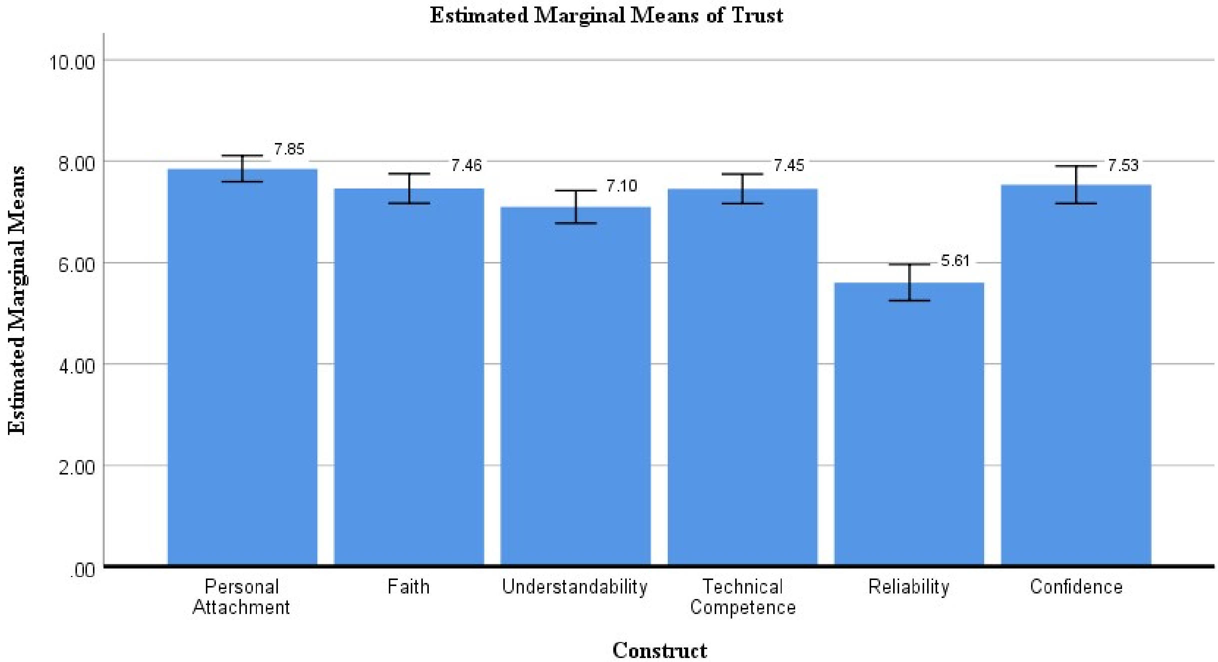
As expected, degraded imagery negatively impacts participants' accuracy (but not decision speed). However, unlike MacMillan et al.'s (1994) findings, this reduction in task accuracy did not coincide with reduced trust in automation under degraded conditions. Additional research is needed to confirm whether there is a relationship between degraded visual conditions and trust. If degraded visual conditions do not affect trust, then negative performance impacts could potentially be mitigated by designing automation to optimize trust levels. In fact, neither the haze nor the transparency of automation conditions were found to predict trust; trust did not significantly vary across conditions. Since performance and trust did not vary together across conditions, the expected mechanism of haze and transparency impacting performance by affecting trust could not be confirmed. Instead, since trust explained 4% of the variance in accuracy without differing significantly across conditions, trust appeared to moderate the effects of haze and transparency. It is also possible that the transparency information, in turn, moderated the effect of haze on decision speed. Trust in automation has been previously placed in role of a causal variable, explaining user performance degradations via complacency, reduced situation awareness, and heuristic decision making (e.g., Wickens et al., 2007). There was no evidence of complacency in the current study, perhaps because the task was sufficiently complex, interesting, and brief that participants remained engaged throughout. If the experiment had been longer and less complex, direct effects of trust on performance via complacency would probably have emerged. The results did at least suggest that priming reliability was effective in calibrating users' trust levels.

Measuring Trust

The measure of trust used averaged items intended to tap into each of five subconstructs of trust: personal attachment, faith, understandability, technical competence, and reliability; and one additional item tapping into the overall definition of human-computer trust (the user's confidence in their decisions while using automation). Post-hoc analyses of the abbreviated

Figure 14

Average Scores on Sub-Constructs of Trust



Note. Error bars represent standard error.

Human-Computer Trust Scale revealed significant differences among many of these sub-constructs, Greenhouse-Geisser $F(3.32, 4{,}528.52) = 56.09$, partial $\eta^2 = .17$, $p < .001$. For example, personal attachment and confidence were significantly higher, while reliability scored significantly lower (Figure 14). Future research should consider the role of these different sub-constructs in predicting performance and optimal usage of automation. Additionally, the limitations of the current study should be kept in mind when interpreting this result. Unequal sample sizes across conditions and the relatively small number of trials may have affected the haze and transparency analyses.

Since workload should be higher in the hazy condition, participants were expected to rely more on the automation if their level of trust was appropriately calibrated (e.g., Bliss, Harden, & Dischinger, 2013). As expected, participants were much (2.24 times) more likely to task shed in the hazy condition compared to the no-haze condition; however, trust did not significantly predict task shedding. Despite the lack of an apparent effect due to trust, haze explained little of the variance in task shedding. Transparency of automation was also expected to affect task shedding, via its effect on trust. Since trust did not vary significantly across transparency levels, it is unsurprising that transparency did not significantly predict task shedding. Some participants may not have chosen to task shed because they had little trouble identifying IED emplacement activity in the videos; however, some (6.25%) of participants commented that degradation in the videos made it difficult to see what was going on. Future research should explore what other factors contribute to participants' motivation to task shed during automation-assisted tasks. Since few participants chose to task shed in any condition, experiments better designed to elicit task shedding may clarify the relationship between haze, transparency, task shedding, and performance. For example, despite a greater tendency to task shed in the hazy condition, participants' accuracy suffered significantly.

Task shedding and decision time also were not supported as good measures of trust. Neither task shedding nor decision time correlated significantly with trust. Decision time was expected to correlate with trust since perceived reliability of automation affects both users' trust and decision-making strategies (e.g., Rovira et al., 2007). The lack of a relationship may have been due to priming reliability, so that reliability was not perceived to vary; or may have been masked by the effects of task complexity. The lack of a linear relationship between task shedding and trust differs from previous findings and is difficult to explain. One reason for this result could have been the relatively few samples in which participants chose to delegate (4.5% of all choices). Additionally, missing data was unequally distributed across conditions. Participants missed (failed to respond within 10 seconds) more of the degraded than non-degraded stimuli, and more of the zero-transparency stimuli compared to medium or high transparency levels. This coincides with the results for accuracy, supporting the hypothesis that participants found degraded visuals and lack of information more challenging. It is possible, however, that this imbalance affected some analyses. Additionally, the experiment may have been too short to allow participants to establish stable trust levels, despite priming reliability. For example, Yang et al. (2017) found that participants' trust levels did not stabilize until they had completed 40-80 trials. In the current study, participants completed just 18 trials.

Attention

One purpose of this study was to validate a predictive model of attention derived from Horrey et al.'s (2006), Johnson et al.'s (2017), and Wickens et al.'s (2003) models. The proposed model adopted bandwidth, relevancy, and value from Wickens' model and value and uncertainty from Johnson's model, because they were the highest-loading components. Neither the proposed model nor Wickens' nor Johnson's models were supported, either in terms of predicting accuracy based on participants' deviation in scanning behavior from optimal, or terms of predicting trust. Johnson's model proved to be a significant predictor of participants' actual scanning behavior. However, additional research is needed to determine a good prescriptive model of attention capable of predicting trust and performance.

Though non-significant, observed attention did appear to depart more from optimal at higher and lower levels of trust. This aligns with the attention model of trust, which predicts higher levels of task-relevant attention at lower levels of trust and lower levels of attention at higher levels of trust (Parasuraman & Manzey, 2010). Additional research is merited to determine the relative efficacy of this model of attention in predicting scanning behavior under different conditions.

The expected relationship between trust and attention allocation was not found. Previous research (Louw & Merat, 2017) suggested a mediated path from degradation to attention via trust. The current results partially supported this model, as attention to ROI 1 was significantly higher in the hazy condition. However, trust did not significantly correlate with observed attention. Considering that trust also did not vary significantly across haze or transparency conditions, it seems likely that some aspect of this experiment resulted in a lack of variance in trust.

Implications

The findings suggest that fast-paced, complex, and interesting tasks such as FMV analysis may be less vulnerable to the negative effects of mistrust and negation of trust in automation. However, there is a need to establish design criteria that will facilitate task shedding. There is limited research on this specific construct, so many of the factors that may affect task shedding behavior have not been identified. Additionally, there appears to be a complex relationship between transparency, degraded visual environments, and trust, which merits additional research.

The current results suggest that optimal levels of transparency may not be dependent on the task or environment. These results also support Chen's (2014) model of transparency; specifically, Chen's proposed Level 2 transparency resulted in the best task performance and efficiency. Therefore, transparency information should include purpose, process, performance, reason, algorithm, and environment information. Chen's recommendation to adapt transparency information to the greatest task relevancy possible seemed to be effective.

Accuracy was the most supported measure of trust, whereas it was unclear whether decision speed may reflect trust levels. These are also areas that would benefit from additional research, especially where trust measures are compared along sub-constructs. For example, perceived reliability likely measures situational trust, personal attachment may reflect dispositional trust, and confidence may reflect state trust (Balliet & van Lange, 2013). These different facets of trust are influenced by different factors and may have a differential effect on performance.

Limitations

Because 15% of all stimuli disappeared before the participant made a choice, an artificial ceiling was imposed on the decision time data. Those missing stimuli were replaced with 10 seconds for analysis purposes, but participants may have taken longer if given the opportunity. This possibility was supported by participant feedback, with frequent (10.4% of participants) comments that the amount of time provided was insufficient. For example, one participant comment that they felt "rushed", and another wrote "countdown timer" in the comment section. Although the investigator tried to stress accuracy and time equally, some participants were more concerned with time, which may have stressed them more. Other participants seemed less concerned with time, with the result that they missed many stimuli. Since this issue did not become apparent until late in data collection, however, the time frame could not be adjusted. However, future research should provide more liberal time limits for complex tasks.

Participants often (23%) reported being confused, especially at the beginning of the experiment. This issue emerged soon after pilot testing and was partially rectified by the addition of an untimed trial and a task walk-through by the investigator. Even with this additional measure, some participants continued to be confused by the task, despite indicating that they were comfortable with the explanation. One participant reported, “This was a little confusing. Honestly I wasn’t sure which side of the screen to focus on.” Another commented, “I felt like I understood the instructions but once I started I felt as though I didn’t know what I was doing.” Participants who reported being confused did not appear to differ significantly on relevant demographic factors from those who did not report being confused.

Participants were mostly (73%) women, which may have affected both performance and trust results. For example, gaming experience has been connected with performance in automated tasks such as UAV control (Lin et al., 2016). In the current sample, women reported one hour per week of gameplay on average, while men reported 6.65 hours per week on average. Sex may also influence trust in automation. For example, Nomura (2016) found that women have more negative attitudes towards automation than men do, on average.

Other limitations include the number of trials, which may have been too few to establish stable trust levels for participants, and unequal sample sizes for the trust measure across conditions.

Conclusion

There is a dearth of empirical studies examining the effects of degraded visual environments within automation, especially in the ISR field. Although Macmillan et al. (1994) explored this construct early on, they and Narayanaswami et al. (2010) considered only transmission factors, such as image distortion, brightness, and resolution, rather than environmental factors such as haze, and they did not include an investigation of transparency. The current research suggests that both imagery degradation and amount of transparency of automation significantly affect accuracy and task shedding behavior, whereas transparency also significantly affects decision speed, and haze significantly affects attention. However, additional research is needed concerning the relationship between visual degradation, transparency of automation, and trust in automation.

Automation within the ISR field is growing in ubiquity and sophistication; however, it is still vulnerable to failures, necessitating a role for human monitoring and intervention (Atwood, 2015; Cardillo, 2016, 2017). An important prerequisite for appropriate use of automation is well-calibrated trust, a concern that will continue as technology moves toward full autonomy. Meanwhile, many in the military community resist automated ISR, largely due to distrust engendered by the small fields of view of modern UAVs (Patrick, 2015). Understanding how to mitigate these effects is essential. Future researchers and user interface designers should carefully consider design criteria to facilitate task shedding and, therefore, proper use of automation as well as more accurate data collection. The current findings suggest that an optimal level of automation transparency is critical to user trust and performance, and should include purpose, process, performance, reason, algorithm, and environment information. The content of these categories will vary for different tasks but should be as relevant as possible. While the current study focused on negation of trust as a holistic construct, future researchers should seek to refine their analyses, targeting the constituent elements of trust.

REFERENCES

Acharya, A., Howes, A., Baber, C., & Marshall, T. (2018). Automation reliability and decision strategy: A sequential decision making model for automation interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 62*, 144–148. Los Angeles, CA: SAGE Publications.

Atwood, C. (2015). Activity-based intelligence: Revolutionizing military intelligence analysis. National Defense University Press. *Joint Forces Quarterly*, *77*. Retrieved from https://ndupress.ndu.edu/Media/News/Article/581866/activity-based-intelligence-revolutionizing-military-intelligence-analysis/

Bailey, N. R., & Scerbo, M. W. (2007). Automation-induced complacency for monitoring highly reliable systems: The role of task complexity, system experience, and operator trust. *Theoretical Issues in Ergonomics Science*, *8*(4), 321–348. Retrieved from https://doi.org/10.1080/14639220500535301.

Balliet, D., & Van Lange, P. A.M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin, 139*(5), 1090-1112. Retrieved from https://doi.org/10.1037/a0030939.

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, *13*(2), 173–189. Retrieved from https://doi.org/10.1023/B:GRUP.0000021840.85686.57.

Bliss, J. P., Harden, J. W., & Dischinger, H. C. (2013). Task shedding and control performance as a function of perceived automation reliability and time pressure. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 635–639. https://doi.org/10.1177/1541931213571136

Chen, J., Lakhmani, S., Stowers, K., Selkowitz, A., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282. https://doi.org/10.1080/1463922X.2017.1315750

Chen, J., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). *Situation awareness-based agent transparency* (No. ARL-TR-6905). Army Research Lab Aberdeen Proving Ground Md Human Research and Engineering Directorate. Retrieved from http://www.dtic.mil/docs/citations/ADA600351

Chen, J. Y. C., & Terrence, P. I. (2009). Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics*, *52*(8), 907–920. https://doi.org/10.1080/00140130802680773

Chung, J., & Wark, S. (2016). *Visualising uncertainty for decision support*. Defence Science and Technology Group, Fisherman's Bend, Victoria, Australia. Retrieved from http://www.dtic.mil/docs/citations/AD1027341

Cohen, J. (1973). Brief notes: Statistical power analysis and research results. *American Educational Research Journal*, *10*(3), 225–229. https://doi.org/10.3102/00028312010003225

Cook, M. J., Angus, C. S., & Adams, C. S. (2004). Human factors of intelligence analysis. *Human Performance, Situation Awareness and Automation: Current Research and Trends*, *2*, 286–293.

Cordova, A., Millard, L. D., Menthe, L., Guffey, R. A., & Rhodes, C. (2013). *Motion Imagery Processing and Exploitation (MIPE)*. Rand Project Air Force Santa Monica CA. Retrieved from https://apps.dtic.mil/docs/citations/ADA591428

Dimoka, A. (2010). What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly*, *34*(2), 373–396. Retrieved from https://doi.org/10.2307/20721433.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*(4), 564–572. Retrieved from https://doi.org/10.1518/001872007X215656.

Dolgov, I., & Kaltenbach, E. K. (2017). Trust in automation inventories: An investigation and comparison of the human-computer trust and trust in automated systems scales. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1271–1275. Retrieved from https://doi.org/10.1177/1541931213601799.

Endsley, M. R. (1996). Automation and situation awareness. In *Automation and human performance: Theory and applications,* 163–181. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492. Retrieved from https://doi.org/10.1080/001401399185595.

Endsley, Mica R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64. Retrieved from https://doi.org/10.1518/001872095779049543.

Gaszczak, A., Breckon, T. P., & Han, J. (2011). Real-time people and vehicle detection from UAV imagery. In *Intelligent Robots and Computer Vision XXVIII: Algorithms and Techniques, 7878*, 78780B. International Society for Optics and Photonics. Retrieved from https://doi.org/10.1117/12.876663.

Getty, D. J., Swets, J. A., Pickett, R. M., & Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of Experimental Psychology: Applied*, *1*(1), 19–33. Retrieved from https://doi.org/10.1037/1076-898X.1.1.19.

Gilmore, J. M. (2016). *Distributed Common Ground System-Army (DCGS-A) Increment 1 Release 2 Follow-on Operational Test and Evaluation (FOT and E) Report*. Washington, DC: Director, Operational Test and Evaluation (DOT and E).

Guevara, M. A., Brockmeier, W. R., Kuyk, T. K., Smith, P. A., Goettl, B. P., & Novar, B. J. (2017). *Light scatter in optical materials: Advanced haze modeling* (No. AFRL-RH-FS-TR-2017-0022). 711th Human Performance Wing JBSA Fort Sam Houston United States. Retrieved from https://apps.dtic.mil/docs/citations/AD1040196

Gunduz, A. E., Ongun, C., Temizel, T. T., & Temizel, A. (2016). Density aware anomaly detection in crowded scenes. *IET Computer Vision*, *10*(5), 376–383. Retrieved from https://doi.org/10.1049/iet-cvi.2015.0345.

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*, *53*(5), 517–527. Retrieved from https://doi.org/10.1177/0018720811417254.

Harguess, J., Shafer, S., & Marez, D. (2017). The role of optical flow in automated quality assessment of full-motion video. In *Applications of Digital Image Processing XL* (Vol. 10396, p. 1039606). International Society for Optics and Photonics. Retrieved from https://doi.org/10.1117/12.2275392.

Hild, J., Voit, M., Kühnle, C., & Beyerer, J. (2018). Predicting observer's task from eye movement patterns during motion image analysis. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications, 58,* 1–5. New York, NY, USA: ACM. Retrieved from https://doi.org/10.1145/3204493.3204575.

Hollock, J. A. (2017). *Flexibility versus expertise: A closer look at the employment of United States Air Force imagery analysts*. Maxwell AFB, AL: Air Command and Staff College, Distance Learning Air University. Retrieved from http://www.dtic.mil/docs/citations/AD1047475.

Hooey, B. L., Toy, J. J. N. T., Carvalho, R. E., Fong, T., & Gore, B. F. (2018). Modeling operator workload for the resource prospector lunar rover mission. In D. N. Cassenti (Ed.), *Advances in Human Factors in Simulation and Modeling,* 183–194. New York, NY: Springer International Publishing.

Horrey, W. J., Wickens, C. D., & Consalus, K. P. (2006). Modeling drivers' visual attention allocation while interacting with in-vehicle technologies. *Journal of Experimental Psychology. Applied*, *12*(2), 67–78. Retrieved from https://doi.org/10.1037/1076-898X.12.2.67.

Itoh, M. (2010). Necessity of supporting situation understanding to prevent over-trust in automation. *International Electronic Journal of Nuclear Safety and Simulation*, *1*(2), 150–157.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. Retrieved from https://doi.org/10.1207/S15327566IJCE0401_04.

Johnson, A. W., Duda, K. R., Sheridan, T. B., & Oman, C. M. (2017). A closed-loop model of operator visual attention, situation awareness, and performance across automation mode transitions. *Human Factors*, *59*(2), 229–241. Retrieved from https://doi.org/10.1177/0018720816665759.

Kahraman, S., & De Charette, R. (2017). *Influence of Fog on Computer Vision Algorithms* (PhD Thesis). Inria, Paris: Hyper Articles en Ligne.

Karpinsky, N. D., Chancey, E. T., Palmer, D. B., & Yamani, Y. (2018). Automation trust and attention allocation in multitasking workspace. *Applied Ergonomics*, *70*, 194–201. Retrieved from https://doi.org/10.1016/j.apergo.2018.03.008.

Khosla, D., Uhlenbrock, R., & Chen, Y. (2017). Automated scene understanding via fusion of image and object features. In *2017 IEEE International Symposium on Technologies for Homeland Security (HST),* 1–4. Retrieved from https://doi.org/10.1109/THS.2017.7943474.

Klein, G. (2008). Naturalistic decision making. *Human Factors*, *50*(3), 456–460. Retrieved from https://doi.org/10.1518/001872008X288385.

Knights, M., & Mello, A. (2017). Defeat by annihilation: Mobility and attrition in the islamic state's defense of mosul. *CTCSentinel*, *10*(4). Retrieved from https://ctc.usma.edu/defeat-by-annihilation-mobility-and-attrition-in-the-islamic-states-defense-of-mosul/.

Kölsch, M., & Zaborowski, R. (2014). Vehicle detection onboard small unmanned aircraft. In B. Kisačanin & M. Gelautz (Eds.), *Advances in Embedded Computer Vision,* 199–215.

Cham, Switzerland: Springer International Publishing. Retrieved from https://doi.org/10.1007/978-3-319-09387-1_9.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., & Nass, C. (2014). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing, 9,* 269-275.

Kopeika, N. S., Dror, I., & Sadot, D. (1998). Causes of atmospheric blur: comment on Atmospheric scattering effect on spatial resolution of imaging systems. *Journal of the Optical Society of America A*, *15*, 3097–3106. Retrieved from https://doi.org/10.1364/JOSAA.15.003097.

Kreitmair, T., & Coman, C. (2014). Video interpretability rating scale under network impairments. In *Image Quality and System Performance XI, 9016*, 90160L. International Society for Optics and Photonics. Retrieved from https://doi.org/10.1117/12.2040862.

Kumari, A., & Sahoo, S. K. (2016). Real time image and video deweathering: The future prospects and possibilities. *Optik*, *127*(2), 829–839. Retrieved from https://doi.org/10.1016/j.ijleo.2015.10.080.

Kuwertz, A., Sander, J., Pfirrmann, U., & Dyck, S. (2017). High-level information management in joint ISR based on an object-oriented approach. In *2017 Sensor Data Fusion: Trends, Solutions, Applications (SDF),* 1–6. Retrieved from https://doi.org/10.1109/SDF.2017.8126360.

Lee, J. D. (2008). Review of a pivotal Human Factors article: “Humans and automation: use, misuse, disuse, abuse.” *Human Factors*, *50*(3), 404–410. Retrieved from https://doi.org/10.1518/001872008X288547.

Li, W., Mahadevan, V., & Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(1), 18–32. Retrieved from https://doi.org/10.1109/TPAMI.2013.111.

Lin, J., Wohleber, R., Matthews, G., Chiu, P., Calhoun, G., Ruff, H., & Funke, G. (2016). Video game experience and gender as predictors of performance and supervisory control of multiple unmanned aerial vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), 746-750. Retrieved from https://doi.org/10.1177/1541931215591175.

Liu, S., Rahman, M. A., Liu, S. C., Yeow Wong, C., Lin, C.-F., Wu, H., & Kwok, N. (2017). Image de-hazing from the perspective of noise filtering. *Computers & Electrical Engineering*, *62*, 345–359. Retrieved from https://doi.org/10.1016/j.compeleceng.2016.11.021.

Long, S. K. (2019). *Effects of control device and task complexity on performance and task shedding during a robotic arm task* (Master's thesis, Old Dominion University, Norfolk, VA). Retrieved from https://digitalcommons.odu.edu/psychology_etds/231/.

Louw, T., & Merat, N. (2017). Are you in the loop? Using gaze dispersion to understand driver visual attention during vehicle automation. *Transportation Research Part C: Emerging Technologies*, *76*, 35–50. Retrieved from https://doi.org/10.1016/j.trc.2017.01.001.

Lu, C., Shi, J., & Jia, J. (2013). Abnormal event detection at 150 FPS in MATLAB. In *2013 IEEE International Conference on Computer Vision,* 2720–2727. Retrieved from https://doi.org/10.1109/ICCV.2013.338.

Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*. Retrieved from https://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5712

Lyons, J. B., & Havig, P. R. (2014). Transparency in a human-machine context: Approaches for fostering shared awareness/intent. In R. Shumaker & S. Lackey (Eds.), *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, 181–190. New York, NY: Springer International Publishing.

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., … Shively, R. (2017). Shaping trust through transparent design: Theoretical and experimental guidelines. In P. Savage-Knepshield & J. Chen (Eds.), *Advances in Human Factors in Robots and Unmanned Systems,* 127–136. New York, NY: Springer International Publishing.

MacMillan, J., Entin, E. B., & Serfaty, D. (1994). Operator reliance on automated support for target recognition. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *38*(19), 1285–1289. Retrieved from https://doi.org/10.1177/154193129403801908.

Madden, D., Choe, T., Deng, H., Gunda, K., Gupta, H., Ramanathan, N., … Hakeem, A. (2014). Mobile ISR: Intelligent ISR management and exploitation for the expeditionary warfighter. In *2014 IEEE Applied Imagery Pattern Recognition Workshop (AIPR),* 1–11. Retrieved from https://doi.org/10.1109/AIPR.2014.7041918

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of the 11 th Australasian Conference on Information Systems,* 6–8.

Marsh, S., & Dibben, M. R. (2005). Trust, untrust, distrust and mistrust: An exploration of the dark(er) side. In P. Herrmann, V. Issarny, & S. Shiu (Eds.), *Trust Management,* 17–33. Berlin, Germany: Springer.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective, 2nd ed*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

McChesney, I., & Bond, R. (2017). Do computer programmers with dyslexia see things differently? A computational eye tracking study. *Eye Movements in Programming: Spring Academy 2017*, 19–21.

McDermott, P. L., Plott, B. M., Ries, A. J., Touryan, J., Barnes, M., & Schweitzer, K. (2015). *Advanced Video Activity Analytics (AVAA): Human Factors Evaluation* (No. ARL-TR-7286). Army Research Lab Aberdeen Proving Ground, MD: Human Research and Engineering Directorate. Retrieved from http://www.dtic.mil/docs/citations/ADA618115

Menthe, L., Hura, M., & Rhodes, C. (2014). The effectiveness of remotely piloted aircraft in a permissive hunter-killer scenario. Santa Monica, CA: Rand Project Air Force. Retrieved from https://www.rand.org/pubs/research_reports/RR276.html.

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

Mishler, S., Chen, J., Sabic, E., Hu, B., Li, N., & Proctor, R. W. (2017). Description-experience gap: The role of feedback and description in human trust in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*, 313–313. Los Angeles, CA: SAGE Publications.

Molenar, J. V., Malm, W. C., & Johnson, C. E. (1994). Visual air quality simulation techniques. *Atmospheric Environment*, *28*, 1055–1063. Retrieved from https://doi.org/10.1016/1352-2310(94)90265-8.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922. Retrieved from https://doi.org/10.1080/00140139408964957.

Muncaster, J., Collins, G., & Waltman, J. (2015). Real-time automated detection, tracking, classification, and geolocation of dismounts using EO and IR FMV. In *Geospatial Informatics, Fusion, and Motion Video Analytics V, 9473*, 94730N. International Society for Optics and Photonics. Retrieved from https://doi.org/10.1117/12.2180396.

Myhr, S., Ax, G., Gill, J., LeClair, L., Sippel, E., Walters, M., … Dixon, R. E. (2018). Mapping and reconnaissance imager, night-enhanced, for sensing of contaminants, oil, and unseen threats (MARINE SCOUT). In *Ocean Sensing and Monitoring X, 10631*, 106310G. International Society for Optics and Photonics. Retrieved from https://doi.org/10.1117/12.2303963.

Narayanaswami, R., Gandhe, A., & Mehra, R. K. (2010). SUMATRA: Supervised Modeling of ATR Algorithms. In *Proceedings of the IEEE 2010 National Aerospace Electronics Conference,* 99–106. Retrieved from https://doi.org/10.1109/NAECON.2010.5712931.

National Research Council. (1993). *Protecting visibility in national parks and wilderness areas*. National Academies Press.

Nicolaou, A. I., & McKnight, D. H. (2006). Perceived information quality in data exchanges: effects on risk, trust, and intention to use. *Information Systems Research*, *17*(4), 332–351. Retrieved from https://doi.org/10.1287/isre.1060.0103.

Nilsson, M., Laere, J. van, Susi, T., & Ziemke, T. (2012). Information fusion in practice: A distributed cognition perspective on the active role of users. *Information Fusion*, *13*(1), 60–78. Retrieved from https://doi.org/10.1016/j.inffus.2011.01.005.

Nomura, T. (2016). Robots and gender. *Gender and the Genome, 1*(1). Retrieved from https://doi.org/ 10.1089/gg.2016.29002.nom.

Olson, S. C., Gaudiosi, D., Beard, A., & Gueler, R. (2018). Statistical evaluation of motion-based MTF for full-motion video using the Python-based PyBSM image quality analysis toolbox. In *Long-Range Imaging III, 10650*, 106500L. International Society for Optics and Photonics. Retrieved from https://doi.org/10.1117/12.2305406.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, *56*(3), 476–488. Retrieved from https://doi.org/10.1177/0018720813501549.

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. Retrieved from https://doi.org/10.1109/3468.844354.

Parasuraman, Raja, & Hancock, P. A. (2001). Adaptive control of mental workload. In *Stress, workload, and fatigue,* 305–320. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Parasuraman, Raja, & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors*, *52*(3), 381–410. Retrieved from https://doi.org/10.1177/0018720810376055.

Parasuraman, Raja, Molloy, R., & Singh, I. (1993). Performance consequences of automation induced complacency. *International Journal of Aviation Psychology*, *3*. Retrieved from https://doi.org/10.1207/s15327108ijap0301_1.

Parasuraman, Raja, Mouloua, M., & Molloy, R. (1997). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors*, *38*, 665–679. Retrieved from https://doi.org/10.1518/001872096778827279.

Parasuraman, Raja, & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. Retrieved from https://doi.org/10.1518/001872097778543886.

Parker, S. (2015). Real-time processing of dual band HD video for maintaining operational effectiveness in degraded visual environments. Retrieved from https://doi.org/10.1117/12.2176958.

Patrick, J. B. (2015). *Pulling Teeth: Why Humans Are More Important Than Hardware in Airborne Intelligence, Surveillance, and Reconnaissance*. Norfolk, VA: National Defense University, Joint Forces Staff College. Retrieved from http://www.dtic.mil/docs/citations/ADA623948

Plott, B. M., McDermott, P. L., & Barnes, M. (2017). *Advanced Video Activity Analytics (AVAA): Human Performance Model Report* (No. ARL-TR-8255). Aberdeen Proving Ground, MD: US Army Research Laboratory. Retrieved from http://www.dtic.mil/docs/citations/AD1043943

Poostchi, M., Palaniappan, K., & Seetharaman, G. (2017). Spatial pyramid context-aware moving vehicle detection and tracking in urban aerial imagery. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS),* 1–6. Retrieved from https://doi.org/10.1109/AVSS.2017.8078504.

Ram, K., Singh, S., Sarin, M. M., Srivastava, A. K., & Tripathi, S. N. (2016). Variability in aerosol optical properties over an urban site, Kanpur, in the Indo-Gangetic Plain: A case study of haze and dust events. *Atmospheric Research*, *174–175*, 52–61. Retrieved from https://doi.org/10.1016/j.atmosres.2016.01.014.

Reddy, V., Sanderson, C., & Lovell, B. C. (2011). Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In *CVPR 2011 Workshops.* 55–61. Retrieved from https://doi.org/10.1109/CVPRW.2011.5981799.

Roshtkhari, M. J., & Levine, M. D. (2013). An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer Vision and Image Understanding*, *117*(10), 1436–1452. Retrieved from https://doi.org/10.1016/j.cviu.2013.06.007.

Ross, J., & Coman, C. (2014). Full Motion Video in a coalition environment. In *2014 10th International Conference on Communications (COMM),* 1–4. Retrieved from

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*(1), 76–87. Retrieved from https://doi.org/10.1518/001872007779598082.

Ruff, H. A., Calhoun, G. L., Draper, M. H., Fontejon, J. V., & Guilfoos, B. J. (2004). *Exploring Automation Issues in Supervisory Control of Multiple UAVs*. Dayton, OH: Sytronics Inc. Retrieved from http://www.dtic.mil/docs/citations/ADA436181

Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Z., & Klette, R. (2018). Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer*

Vision and Image Understanding, *172*, 88–97. Retrieved from https://doi.org/10.1016/j.cviu.2018.02.006.

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*, *58*(3), 377–400. Retrieved from https://doi.org/10.1177/0018720816634228.

Schweitzer, K. M., Ries, A. J., McDermott, P. L., Plott, B. M., Wilson, E. A., & Morrow, G. P. (2018). *Human Factors Evaluation of Advanced Video Activity Analytics (AVAA) Functionality* (No. ARL-TR-8301). San Antonio, TX: US Army Research Laboratory. Retrieved from http://www.dtic.mil/docs/citations/AD1048161.

Senders, J. W. (1983). *Visual sampling processes*. Tilburg, Netherlands: Tilburg University. Retrieved from https://research.tilburguniversity.edu/en/publications/6236d393-ab34-4010-84c6-862edc525d14.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators*. Cambridge, MA: Massachusetts Inst. of Tech., Man-Machine Systems Lab. Retrieved from http://www.dtic.mil/docs/citations/ADA057655

Sohn, Y. W., & Doane, S. M. (2004). Memory processes of flight situation awareness: interactive roles of working memory capacity, long-term working memory, and expertise. *Human Factors*, *46*(3), 461–475. Retrieved from https://doi.org/10.1518/hfes.46.3.461.50392.

Thissell, W. R., Czajkowski, R., Schrenk, F., Selway, T., Ries, A. J., Patel, S., … Palaniappan, K. (2015). A scalable architecture for operational FMV exploitation. In *Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 1062–1070. Washington, DC: IEEE Computer Society. Retrieved from https://doi.org/10.1109/ICCVW.2015.139.

Tseng, C.-H., & Shao, Y. (2012). Sample size growth with an increasing number of comparisons. *Journal of Probability and Statistics*. Retrieved from https://doi.org/10.1155/2012/935621.

Wang, C., Zhang, C., & Yang, X. J. (2018). Automation reliability and trust: A Bayesian inference approach. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 62*, 202–206. Los Angeles, CA: SAGE Publications.

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. Retrieved from https://doi.org/10.1080/14639220500370105.

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. Retrieved from https://doi.org/10.1518/hfes.45.3.360.27250.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering Psychology & Human Performance*. London, England: Psychology Press.

Wickens, C. D., McCarley, J. S., Alexander, A. L., Thomas, L. C., Ambinder, M., & Zheng, S. (2007). Attention-Situation Awareness (A-SA) model of pilot error. In D. C. Foyle & B. L. Hooey (Eds.), *Human Performance Modeling in Aviation*. Boca Raton, FL: CRC Press.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for Mental workload measurement in a test and evaluation environment. *Human Factors*, *35*(2), 263–281. Retrieved from https://doi.org/10.1177/001872089303500205.

Wright, J. L., Chen, J. Y. C., Barnes, M. J., & Hancock, P. A. (2017). The effect of agent reasoning transparency on complacent behavior: An analysis of eye movements and response performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1594–1598. Retrieved from https://doi.org/10.1177/1541931213601762.

Xiao, T., Zhang, C., & Zha, H. (2015). Learning to detect anomalies in surveillance video. *IEEE Signal Processing Letters*, *22*(9), 1477–1481. Retrieved from https://doi.org/10.1109/LSP.2015.2410031.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 408–416. New York, NY: Association for Computational Machinery. Retrieved from https://doi.org/10.1145/2909824.3020230.

Yarovinskiy, A. (2017). *Intelligence Surveillance and Reconnaissance Full Motion Video Automatic Anomaly Detection Of Crowd Movements: System Requirements For Airborne Application*. Maxwell AFB, AL: Air Command and Staff College, Air University. Retrieved from http://www.dtic.mil/docs/citations/AD1054570

Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, *43*(3), 355–365. Retrieved from https://doi.org/10.1518/001872001775898269.

Zhang, X., Li, H., Qi, Y., Leow, W. K., & Ng, T. K. (2006). Rain removal in video by combining temporal and chromatic properties. In *2006 IEEE International Conference on Multimedia and Expo,* 461–464. Retrieved from https://doi.org/10.1109/ICME.2006.262572.

APPENDIX A
INFORMED CONSENT DOCUMENT
OLD DOMINION UNIVERSITY

PROJECT TITLE: Effects of Transparency and Degraded Visual Environments on Trust and Performance during a Full Motion Video Analysis Task

INTRODUCTION

The purpose of this form is to give you information that may affect your decision whether to say YES or NO to participation in this research, and to record the consent of those who say YES.

RESEARCHERS

James P. Bliss, Ph.D., Old Dominion University, Psychology Department, Responsible Project Investigator
Sarah C. Leibner, B.S., Old Dominion University, Psychology Department, Graduate Student

DESCRIPTION OF RESEARCH STUDY

In this experiment, you will complete a Background Information Form. Following this, you will be asked to perform a practice session to familiarize yourself with the information system. After training, you will be asked to evaluate intelligence information from full motion video, deciding whether and how to decide about the information provided. You will have the option to offload decision making tasks to an automated system. Periodically during this session, you will complete questionnaires evaluating your trust. Additionally, an eye tracking system will be used to evaluate where you are looking throughout the experiment. After the experimental sessions, you will complete an Opinion Questionnaire to indicate your strategy for responding. You will then be debriefed and dismissed.

You will receive 1 SONA credit for participating in this study.

If you say YES, then your participation will last for approximately 35 minutes in MGB 326. Approximately 46 subjects will be participating in this study.

EXCLUSIONARY CRITERIA

To be eligible for this study, you must be at least 18 years of age or older and must not have participated in the study "Investigation of Alternative Real-Time Measures of Human-Automation Trust".

RISKS AND BENEFITS

RISKS: If you decide to participate in this study, then you may face a risk of eyestrain from using a computer, but it will be no more than from playing a video game. The researcher tried to reduce these risks by minimizing the amount of time in the study to sixty minutes. As with any research, there is some possibility that you may be subject to risks that have not yet been identified.

BENEFITS: There are no known benefits from this study. Others may benefit by experiencing higher-quality technology systems.

COSTS AND PAYMENTS

The main payment to you for participating in this study is the extra credit or course credit points that you will earn for your class. Although researchers are unable to give you payment for participating in this study, if you decide to participate in this study, you will receive 1.5 Psychology Department research credit, which may be applied to course requirements or extra credit in your Psychology course. Equivalent credits may be obtained in other ways. You do not have to participate in this study, or any Psychology Department study, in order to obtain this credit.

NEW INFORMATION

If the researchers find new information during this study that would reasonably change your decision about participating, then they will give it to you.

CONFIDENTIALITY

The researchers will take reasonable steps to keep your private information, such as questionnaires, confidential. The researchers will store the information in a locked filing cabinet for five years, after which the data will be destroyed. The results of this study may be used in reports, presentations, and publications, but the researcher will not identify you.

WITHDRAWAL PRIVILEGE

It is OK for you to say NO. Even if you say YES now, you are free to say NO later, and walk away or withdraw from the study -- at any time. Your decision will not affect your relationship with Old Dominion University, or otherwise cause a loss of benefits to which you might otherwise be entitled. The researchers reserve the right to withdraw your participation in this study, at any time, if they observe potential problems with your continued participation.

COMPENSATION FOR ILLNESS AND INJURY:

If you agree to participate, then your consent in this document does not waive any of your legal rights. However, in the event of harm, injury, or illness arising from this study, neither Old Dominion University nor the researchers can give you any money, insurance coverage, free medical care, or any other compensation for such injury. In the event that you suffer injury as a result of participation in any research project, you may contact Dr. James P. Bliss at 757-683-4051, Dr. Tancy Vandecar-Burdin (ODU IRB Chair) at 757-683-3802, or the ODU Office of Research, 757-683-3460.

VOLUNTARY CONSENT

By signing this form, you are saying several things. You are saying that you have read this form or have had it read to you, that you are satisfied that you understand this form, the research study, and its risks and benefits. The researchers should have answered any questions you may have had about the research. If you have any questions later on, please contact the researcher at the number above.

If at any time you feel pressured to participate, or if you have any questions about your rights or this form, then you should call Dr. Tancy Vandecar-Burdin, the current IRB chair, at 757-683-3802, or the Old Dominion University Office of Research, at 757-683-3460.

And importantly, by signing below, you are telling the researcher YES, that you agree to participate in this study. The researcher should give you a copy of this form for your records.

Participant's Printed Name & Signature	**Date**
Investigator's Printed Name & Signature	**Date**

APPENDIX B
DEMOGRAPHICS QUESTIONNAIRE

Participant #_____ Date:__________ Time:__________

The purpose of this questionnaire is to collect background information for participants in this experiment. This information will be used strictly for this experiment and for research purposes only. Please complete or circle each item to the best of your knowledge.

1. Age _________

2. Male
   Female

3. Are you a current or former member of the Armed Services? ________

   0=No
   1=Yes

4. Have you ever been diagnosed as color blind or color deficient? ________

   0 = No
   1 = Yes

6. Have you ever been diagnosed as being nearsighted (myopic)? _______

   0=No
   1=Yes

7. Have you ever been diagnosed as being farsighted (hyperopic)?_______

   0=No
   1=Yes

8. If you answered yes to either #6 or #7, do you have correction with you (i.e. glasses, contact lenses, etc.)? _______

   0=No
   1=Yes

9. How many hours per week do you play video/simulation games? _______

10. How many hours per week do you use a computer (work and recreation combined)?______

APPENDIX C
ONLINE TRUST QUESTIONNAIRE

Human-Computer Trust Scale - Madsen and Gregor (2000)

Part. #: _______ Group: ________ Session: ________ Date:______ Time:_______

Below is a list of statements for evaluating trust between people and automated systems. Please circle the number that best describes your feeling or your impression of the automated video analysis aid you used during the task.

1. I can rely on the system to function properly.

Not descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 : Very Descriptive

2. The system has sound knowledge about the key identification features of IED emplacers.

Not descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 : Very Descriptive

3. Although I may not know exactly how the system works, I know how to use it to perform well.

Not descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 : Very Descriptive

4. Even if I have no reason to expect the system will be able to identify IED emplacement activity, I still feel certain that it will.

Not descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 : Very Descriptive

5. I feel a sense of attachment to using the system.

Not descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 : Very Descriptive

6. I am confident in the decisions that I made.

Not descriptive: 1 2 3 4 5 6 7 8 9 10 11 12 : Very Descriptive

APPENDIX D
OFFLINE EXPERIENCE QUESTIONNAIRE

Participant No.__________ Date:____________

Please answer the following questions about yourself by circling the most appropriate response. The information you provide will be kept completely confidential and will not be linked backed to you in any way.

Please circle only one answer per question.

1. This experiment was time consuming.

Disagree strongly Disagree Neutral Agree Agree Strongly

2. This experiment was confusing.

Disagree strongly Disagree Neutral Agree Agree Strongly

3. I did not feel like I had a good grasp on the instructions for this experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

4. I feel like I performed well on this experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

5. I feel like I performed poorly on this experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

6. This experiment was easy to understand

Disagree strongly Disagree Neutral Agree Agree Strongly

7. This experiment was enjoyable.

Disagree strongly Disagree Neutral Agree Agree Strongly

8. I did not enjoy this experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

9. I am glad that I participated in this experiment

Disagree strongly Disagree Neutral Agree Agree Strongly

10. I felt engaged in the tasks for this experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

11. I felt like I received adequate time to train and get comfortable with the experimental task before beginning the actual experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

12. I felt like I did not receive adequate time to train and get comfortable with the experimental task before beginning the actual experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

13. I felt motivated to perform to the best of my ability in this experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

14. I did not care how well I performed in this experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

15. I tried my best to perform well on this experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

16. I did not try my best to perform well on this experiment.

Disagree strongly Disagree Neutral Agree Agree Strongly

17. Overall, I would recommend this experiment to other students.

Disagree strongly Disagree Neutral Agree Agree Strongly

18. Did you have a strategy for responding to the experimental task?

Yes No

If yes, please describe

____

____

____

19. Do you have any other thought, feelings, or comments about the experiment?

____

____

____

APPENDIX E
PARTICIPANT INSTRUCTION VIGNETTE

You have received intelligence reporting indicating heightened bomb-making activity in your area of responsibility. Your task is to analyze Full Motion Video feed, looking for potential Improvised Explosive Device (IED) emplacement activity in preparation for a U.S. military presence patrol in the area. IED emplacement activity consists of small groups stopping in or near a road, digging and/or unloading objects on the road, and then hastily departing. You will analyze short FMV clips with the aid of an automated analyst designed to identify probable IED emplacements along convoy routes. The automated analyst is able to detect objects such as people and vehicles 95% of the time and can correctly classify IED emplacement activity 78% of the time. The automated analyst will indicate whether it has identified IED emplacement activity. Your task is to choose whether to agree or disagree that there is IED emplacement activity present. You may also choose to request additional information or to delegate decision-making to the automated system. Keep in mind that the automated system is 78% reliable. You must ensure that no IED emplacements are missed. Missed IED emplacements could result in friendly and civilian casualties.

While you are analyzing the FMV feed, you must also monitor a U.S. military convoy. The convoy feed will be on the right side of the screen. **You will have 10 seconds to analyze the activity. If you do not choose within this time, you will be presented with the next video.**

You will have an opportunity to familiarize yourself with this task. When you are comfortable with the task, let the researcher know that you are ready to proceed.

CURRICULUM VITAE

Sarah C. Leibner

Department of Psychology
Mills Godwin Building, Room 250
Old Dominion University
Norfolk, VA 23529-0267

EDUCATION:

2009-2012	A.S., Intelligence Operations, Cochise College, Sierra Vista, AZ
2011-2013	A.S., Science, Tidewater Community College, Virginia Beach, VA
2016-2017	B.S., Psychology, Old Dominion University, Norfolk, VA
2017 – Present	Ph.D., Human Factors Psychology, Old Dominion University, Norfolk, VA

RESEARCH EXPERIENCE:

11/2016 – 03/2019: Project Augmented Reality User Interfaces for Tactical Drones, Old Dominion University
Supervisor: James P. Bliss, Ph.D.; Peter Crane, Ph.D. (PI)
Company: Virtual Reality Rehab (VRR) develops emerging technologies to improve military effectiveness.

Purpose was to determine the optimal commercial off-the-shelf drone platform for testing and implementing augmented reality interfaces for drone pilots. Collaborated on use case scenario, provided Subject Matter Expert (SME) consultation, and created a concept of operations for field use of miniaturized drones. Conducted software build testing, usability testing while manipulating user interface and environmental variables, and analyzed data. Assisted with planning research. Two experimental studies were conducted for this project.

05/2018 – 11/2018: Project Synthetic Vision for Ground Forces, Old Dominion University
Supervisor: James P. Bliss, Ph.D.; Peter Crane, Ph.D. (PI)
Company: Virtual Reality Rehab (VRR) develops emerging technologies to improve military effectiveness.

Purpose was to develop, test, and study the effects on user performance of an augmented reality system for Joint Tactical Air Controllers using a head mounted display. Conducted literature reviews; led, contributed to, and organized SME team consultation efforts; assisted in designing prototypes and streamlining user interfaces. Conducted software build testing. Assisted with writing quarterly and final reports.

11/2016 - 08/2017: Air Force Research Laboratory Project Trust in Automation of Data Fusion, ODU
Supervisor: James P. Bliss, Ph.D. (PI)

Purpose was to investigate alternative measures of trust in automation using an intelligence fusion task. Conducted literature review, assisted experimental design development, completed data analysis, and collaborated in preparing the final research paper. One experimental study was conducted for this project.

PUBLICATIONS:

Leibner, S., Proaps, A., & Bliss, J. P. (2019). Effects of transparency and controller type on performance in a multi-layer display. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 63(1), 1772-1776.

Proaps, A., Leibner, S., & Bliss, J. P. (2019). Effects of transparency level, controller type, and degraded conditions on performance using augmented reality and synthetic vision. In *Winning the War of Cognition by Pushing Readiness and Lethality Boundaries*. Paper presented at I/ITSEC: Interservice/Industry Training, Simulation & Education Conference, Orlando, FL.

Tiller, L., Angelini, C., Leibner, S. C. & Still, J. (2019). Explore-a-Nation: Combining graphical and alphanumeric authentication. In A. Moallem (Ed.), *HCI for Cybersecurity, Privacy and Trust*. Paper presented at HCII: International Conference on Human-Computer Interaction, Orlando, FL (81-95). Cham, Switzerland: Springer.

HONORS, AWARDS AND PRIZES:

2019 Most Practical Student Research Award, HFES