2020

Eye-tracking to Evaluate Trust in Human-ATR Interaction

Samuel Francis Adelman
Wright State University

EYE-TRACKING TO EVALUATE TRUST IN HUMAN-ATR INTERACTION

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Industrial and Human Factors Engineering

By

SAMUEL FRANCIS ADELMAN

B.S., Wright State University, 2017

Wright State University

2020

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

April 24, 2020

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Samuel Francis Adelman ENTITLED Eye-tracking to Evaluate Trust in Human-ATR Interaction. BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Industrial and Human Factors Engineering

_____
Mary Fendley, Ph.D., Thesis Director

_____
John Gallagher, Ph.D., B.I.E. Interim
Department Chair

Committee on Final Examination

_____
Mary Fendley, Ph.D.

_____
Subhashini Ganapathy, Ph.D.

_____
Josh Ash, Ph.D.

_____
Barry Milligan, Ph.D., Interim Dean of the
Graduate School

ABSTRACT

Adelman, Samuel Francis. M.S.I.H.E., Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, 2020. EYE-TRACKING TO EVALUATE TRUST IN HUMAN-ATR INTERACTION

Human collaboration with targeting aids have allowed analysts to achieve a greater level of coordination and productivity in a variety of fields. This project investigates the impact that an Assisted Target Recognition (ATR) algorithm's false alarm rate and the task Target of Interest (TOI) level has on user-system trust and use in a targeting decision task. Previous studies suggest that an increased number of false alarms in an ATR task negatively impacts analyst trust in the system. This study will further contribute to this research, aiming to provide a better framework for appropriate tolerance levels within ATR algorithms, utilizing pre-truthed ATR footage. Two studies, a pilot and a main study, were conducted. Participants performed computer simulated search tasks with or without the help of the detection aid at four false alarm rates. Trust and use in the decision aid were recorded by participant gaze behavior and a trust in automation scale.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This research was made possible by the National Science Foundation under the sponsorship of I/UCRC Center for Surveillance Research. This research was also supported with contributions from Etegent Inc. We thank Adam Nolan and Jameson Morgan for their patronage, support and advice.

Introduction

Eye-tracking (ET), though routinely utilized in fields such as marketing research (e.g., marketing stimuli response, visual brand attention; (Wedel, 2017; Khushaba, 2013; Chandon, 2006), and psychology (e.g., infancy behavior, personality disorders: Mele, 2012; Gredeback, 2009; Iocono, 1982), is still a relatively new approach in assessing trust in the field of automation, though has become increasingly prevalent, (O'Meara et al., 2015; Van de Merwe, Van Dijk, H., & Zon, R., 2012; Ratwani, R. M. & McCurry, J, 2010). Trust in automation has been widely studied throughout the engineering domain and its relation to user compliance and reliance. The failure of an automated system can be seen to impact a limit or threshold of trust a user has in that system when presented with a type of error. For instance, Dixon (2006) states that once a user's ability to perform a task exceeds the capabilities of the automation system, they will begin to disregard it. Understandably, this is the most obvious consequence, but should not be overlooked. Errors from an aid have and will continue to mislead a user into misuse or disuse of a system, (Parasuraman, 2000). Being psychological in nature, trust in automation is typically measured by various self-report surveys or questionnaires (e.g. Jian, et al., 2000; Chancey et al., 2017), despite their well-known limitations revolving around various response-biases (Miller, 2011; Ezzati, 2006). With this understanding, eye-tracking offers researchers the unique ability to monitor the frequency and/or pattern of a users' gaze during a task outside of subjective means, highlighting information recognition and processing as it may relate to the cognitive workload or situational awareness of the human-operator.

The investigation into the utility of monitoring a user's gaze behavior as an indication of other factors such as situational awareness has yielded promising results. For example, Moore (2010) and Hauland (2008) both showed how frequently scanned aircraft Area of Interest (AOI's) locations during Air Traffic Controller simulations significantly predicted high situational awareness scores and led to fewer recall errors. This is an important finding that is in line with similar studies showing that active visual scanning and/or attention allocation led to increased detection rates and subjective situational awareness scores, (Wickens, 2004; Ratwani, 2010). Specifically, regarding trust in automation, Korber (2018), Hergeth (2016) and Walker (2018) effectively demonstrated that higher subjective automation trust ratings positively correlated with increased non-driving related task attention and reduced automation monitoring during automated roadway simulations, indicating that an increase in user trust in the automated driving systems actually allowed participants to deviate their viewing from the road and complete secondary tasks. These findings prompt the use of gaze behavior as a potentially reliable measure for trust in automation apart of subjective means.

The construct of gaze behavior and eye monitoring involves measures such as fixation duration and count, as well as saccade movements (**Table 1**). A fixation has been defined as a foveal-directed visual focus towards a stimulus lasting up to 200 milliseconds, whilst a saccade is defined as the eye-movement between fixation points, (Greef, 2009; Meißner, 2019). As mentioned before, these measurements have proven beneficial in the fields of marketing research and psychology as they provide researchers a glimpse into the cognitive processes of the individual. Additional measurements such as pupillary response have also been explored,

demonstrating the relationship between the magnitude of pupillary dilation and a subject's mental processing load, (Iqbal, 2004; Beatty, 1982).

Table 1: Definitions of Study Terms

Fixation Duration	Time (In seconds) that a participant looked at an AOI
Fixation Count	Number of instances that a participant fixated on an AOI
Transitions (Saccade)	An eye movement between two AOI's
False Alarm	An incorrect observation by a user or system that a signal is present when in fact it is absent
Miss	A failure to detect a signal when it is present
Target of Interest (TOI)	Vehicle(s) of color, model, or direction that participant was tasked with tracking
Area of Interest (AOI)	Designated boundaries that allows the eye tracking researcher or analyst to calculate quantitative eye movement measures

This thesis contains the research, procedures, methodologies, and results of two IRB approved studies. The first experiment, termed as the "Waldo" study, enabled us to gain a better familiarity with Eye Tracking software and its measures, and provided us with an appropriate framework for the main "ATR" experiment.

Literature Review

Assisted Target Recognition

The broadening of automation has prompted the design and implementation of computerized decision aiding software in a variety of fields. Designed to assist users in various applications, benefits can be seen especially in increasingly complex and critical tasks, (Jian et al., 2000; Kaber, 1997) and in respect to various forms of decision making, (Morrisson, 1998; Parasuraman, 2008). Consequently, this new focus on human capabilities alongside automation revealed unexpected changes in human performance resulting in a manifestation of various cognitive demands, stemming from managing the interface itself either during setup, operation or during performance analysis, (Parasuraman, 2008; Woods, 1996).

In the military domain, Assisted Target Recognition is a decision aid technology employed often to analyze large amounts of geospatial intelligence and assist analysts by providing useful information when searching imagery, (Irvine, 2008). The potential benefits of this technology are significant, particularly regarding threat-detection and identification capabilities of military vehicles or installments – an area of increasing interest. ATR technology to date uses synthetic aperture radar (SAR) imagery that often is slow and inaccurate, (Clemente, 2017; Wang, 2017), but has demonstrated superior usability in detection of marine ships, SCUD missile launchers, and the airborne detection of mines, augmenting the role of what typically would be filled by soldiers on the ground, (Zhao, 2018; Jones, 1999; Rajagopal, 2005).

Evaluation of an ATR technology's performance typically results from a comparison to a previously truthed imagery dataset where its detection probability and false alarm rate can be manipulated and determined, (Irvine, 2008). Detection and False Alarms are directly related, and when graphed on a Receiver Operating Characteristic (ROC) curve, allows a systems designer or analyst to select the appropriate threshold based on the environment or the task at hand, (Dougherty, 2005). In the minefield example previously mentioned, analysts could choose to encounter more false alarms to maximize their detection, as the alternative could prove dangerous. The current study aims at determining the impact of this threshold change on a user's trust in the system.

Compliance and Reliance

The degree of automation trust is a key factor of user reliance and compliance. If the users' ability to perform a task begins to outweigh the capability of automation, they will likely begin to disregard it, (Dixon, 2006). Interestingly, people seem to place a substantial amount of trust in automation performance, and often rate automated systems as more trustworthy than human alternatives (Lyons, 2012; Dijkstra, 1999) to such a degree that initial errors hastily degrade user trust, (Dzindolet, 2003; Merrit, 2015). Understandably, some degree of failure is anticipated in these systems, but the larger question remains as to the location of the psychological threshold that an operator possesses and considers when determining whether the automation should be trusted or used. This threshold influences the level of compliance or reliance in a system.

Dixon (2006) refers to reliance as the response of an operator when no alarm is present. Operators who rely on a system in turn can sub-divide cognitive resources in simultaneous or concurrent tasks, confident that the system will alert them when attention is needed. However, if the system begins to miss these alerts, operator reliance of the system rapidly decreases, (Chancey, 2017). Conversely, compliance refers to the response of an operator when there is an alarm present. Compliant operators will cease attending to separate tasks when alerted to a fault or warning and proceed accordingly. Thus, if the system often incorrectly alerts (false alarm) the operator, compliance is negatively affected, leading to an increase in response time, Rice (2011), or ignoring signals in cognitively taxing conditions, Bliss (1998).

As to the degree that Miss or False Alarm prone systems negatively affect trust in automation, findings vary. Both Dixon (2006) and Rice (2011) found false-alarm prone systems to have a more negative impact on task performance than miss-prone systems, whilst Chancey (2017) found task performance worse in miss-prone systems, so much so that it was postulated that false-alarms caused operators to pay more attention to the task. Overall, both types of mistakes have a negative impact on automation trust, and it is a generally accepted principle that the level of trust in automation is positively correlated with its utilization, (Geels-Blair, 2013).

Measures of Trust

Trust has been traditionally difficult to measure due to its multidimensional nature, (Jian et al., 2000). Similar constructs without concrete distinctions, such as risk, can prevent a clearer understanding of what precisely trust is and its relationship within people or machines, (Mayer,

Davis, & Schoorman, 1995), and prompts the question as to whether it is to be viewed or measured in a static or comprehensive way, (Rousseau et al., 1998). Cognitive processes often take precedent in determining trust in automation, in that much of the concern lies in whether an operator believes that the automation does what it was expected to do, (Chien at al., 2014), rather than the affectual factors seen in interpersonal trust research, (McKnight, Choudhury, & Kacmar, 2002) in psychological domains.

Despite a rather large amount of trust research surrounding fields such as security inspection (Kraemer, Carayon, & Sanquist, 2009), or command and control instances (Rovira, McGarry & Parasuraman, 2007), many of these scales were not empirically founded for measuring trust in automation. In response, (Jian et al., 2000) devised a multi-item scale for operators, examining the similarities and differences between general and human-machine trust. Later validated by (Safar & Turner, 2005) and (Spain, Ernesto, & Bliss, 2008), their scale now offers researchers an empirically driven measure for trust in automation.

Preface: Waldo Study

The Waldo pilot study was conducted to serve as a proving period to shed light on and fine tune the framework for the ATR experiment. It was important to understand how the length and complexity of certain search tasks could induce fatigue in the participants, and to fine tune the quantitative response questions in such a way that aided with data analysis to effectively detect and thus illustrate the change in trust and use of a decision aid.

Methods

Participants were 23 (11 men and 12 women) undergraduate and graduate students at a University in Southwest Ohio, U.S., who completed a visual search task and subsequent trust surveys with no compensation. Participants' were between the ages of 18 and 45 with a mean age of 25.09 years (SD = 6.39) and described their nationality as American (47.8%), Indian (30.4%), British (4.3%), French (4.3%), Nigerian (4.3%), Mauritanian (4.3%), and Iraqi (4.3%). Participation was on a voluntary basis, and due to the visual nature of this study, we restricted eligibility to individuals with actively corrected vision and no visual or motor dysfunction. No participants failed to meet these requirements.

Apparatus and Stimuli

This experiment utilized 20 different images sampled from various "Where's Waldo" puzzle search books with or without the presence of a decision support system (DSS). The DSS-aid conditions **(Figure 1)** display an enlarged 1.5 x 1.5cm 'chip' from the original 26.5 x 19cm parent image, equating to roughly 1/200th of the task image. The correct aid's chip displays an

area within 1 cm to the location of the target, whilst the incorrect decision aid displays an area 10 – 25cm apart from the target.

Figure 1: DSS-aid Condition

Participants were seated in a sound and light controlled room and completed the visual search tasks using a Tobii T120 eye tracker. The T120 has a tracking distance between 50 to 80 centimeters and services gaze angles of 35 degrees in either direction, therefore allowing minimal head movement of the participants. The data rate of the T120 ranges from 60 Hz to 120 Hz, with a screen resolution of 1280 x 1024 pixels on a 17-inch display. Subjects' eye movements and fixation times were measured by heat maps and directional analytics. Data output includes timestamps, fixation duration and frequency, and saccade frequency.

Design and Procedure

The experiment had a 2 x 4 within-subjects design, (**Table 2**). The experimental blocks were produced by the combination of the presence of the decision aid (Yes DSS, no DSS), and

the false-alarm rate of the DSS (0-3 False Alarms). Each participant in the DSS condition completed 5 puzzles under 4 conditions for a total of 20 trials, with a different puzzle each trial. Each puzzle was limited to 90 seconds. Subjects progressed linearly through the 100%, 80%,

Table 2: Experimental Conditions

	Experimental Conditions				
	Decision Aid				No Decision Aid
	(0 FA – 100% identification)	(1 FA – 80% identification)	(2 FA – 60% identification)	(3 FA – 40% identification)	
Image (1-5)					

60% and 40% DSS detection rates. Non-DSS condition subjects completed the same 20 puzzles without the aid. In both the DSS and non-DSS groups, puzzles were in no specific order and would be viewed just once.

(**Figure 2**) shows after completing a consent form and a pre-questionnaire to assess adequate visual function and basic demographics, participants were given a walkthrough of the search task with an example. Participants were tasked with searching for a single unique character known as “Waldo,” which has been already been randomly placed within the scene. The goal is to locate “Waldo,” a uniquely dressed character, as quickly as possible in the surrounding environment.

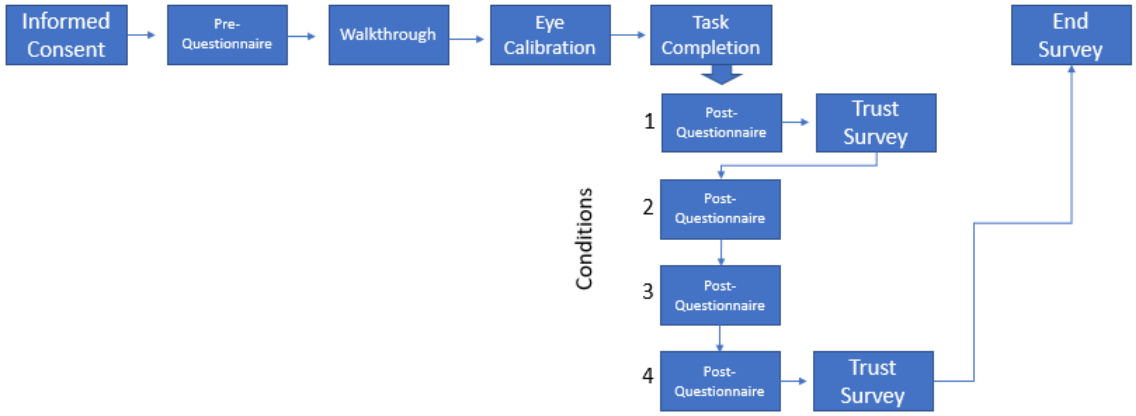


Figure 2: Procedure

Following eye-tracker calibration, participants were told they were free to use or not use the DSS if present and were not informed as to the nature of its accuracy. Upon finding 'Waldo,' participants were asked to press a key on a nearby keyboard to advance to the next puzzle. A post-questionnaire was administered after each condition that consisted of four questions that pertained to their trust, use, and comfort of the DSS-aid. Responses are coded on a scale from 1 to 5 ranging from 'Strongly disagree' to 'Strongly agree,' (**Table 3**). Separately, after the first and last condition, a measure of trust in automation was obtained using the Checklist for Trust between People and Automation (Jian, Bisantz, & Drury, 2000).

Table 3: Post Questionnaire

1. I became more comfortable with the decision aid as I continued to use it.
2. My identification performance improved with the help of the decision aid.
3. I have trust in the decision aid thus far.
4. I used the decision aid at this condition.

This 12-item measure asks participants to assess their level of confidence and reliability in the system. Responses are coded on scale from 1 to 7 ranging from 'Not at all' to 'Extremely.' **Table**

4 provided an example. Finally, a qualitative free-report survey was given at the end of all conditions that asked participants how they thought the DSS impacted their decision making, and at what point during the experiment they began to lose trust in the DSS – if at all. Participant gaze behavior was monitored simultaneously to examine use of the DSS.

Table 4: Trust between People and Automation Survey

1. The system is misleading
2. I suspect the system is incorrect
3. The system is accurate
4. The system has gained my trust

Results

(**Table 5**) presents the descriptive statistics for fixation duration and saccade instances with the DSS, as well as the post-questionnaire responses across all four conditions and trust between people and automation results after condition one and four. Contrary to (Wang, Jamieson, & Hollands, 2009), we divided the Trust between People and Automation survey into two groups for analysis. The first group, which consisted of questions one through five, were labeled 'negative connotated questions,' as they centered around deception and suspicions of the aid. Conversely, the second group consisted of questions six through eleven which were labeled 'positive connotated questions,' as they centered around the integrity and dependability of the aid.

Table 5: Descriptive Statistics

			Condition 1	Condition 2	Condition 3	Condition 4
Gaze Behavior	Fixation	*M*	36.79	43.85	28.7	31.42
		SD	5.46	17.84	5.73	9.62
	Saccade	*M*	49.4	58.4	44.6	51
		SD	6.31	20.95	4.72	16.57
Post-Questionnaire	Question 1	*M*	3.55	3.8	3.6	2.95
		SD	0.95	1.11	1.23	1.47
	Question 2	*M*	3.6	3.95	3.55	2.9
		SD	1.05	1.15	1.28	1.41
	Question 3	*M*	2.9	3.35	3.15	2.65
		SD	1.02	1.23	1.14	1.18
	Question 4	*M*	3.65	3.75	3.6	3.4
		SD	1.14	0.97	1.35	1.43
Bisantz	Negative	*M*	3.03			3.85
		SD	0.95			1.04
	Positive	*M*	4.02			3.45
		SD	0.97			0.8

Eye fixation times on the decision aid had the greatest difference between conditions two (M = 43.85, SD = 17.84) and three (M = 28.7, SD = 17.84), respectively. However, their difference was not significant, $t(4) = 1.79$, $p = .07$. Similarly, saccade instances between conditions two (M = 58.4, SD = 20.95) and three (M = 44.6, SD = 4.72) had the greatest difference, yet failed to reach significance, $t(4) = 1.35$, $p = .12$.

(**Table 6**) displays the post questionnaire significance results. Between the first and last conditions, question one (DSS comfort level) with (M = 3.55, SD = .95) and (M = 2.95, SD = 1.47) was found to be significant, $t(20) = 1.75$, $p = .048$. Additionally, question two (Identification performance) with (M = 3.6, SD = 1.05) and (M = 2.9, SD = 1.41) was also found to be significant,

t(2.33), *p* = .015. Question three (Trust) and question 4 (Usage) failed to achieve significance between the first and last conditions.

Table 6: Post-Questionnaire Significance Table

	t=	*p*=
Q1	1.75	0.048*
Q2	2.33	0.015*
Q3	0.72	0.24
Q4	0.93	0.18

The negatively connotated group within the trust in automation survey, with (*M* = 3.03, *SD* = .95) and (*M* = 3.85, *SD* = 1.04) achieved significance, *t*(19) = -3.41, *p* = .001, whilst the positively connotated group with (*M* = 4.02, *SD* = .97) and (*M* = 3.45, *SD* = .8) 4.02 (.97) did not, *t*(19) = 1.61, *p* = .06.

Table 7: Regression Analyses Results

		Fixation	Saccade
Q1	R Squared	0.31	0.05
	p =	0.44	0.76
Q2	R Squared	0.47	0.14
	p =	0.32	0.62
Q3	R Squared	0.28	0.11
	p =	0.47	0.65
Q4	R Squared	0.52	0.17
	p =	0.28	0.6

(**Table 7**) displays the regression analyses that were conducted. Unfortunately, fixation duration and saccade instance failed to be significant predictors of post-questionnaire responses.

Discussion

The present study aimed to investigate to what effect, if any, that a false-alarm rate of a decision support system (DSS) could have on user trust, usage, and gaze behavior during a cognitive search task, in order to propose a human-decision aid trust interaction model. The investigation of participant gaze behavior allowed us to estimate user usage of the accompanying decision aid during the 20 search tasks. Four false alarm rates were investigated, including the average fixation time and response along the four conditions (**Figure 3**). Although there is evidence towards a formation of a trend, no significant relation was found between fixation time and the corresponding responses. Still, these findings seem to mirror past investigations into the effect of various false alarm rates on participant trust where a single instance of a false alarm significantly reduced participant trust, (Cafarelli, 1998).

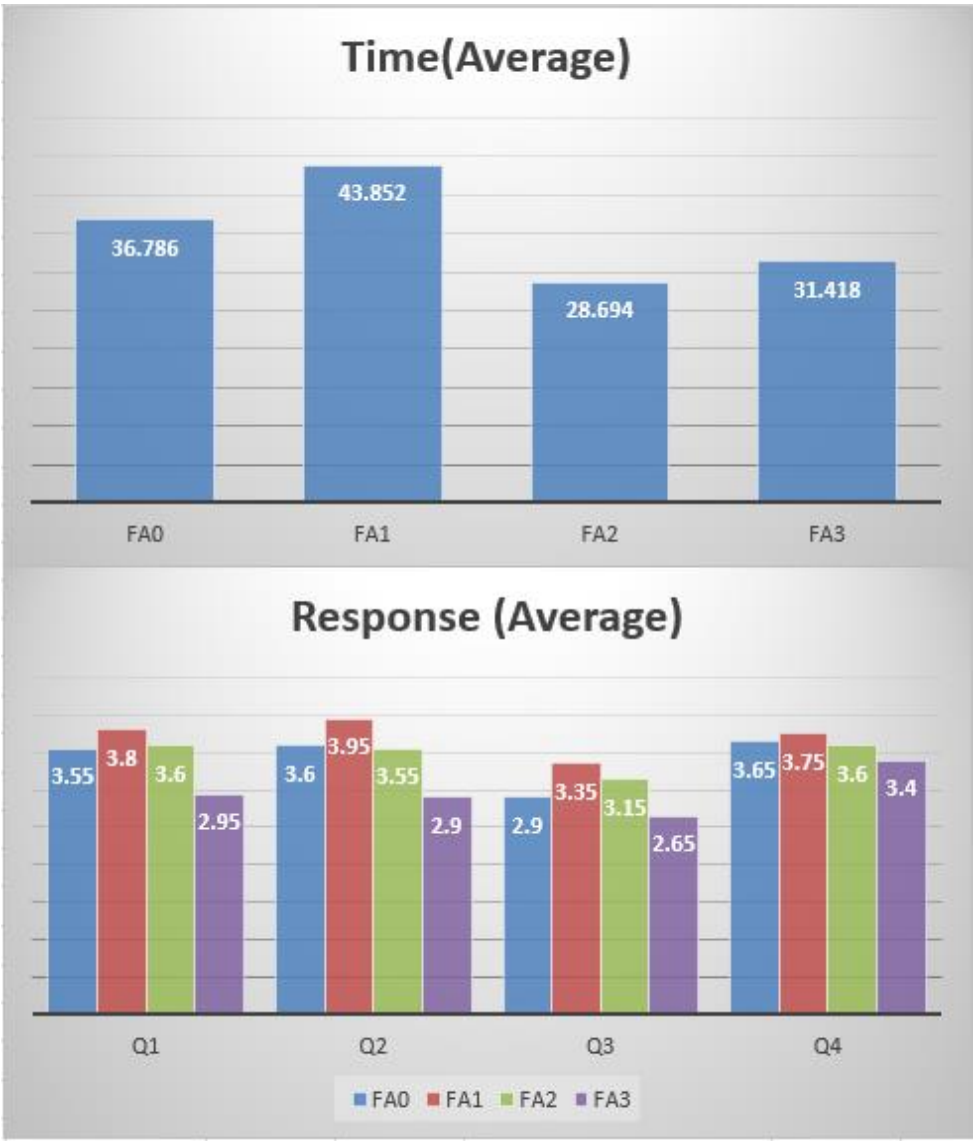


Figure 3: Average Fixation Times and Response

In the present study, an examination into the means of fixation time and participant response after the first instance of a false alarm show them both decreasing, though not to a significant degree. In fact, further examination reveals that the average response for every item of the post-questionnaire reduces, indicating the possibility of a trend.

Finally, within the Trust between People and Automation survey, a comparison of the negative and positive connotation groups shows promise. Between the first and last condition, the negative connotation scores significantly increased (p=.001) whilst the positive connotation approached a significant decrease in scores (p=.06). These results seemingly indicate there is in fact a decrease in trust and confidence in the system as the false alarm rate increases.

Information Processing Model Adaptation

A cognitive model is a visual map of human problem solving and/or mental processing that revolves around a task or set of tasks that can be utilized for predicting human behavior and performance. For our research, we searched for processing models that could accommodate the addition of trust and decision aid-human interaction within automation. Wickens' (2003) information processing model was one such model, which focused on an event stimulus, user perception, decision, and their response.

Our proposed decision model consists of two major changes to Wickens' original model of information processing. First, is the correspondence of trust between the human operators' perception and the DSS (Decision Support System). Based on the result of the response from DSS and the execution of that response, the human operator begins to develop a judgment about the reliability of the DSS. The second major change is the beforementioned judgement, which is shown with the user's memory in terms of their future compliance. Our proposed model provides a cognitive framework of trust between the user and the DSS, (**Figure 4**).

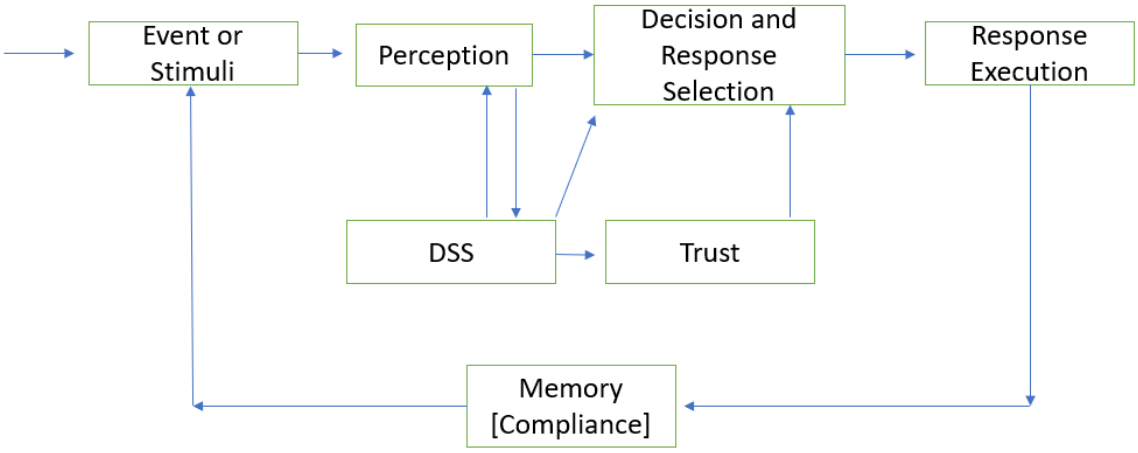


Figure 4: Proposed Processing Model

The examination of user gaze behavior indicated that following the perception of the stimuli, there are two user paths: Consultation and Non-Consultation. Consulters will examine the DSS and modify their perception of the stimuli when searching for a target (**Figure 5**). On the contrary, Non-consulters will bypass the DSS entirely and subsequently neither gain nor lose trust in the system, while still executing a response.

Our model proposal was built on one main assumption: That the accuracy of the DSS would impact user trust in the system, and ultimately effect compliance or future consultation. The present findings provide modest support of this assumption. Our examination of gaze behavior did not reveal a significant decrease in DSS consultation as the false-alarm rate increased, nor did self-reported trust or usage decrease to any significant degree. Interestingly however, user comfort with the DSS and identification performance both significantly decreased indicating participants began to manifest a negative relationship with the DSS.

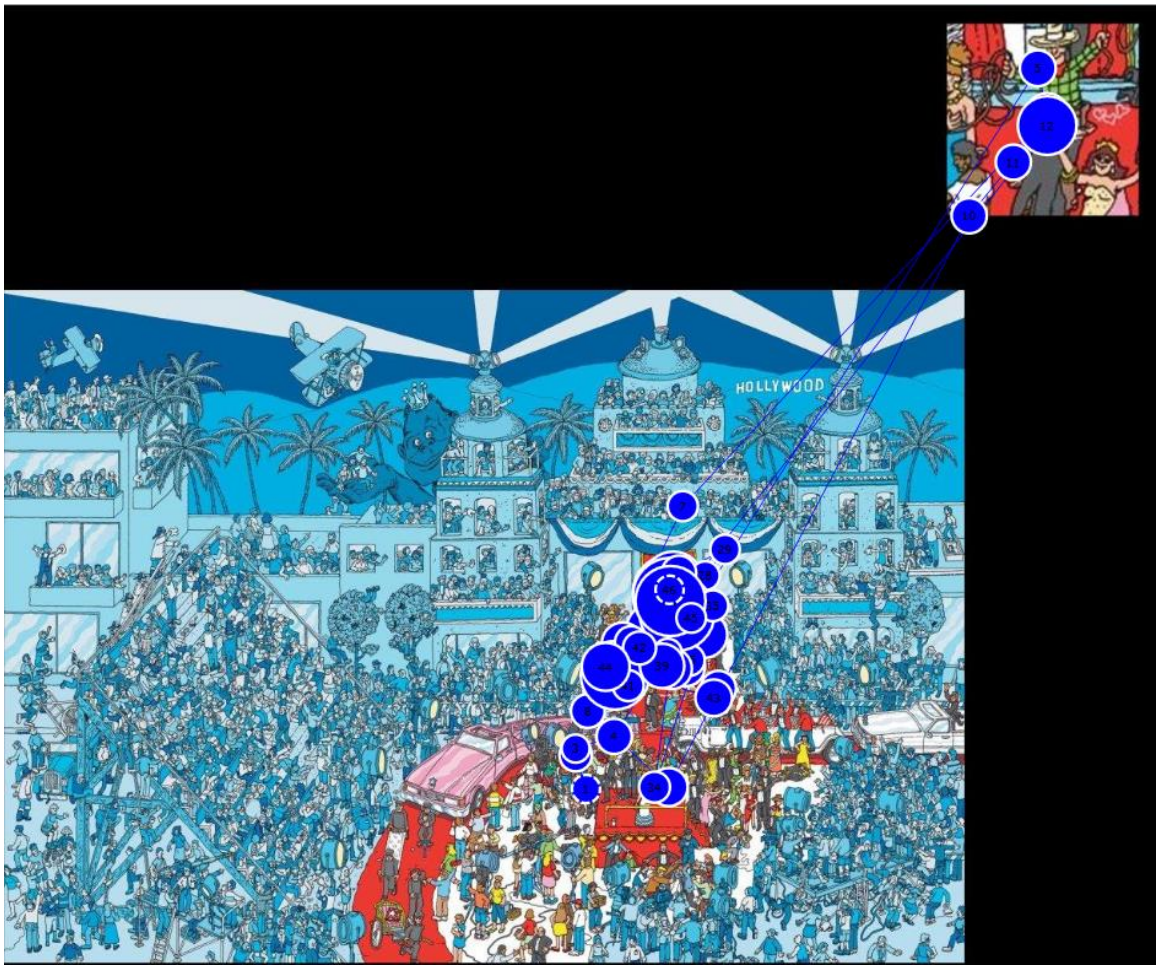


Figure 5: DSS Consulter Gaze Behavior

Previously mentioned, the Trust between People and Automation survey, (Jian, Bisantz & Drury, 2000) moderately supported our assumption. The significant increase in negative connotation scores and the near significant decrease in positive scores indicate that participants did become suspicious of the DSS after experiencing false alarms and did not trust the system. These findings would support our proposed processing model. Future research is warranted.

Limitations and Considerations for the ATR Study

During the Waldo pilot experiment, head movement of the participants was not restricted. Following calibration, participants could shift position in their seat, preventing the Tobii eye tracker from maintaining eye tracking. In one such instance, a participant abruptly lowered their seat, forcing a restart and recalibration with that condition. Additionally, although marketed as being able to track eye movement through eye-glasses, multiple calibrations were required on one instance where a participant's glasses were preventing successful calibration results. Most importantly due to the nature of the task and the size of the testing screen, participants unconsciously leaned in towards the screen, placing them outside of the 50 – 80cm tracking distance that the Tobii T120 was capable of. This was regardless of the users corrected vision. The structure of the Post questionnaire being at the end of each false alarm condition also provided us with less data than we would have liked.

To account for these limitations, the ATR study was to be conducted with a couple different circumstances. First, the testing seat would have to be affixed to the ground in order to prevent the participant from shifting or lowering their position. The experimenter will also make a point to play close attention to the distance that the participant is from the Tobii testing screen in order to keep calibration. As for the post-questionnaire, we proposed its administration after each task as opposed to each condition, allowing us to measure any fluctuations within the conditions. Additionally, due to its repeat administration, we decided to reduce the number of questions from four to two, focusing solely on the Trust and Usage aspects of the algorithm as the false alarm rate and task difficulty changed.

The Waldo Pilot experiment lasted roughly 35 minutes from start to finish per participant. Based on participant feedback, this time frame was rated quite comfortably. This fact was considered for the development of the ATR experiment to prevent fatigue and promote task specific attention.

Eye-tracking to Evaluate Impact of Trust in Human-ATR Interaction

The current study aims at uncovering the relationship between a systems' false alarm rate and the impact on its user's reported trust and usage. Twelve tracking tasks with the presence of an ATR algorithm were completed with various false alarm rates and difficulties. We hypothesized that an increase in the number of false alarms would negatively impact user reported trust and use in the targeting algorithm. Additionally, we expect a decline in the fixation duration and fixation count on the targeting algorithm with the increased false alarm rate. The results from this research provide implications in the design of targeting and tracking technologies.

Methods

Participants were 25 (14 male and 11 female) undergraduate and graduate students between the ages of 18 and 45 years of age (M=25.09, SD = 6.39) and were required to hold U.S. citizenship. Participation was on a voluntary basis, and due to the visual nature of this study, we restricted eligibility to individuals that disclosed they had actively corrected vision and no visual or motor dysfunction. No participants failed to meet these requirements prior to their involvement. Participants were not compensated for this experiment.

Apparatus and Stimuli

This experiment utilized twelve pre-recorded videos from a Tower Data Collection Set, provided by Etegent Inc. Each video provided two angles of security camera footage overlooking an entrance and parking lot to a building compound as well as a moderately

trafficked roadway in Dayton, OH. The two camera feeds were oriented vertically and positioned on the left half of the simulation, with the entire right half consisting of the targeting algorithm, (**Figure 6**). The targeting algorithm consisted of a map, output, and record component. The map component, see in the upper half of the algorithm, provided the user with a birds-eye view of the compound/roadway, as well as two highlighted areas that corresponded with each camera view. Additionally, this component illustrated all instances that the algorithm had detected the task TOI's with red and blue indicators, providing the user with a tracking reference during the experiment. The output component in the center displayed the total number of detections by the algorithm for that task, and the records component provided further details of these detections such as the time they were detected, type of vehicle, and their respective color.

Participants were seated in a sound and light controlled room and completed the visual search tasks using a Tobii T120 eye tracker. The T120's monitor has a tracking distance between 50 to 80 centimeters and services a gaze angle capability of 35 degrees in all directions allowing

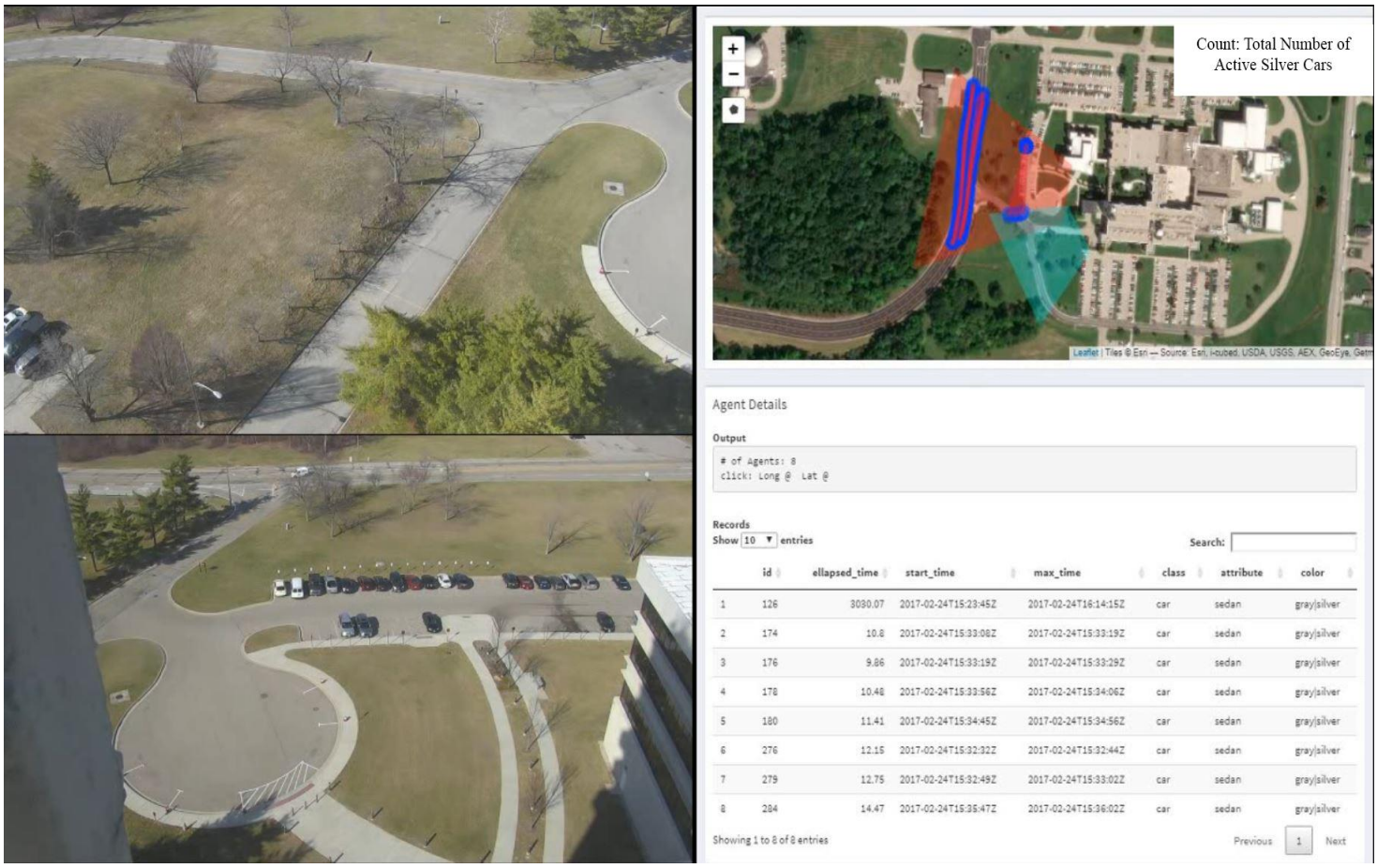


Figure 6: Experimental Interface

a moderate degree of head movement from the participants. The data rate of the T120 ranges from 60 Hz to 120 Hz, with a screen resolution of 1280x 1024 pixels on a 17-inch display. Subjects' eye movements and fixation times were measured by heat maps and directional analytics. Data output includes timestamps, fixation duration and frequency, and saccade frequency.

Design and Procedure

The current experiment was treated as a 3x4 within-subjects design. The primary independent variable was the false alarm rate (0-3 False Alarms). Each participant progressed linearly through the conditions which consisted of three separate tasks of tracking one, two or three targets of interest (TOI) for a total of twelve tracking tasks. Each trial had a duration of 90 seconds. Participants were asked to count the number of active vehicles of a specified color, entering a designated area, or traveling a specified direction. (**Table 8**) displays a sample tasking that was used for one such condition.

Table 8: Sample Tasking

Number of Target(s) of Interest (TOI)	Example Target Description(s)
1	Count: Number of total active Silver Cars
2	Count: Number of active Black Trucks on roadway and Vehicles entering parking lot
3	Count: Number of total White SUV's, Vehicles entering parking lot, and Vehicles leaving compound traveling West.

(**Figure 7**) shows that after completing an informed consent form and pre-questionnaire which included the assessment of age, gender, visual function and citizenship status, participants were shown an introductory walkthrough/training video with a sample tasking to ensure understanding of the targeting algorithm, and the nature of their role in the experiment. Following calibration of the Tobii T120 eye tracker, participants then began the experiment

with no knowledge of the algorithm's accuracy. Prior to each trial video, participants were shown an instructions screen that listed the TOI's for the coming task as well as an

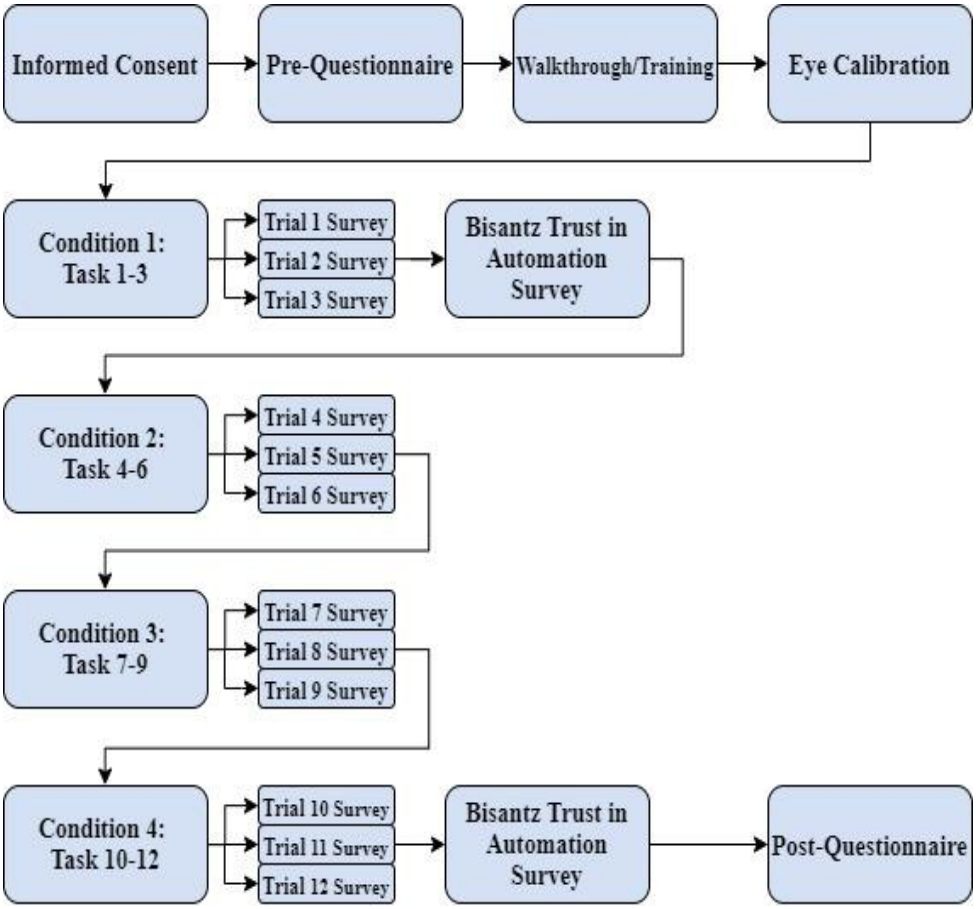


Figure 7: Procedure

illustration depicting the differences between cars, trucks, and SUV's and were given unlimited time for review. Task TOI's were also made available on-screen during the experiment to ensure understanding. Participants were asked to communicate with the lead investigator a total verbal tally when a designated target(s) of interest was detected. Additionally, any comments during the task were encouraged to be relayed to the lead investigator for recording. Upon completion of each trial, a post questionnaire was administered that consisted of two

questions that pertained to the users trust and usage of the algorithm on scale from 1 to 5 ranging from “Strongly Disagree’ to ‘Strongly Agree.’ Additionally, after the first and last condition a measure of trust in automation was obtained using the Checklist for Trust between People and Automation (Jian et al [7]). This 12-item measure of five positive and seven negatively framed questions asks participants to assess their level of confidence and reliability

Table 9: Descriptive Statistics

			Condition 1 (0FA)	Condition 2 (1FA)	Condition 3 (2FA)	Condition 4 (3FA)	(1 TOI)	(2 TOI)	(3 TOI)
Fixation Duration (s)	Algorithm	*M*	9.34	4.54	6.97	8.89	4.58	6.52	11.23
		SD	7.18	3.29	4.78	6.94	5.02	4.26	6.63
	Map	*M*	1.28	0.50	0.45	0.63	0.68	0.52	0.94
		SD	1.95	0.74	0.75	1.22	1.30	0.78	1.65
	Output	*M*	1.92	1.01	0.86	0.88	1.00	1.48	1.03
		SD	1.97	1.05	0.87	0.95	1.65	1.56	1.34
	Records	*M*	3.01	1.20	1.53	2.73	2.22	2.05	2.09
		SD	3.67	1.93	2.70	4.76	4.05	2.73	3.64
Fixation Count (N)	Algorithm	*M*	28.32	14.63	24.89	29.38	13.99	20.98	38.01
		SD	21.28	11.12	18.23	23.77	14.74	13.00	22.63
	Map	*M*	4.88	2.20	1.93	2.46	2.35	2.19	4.07
		SD	5.82	2.76	2.78	3.57	3.81	2.94	5.00
	Output	*M*	4.85	2.67	2.42	2.46	2.73	3.80	2.77
		SD	4.72	2.80	2.50	2.64	3.10	3.83	3.24
	Records	*M*	8.65	3.84	5.00	8.12	6.69	6.04	6.51
		SD	10.17	6.02	8.406	13.76	11.67	7.77	10.77
Post-Questionnaire	Trust	*M*	3.85	3.52	3.58	3.04			
		SD	0.87	0.95	0.97	1.03			
	Use	*M*	3.73	3.91	4.03	3.72			
		SD	1.11	0.97	0.99	1.05			
Automation Survey	Negative	*M*	2.43			2.84			
		SD	1.38			1.60			
	Positive	*M*	4.45			4.15			
		SD	1.35			1.50			

in the system. Responses are coded on scale from 1 to 7 ranging from 'Not at all' to 'Extremely.' Finally, a qualitative free-report questionnaire was given at the end of all conditions that asked participants how they thought the targeting algorithm impacted their decision making, and at what point trust and use may have begun to deteriorate. Participant gaze behavior was monitored throughout the duration of the experiment to examine use in the algorithm.

Results

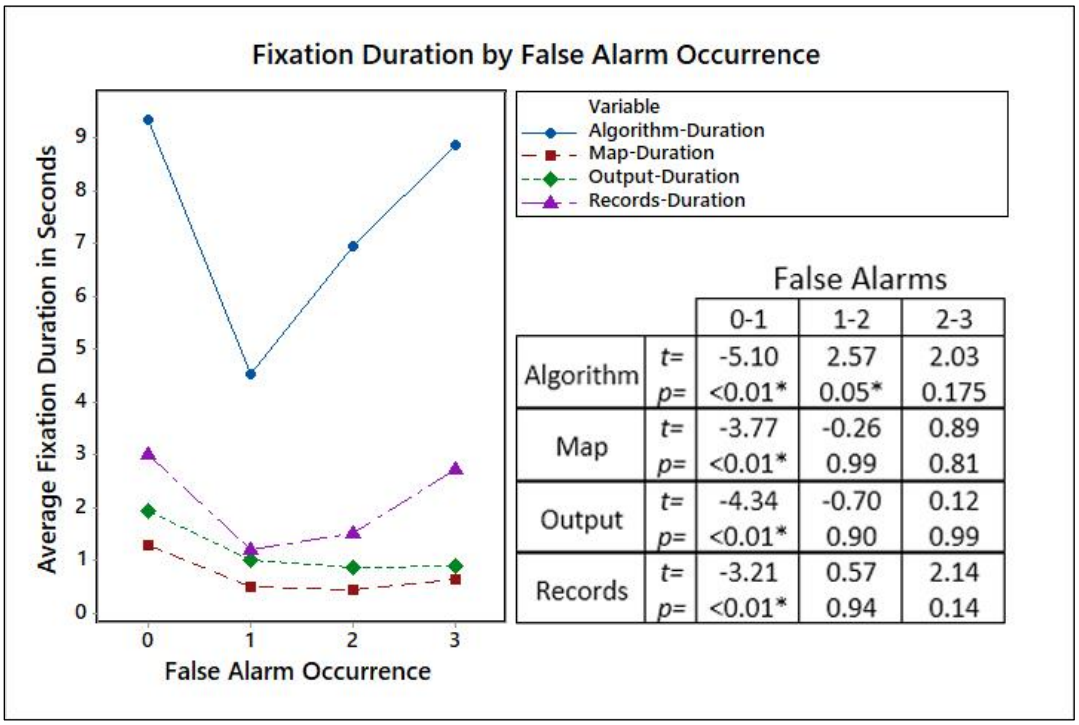
(**Table 9**) displays basic descriptive statistics for the fixation duration and count on the algorithm and its components, as well as the post-questionnaire and Trust in Automation scores

Table 10: Analysis of Variance Results

	Component	False Alarm Occurrence	Target(s) of Interest	Trust	Use
Fixation Duration	Algorithm	F(3,296) = 10.80, p<0.01	F(2,297) = 40.26, p<0.01	F(4,295) = 1.08, p=0.37	F(4,295) = 2.29, p=0.06
	Map	F(3,296) = 6.93, p<0.01	F(2,297) = 2.74, p=0.06	F(4,295) = 0.29, p=0.88	F(4,295) = 2.06, p=0.09
	Output	F(3,296) = 11.53, p<0.01	F(2,297) =3.89, p<0.05	F(4,295) = 1.94, p=0.10	F(4,295) = 2.40, p=0.05
	Records	F(3,296) = 4.96, p<0.01	F(2,297) = 0.07, p=0.93	F(4,295) = 1.98, p=0.09	F(4,295) = 1.87, p=0.12
Fixation Count	Algorithm	F(3,296) = 9.22, p<0.01	F(2,297) = 51.01, p<0.01	F(4,295) = 0.81, p=0.52	F(4,295) = 2.48, p<0.05
	Map	F(3,296) = 8.92, p<0.01	F(2,297) = 6.75, p<0.01	F(4,295) = 0.40, p=0.81	F(4,295) = 1.49, p=0.21
	Output	F(3,296) = 9.54, p<0.01	F(2,297) = 3.17, p<0.05	F(4,295) = 3.44, p<0.01	F(4,295) = 2.86, p<0.01
	Records	F(3,296) = 4.12, p<0.01	F(2,297) = 0.11, p=0.90	F(4,295) = 2.20, p=0.07	F(4,295) = 2.23, p=0.06

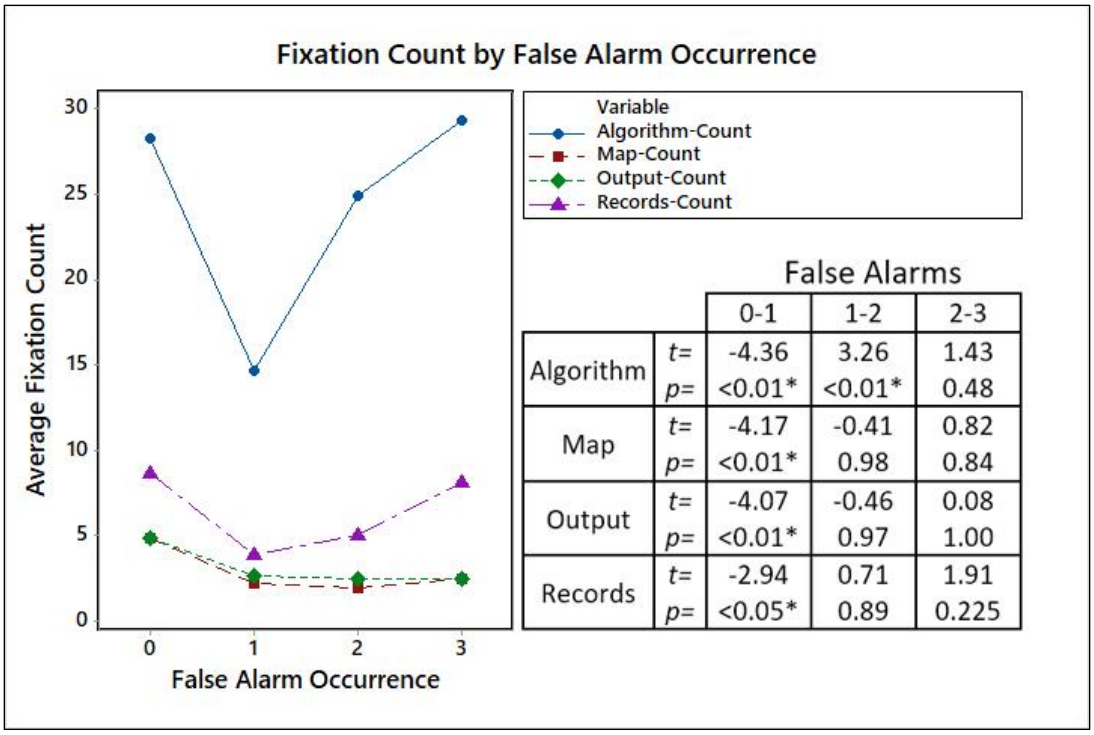
Grayed areas denote insignificant findings (p>0.05)

Figures 8 and 9 display the Tukey's pairwise comparisons findings across each false false alarm condition in relation to the respective measurements. As seen, between the zero and one false alarm conditions, a significant decline in both measurements on all components of the algorithm were detected suggesting a rapid degradation of user-algorithm interaction following the witness of the first false alarm. Subsequently, between the one and two false alarm conditions, a significant increase was detected regarding the algorithm as a whole, indicating a substantial resume of human-system interaction, $t(3) = 2.57$ and 3.26, $p<0.05$ and $p<0.01$.



		False Alarms		
		0-1	1-2	2-3
Algorithm	*t*=	-5.10	2.57	2.03
	p=	<0.01*	0.05*	0.175
Map	*t*=	-3.77	-0.26	0.89
	p=	<0.01*	0.99	0.81
Output	*t*=	-4.34	-0.70	0.12
	p=	<0.01*	0.90	0.99
Records	*t*=	-3.21	0.57	2.14
	p=	<0.01*	0.94	0.14

Figure 8: Fixation Duration by False Alarms

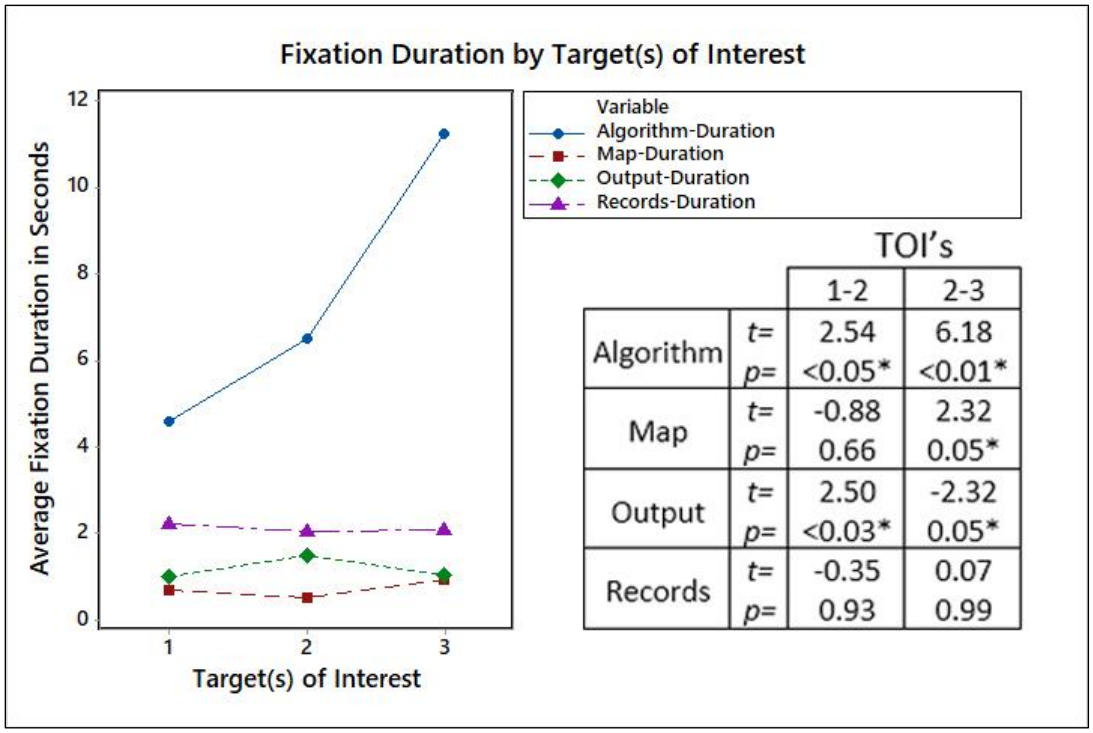


		False Alarms		
		0-1	1-2	2-3
Algorithm	*t*=	-4.36	3.26	1.43
	p=	<0.01*	<0.01*	0.48
Map	*t*=	-4.17	-0.41	0.82
	p=	<0.01*	0.98	0.84
Output	*t*=	-4.07	-0.46	0.08
	p=	<0.01*	0.97	1.00
Records	*t*=	-2.94	0.71	1.91
	p=	<0.05*	0.89	0.225

Figure 9: Fixation Count by False Alarms

Figure 10 and 11 display the pairwise comparisons across the three target(s) of interest levels. In both measures between all target levels, there was a significant positive increase found with the algorithm, more so between the second and third levels, $t(2) = 6.18$ and 6.96, $p<0.01$, indicating substantial system-consultation at higher workloads. There were similar increases detected with the Output component between the first and second levels, which conversely decreased between the second and third. A significant increase in fixation on the map component was also detected.

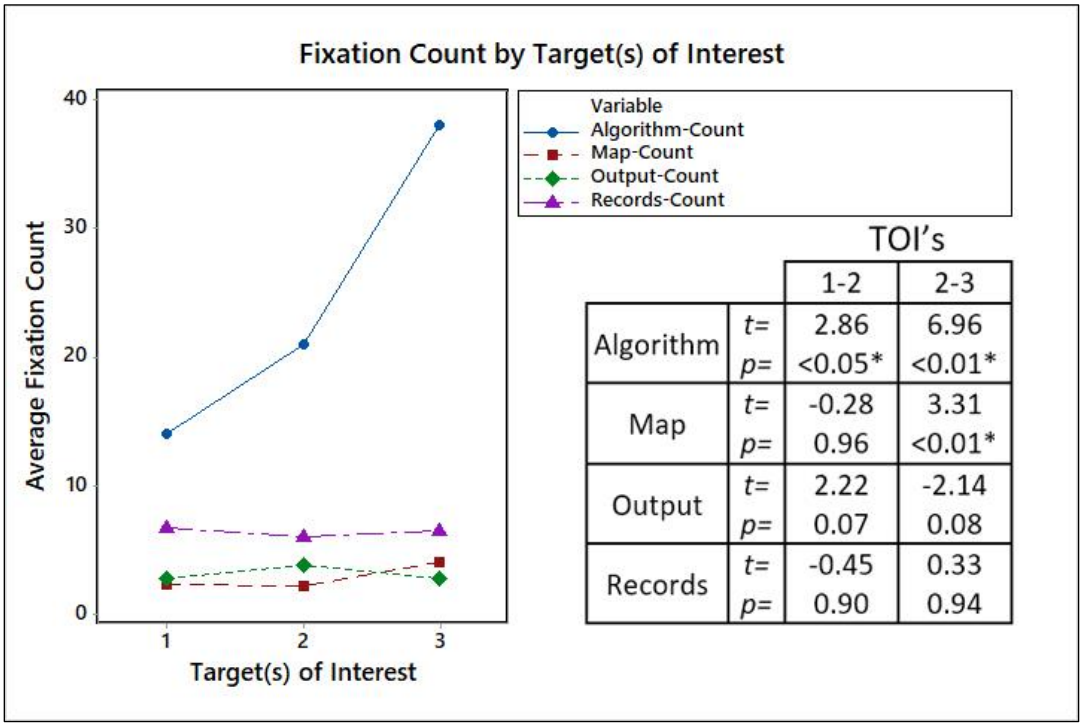
Figure 12 and 13 display the pairwise comparison of both the false alarms and target(s) of interest levels with respect to participant submitted trust and usage levels. Regarding the false alarm rate, a steady decline in user trust ratings is seen in response to the



TOI's

		1-2	2-3
Algorithm	*t*=	2.54	6.18
	p=	<0.05*	<0.01*
Map	*t*=	-0.88	2.32
	p=	0.66	0.05*
Output	*t*=	2.50	-2.32
	p=	<0.03*	0.05*
Records	*t*=	-0.35	0.07
	p=	0.93	0.99

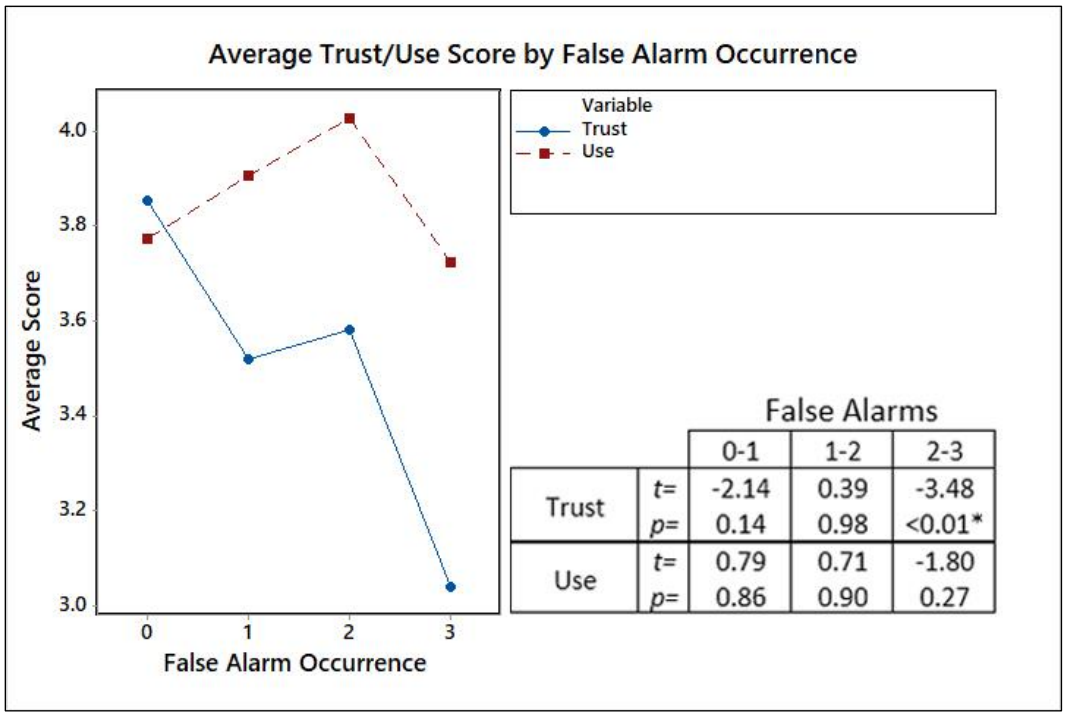
Figure 10: Fixation Duration by Target(s) of Interest



TOI's

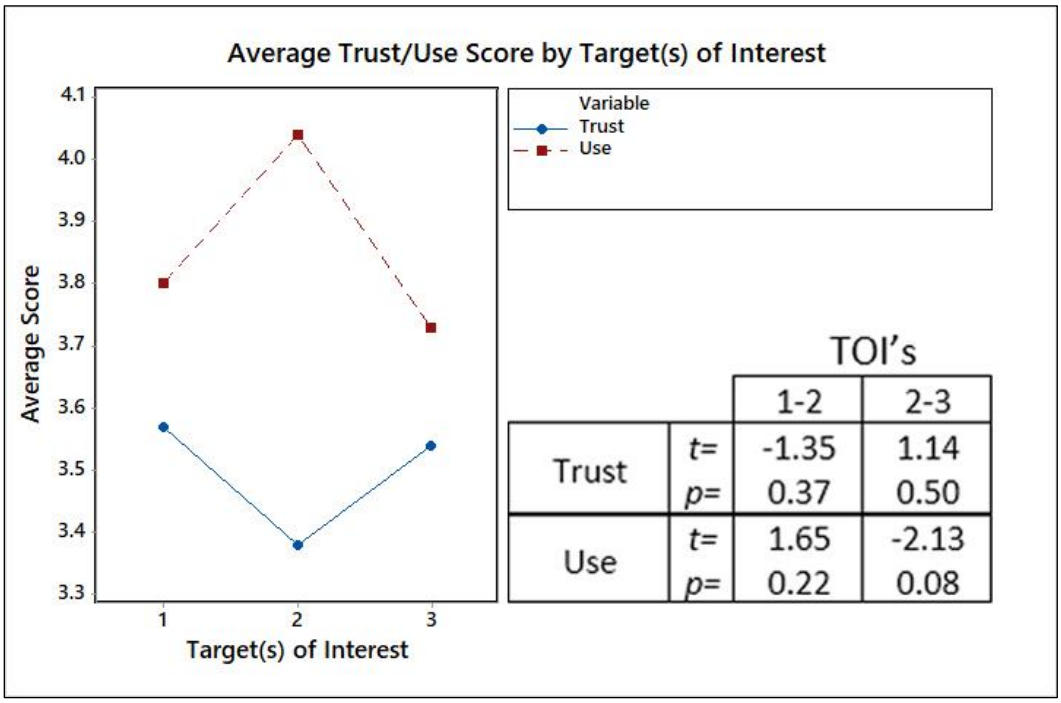
		1-2	2-3
Algorithm	*t*=	2.86	6.96
	p=	<0.05*	<0.01*
Map	*t*=	-0.28	3.31
	p=	0.96	<0.01*
Output	*t*=	2.22	-2.14
	p=	0.07	0.08
Records	*t*=	-0.45	0.33
	p=	0.90	0.94

Figure 11: Fixation Count by Target(s) of Interest



		False Alarms		
		0-1	1-2	2-3
Trust	*t*=	-2.14	0.39	-3.48
	p=	0.14	0.98	<0.01*
Use	*t*=	0.79	0.71	-1.80
	p=	0.86	0.90	0.27

Figure 12: Trust and Use by False Alarms



		TOI's	
		1-2	2-3
Trust	*t*=	-1.35	1.14
	p=	0.37	0.50
Use	*t*=	1.65	-2.13
	p=	0.22	0.08

Figure 13: Trust and Use by Target(s) of Interest

alarms, even more so between the second and third encounter, $t(3) = -3.48$, $p<0.01$. The second to third encounter also appeared to negatively impact reported usage ratings, but not to any significant degree. Trust scores did not seem to be influenced by the change in target of interest levels, but usage scores did decline considerably in the third level $t(2) = -2.13$, $p=0.08$, suggesting a depart from system-consultation in higher cognitive workloads despite the beforementioned fixation findings.

Similarly, to (Bisantz & Seong, 2001), we divided the automation survey into two groups for analysis. The first group, which consisted of questions one through five, were labeled 'negatively framed questions,' as they centered around deception and suspicions of the aid. The second group consisted of questions six through twelve which contrarily were labeled 'positively framed questions,' as they centered around the integrity and dependability of the aid. The average of these scores were collected after the first and last condition for measurement. There was no significant difference found between the negatively framed responses from Condition 1 ($M = 2.43$, $SD = 1.38$) and Condition 4 ($M = 2.84$, $SD = 1.60$), $t(7) = -1.28$, $p=0.241$, nor was any significant difference found ($M = 4.45$, $SD = 1.35$) and Condition 4 ($M = 4.15$, $SD = 1.50$),

$t(7) = 1.61$, $p=0.151$, though their trends followed expected directions.

Discussion

Using pre-truthed video footage, it was investigated how false alarms would impact user trust and use in an ATR algorithm. Throughout the experiment, participants completed various

tasks where they had to locate and count specified vehicles. We compared user eye movement as well as self-reported trust and use scores between a perfect targeting algorithm and an algorithm with one, two or three false alarms. Our findings resemble the cry wolf effect, a phenomenon that has been detailed in several different scenarios where the occurrence of a single false alarm provoked a deleterious effect on compliance and use in a system (Roulston & Smith, 2004; Bliss, 1993). Contrarily, in correspondence with McBride, instead of detecting a decline in usage rates, we found that a single false alarm occurrence rapidly degraded user trust in the targeting algorithm, as evidenced by the reported decline of trust scores and corresponding fixation measures. These trust scores did not significantly revert throughout the duration of the experiment, demonstrating a lasting distrust in the system. A significant resumption was seen, however, in both fixation measures for all components following the occurrence of the second false alarm. In terms of compliance, usage scores actually increased after the occurrence of a false alarm- in fact, it was not until users detected three false alarms where algorithm consultation dropped. With this, we submit that users continued to consult the algorithm for quite some time regardless of their loss of trust.

The task TOI level was also manipulated during the experiment into three workloads, (Low, Moderate, High). This was achieved by requiring participants to identify and count one, two or three types of vehicles, respectively. The TOI level demonstrated a significant positive relationship with both fixation measures on the detection aid suggesting an increased level of algorithm consultation at higher workloads but possessed no definitive relationship between the self-reported trust and use scores. Similarly to (McBride, 2011), the increase in workload between the one and two TOI tasks led to higher usage scores in our experiment, but differed

in demonstrating lower usage scores in the three TOI tasks. This decline in usage was paired with a nearly equivalent increase in trust scores, mirroring (Yuan, 2017) suggestions that there is little to no effect of workload on trust in automation.

We were unable to find statistically significant changes in the Trust in Automation survey scores between the first and last conditions, but these demonstrated appropriate trends that fit our initial assumptions. The increased scores of the negatively framed questions and the decreased scores of the positively framed questions indicate that participants became more suspicious, more distrusting, and less dependent on the ATR algorithm as the false alarm rate increased throughout the duration of the experiment. In fact, the free-form qualitative responses collected at the end of the study reflected that very sentiment. Key words such as skeptical, suspicious and distrust were commonly described in these responses, and fittingly, the majority of participants disclosed that they had lost the most trust in the algorithm at the final three false alarm rate condition, further supporting our findings.

Conclusions and Future Work

Two studies were conducted to investigate how false alarms impacted user trust and use with an DSS and ATR algorithm, exploring the use of eye-tracking metrics as an objective measure to validate the use of surveys. In both experiments, but more so in the ATR study, the fixation measurements on the DSS and the ATR Algorithm decreased following the increase in false alarms. When these eye fixation measures were compared concurrently with the self-reported trust scores, it was seen that initially they decreased with one another, but that some level of consultation was maintained throughout the duration of the experiment as evident by

the steady usage scores and resumption of fixation ratings on the algorithm. These findings suggest various eye tracking measures could be used to identify the level of trust and use of a system by an operator.

The findings of these studies contribute to the designs of targeting algorithms in a variety of circumstances. By demonstrating how the false alarm rate affects operators' trust and use in the algorithm, developers can consider to what extent the intensity of their detection software should operate at in order to optimize user compliance. Additionally, our work examining the trends in usage of the components of the algorithm gives reference for the future development of these tracking tools. Further investigation into these components and their influence on task accuracy and performance is warranted. Our work contributes to the body of knowledge regarding use of eye-tracking measures to understand underlying cognitive mechanisms that impact human-machine interaction.

References

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276.

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

Bliss, J. P., & Gilson, R. D. (1998). Emergency signal failure: Implications and recommendations. *Ergonomics*, *41*(1), 57-72.

Bliss, James P., "The cry-wolf phenomenon and its effect on alarm responses" (1993). *Retrospective Theses and Dissertations*. 3614.
https://stars.library.ucf.edu/rtd/3614

Cafarelli, D. A. (1998). *Effect of false alarm rate on pilot use and trust of automation under conditions of simulated high risk* (Doctoral dissertation, Massachusetts Institute of Technology).

Chancey, E. T., Yamani, Y., Brill, J. C., & Bliss, J. P. (2017, September). Effects of Alarm System Error Bias and Reliability on Performance Measures in a Multitasking Environment: Are False Alarms Really Worse than Misses?. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1621-1625). Sage CA: Los Angeles, CA: SAGE Publications.

Chandon, P., Hutchinson, J., Bradlow, E., & Young, S. H. (2006). Measuring the value of point-of-purchase marketing with commercial eye-tracking data. *INSEAD Business School Research Paper*, (2007/22).

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

Clemente, C., Pallotta, L., Gaglione, D., De Maio, A., & Soraghan, J. J. (2017). Automatic Target Recognition of Military Vehicles With Krawtchouk Moments. *IEEE Trans. Aerospace and Electronic Systems*, *53*(1), 493-500.

de Greef, T., Lafeber, H., van Oostendorp, H., & Lindenberg, J. (2009, July). Eye movement as indicators of mental workload to trigger adaptive automation. In *International Conference on Foundations of Augmented Cognition* (pp. 219-228). Springer, Berlin, Heidelberg.

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology*, *18*(6), 399-411.

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

Dougherty, L. (2005). *Automatic Target Recognition, Executive Summary and Annotated Brief (PR)* (No. SAB-TR-05-02-PR). SCIENTIFIC ADVISORY BOARD (AIR FORCE) WASHINGTON DC.

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

Ezzati, M., Martin, H., Skjold, S., Hoorn, S. V., & Murray, C. J. (2006). Trends in national and state-level obesity in the USA after correction for self-report bias: analysis of health surveys. *Journal of the Royal Society of Medicine*, *99*(5), 250-257.

Geels-Blair, K., Rice, S., & Schwark, J. (2013). Using system-wide trust theory to reveal the contagion effects of automation false alarms and misses on compliance and reliance in a simulated aviation task. *The International Journal of Aviation Psychology*, *23*(3), 245-266.

Gredebäck, G., Johnson, S., & von Hofsten, C. (2009). Eye tracking in infancy research. *Developmental Neuropsychology*, *35*(1), 1-19.

Hauland, G. (2008). Measuring individual and team situation awareness during planning tasks in training of en route air traffic control. *The International Journal of Aviation Psychology*, *18*(3), 290-304.

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

Iacono, W. G., Peloquin, L. J., Lumry, A. E., Valentine, R. H., & Tuason, V. B. (1982). Eye tracking in patients with unipolar and bipolar affective disorders in remission. *Journal of Abnormal Psychology*, *91*(1), 35.

Iqbal, S. T., Zheng, X. S., & Bailey, B. P. (2004, April). Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on Human Factors in Computing Systems* (pp. 1477-1480).

Irvine, J. M., Leonard, J., Doucette, P., & Martin, A. (2008, May). An approach for evaluating assisted target detection technology. In *Signal Processing, Sensor Fusion, and Target Recognition XVII* (Vol. 6968, p. 69680I). International Society for Optics and Photonics.

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.

Jones, G., & Bhanu, B. (1999). Recognition of articulated and occluded objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *21*(7), 603-613.

Kaber, D. B., & Endsley, M. R. (1997, October). The combined effect of level of automation and adaptive automation on human performance with complex, dynamic control systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 41, No. 1, pp. 205-209). Sage CA: Los Angeles, CA: SAGE Publications.

Khushaba, R. N., Wise, C., Kodagoda, S., Louviere, J., Kahn, B. E., & Townsend, C. (2013). Consumer neuroscience: Assessing the brain response to marketing stimuli using electroencephalogram (EEG) and eye tracking. *Expert Systems with Applications*, *40*(9), 3803-3812.

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

Kraemer, S., Carayon, P., & Sanquist, T. F. (2009). Human and organizational factors in security screening and inspection systems: conceptual framework and key research needs. *Cognition, Technology & Work*, *11*(1), 29-41.

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

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

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2011). Understanding the effect of workload on automation use for younger and older adults. *Human Factors*, *53*(6), 672-686.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334-359.

Meißner, M., & Oll, J. (2019). The promise of eye-tracking methodology in organizational research: A taxonomy, review, and future avenues. *Organizational Research Methods*, *22*(2), 590-617.

Mele, M. L., & Federici, S. (2012). Gaze and eye-tracking solutions for psychological research. *Cognitive Processing*, *13*(1), 261-265.

Merritt, S. M., Sinha, R., Curran, P. G., & Ilgen, D. R. (2015). Attitudinal predictors of relative reliance on human vs. automated advisors. *International Journal of Human Factors and Ergonomics*, *3*(3-4), 327-345.

Miller, A. L. (2011). Investigating Social Desirability Bias in Student Self-Report Surveys. *Association for Institutional Research (NJ1)*.

Moore, K., & Gugerty, L. (2010, September). Development of a novel measure of situation awareness: The case for eye movement analysis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 54, No. 19, pp. 1650-1654). Sage CA: Los Angeles, CA: SAGE Publications.

Morrison, J. G., Kelly, R. T., Moore, R. A., & Hutchins, S. G. (1998). Implications of decision-making research for decision support and displays.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors*, *50*(3), 511-520.

Rajagopal, A., Agarwal, S., & Ramakrishnan, S. (2005, December). Simulation-based performance modeling for war fighter in loop minefield detection system. In *Proceedings of the Winter Simulation Conference, 2005.* (pp. 1160-1169). IEEE.

Ratwani, R. M., McCurry, J. M., & Trafton, J. G. (2010, March). Single operator, multiple robots: an eye movement based theoretic model of operator situation awareness. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 235-242). IEEE.

Rice, S., & McCarley, J. S. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied*, *17*(4), 320.

Roulston, M. S., & Smith, L. A. (2004). The boy who cried wolf revisited: The impact of false alarm intolerance on cost–loss scenarios. *Weather and Forecasting*, *19*(2), 391-397.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393-404.

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.

Safar, J. A., & Turner, C. W. (2005, September). Validation of a two factor structure for system trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, No. 3, pp. 497-501). Sage CA: Los Angeles, CA: SAGE Publications.

Spain, R. D., Bustamante, E. A., & Bliss, J. P. (2008, September). Towards an empirically developed scale for system trust: Take two. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 52, No. 19, pp. 1335-1339). Sage CA: Los Angeles, CA: SAGE Publications.

Thomas, L. C., & Wickens, C. D. (2004, September). Eye-tracking and individual differences in off-normal event detection when flying with a synthetic vision system display. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, No. 1, pp. 223-227). Sage CA: Los Angeles, CA: Sage Publications.

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

Wang, Z., Du, L., Zhang, P., Li, L., Wang, F., Xu, S., & Su, H. (2017). Visual attention-based target detection and discrimination for high-resolution SAR images in complex scenes. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(4), 1855-1872.

Wedel, M., & Pieters, R. (2017). A review of eye-tracking research in marketing. In *Review of Marketing Eesearch* (pp. 123-147). Routledge.

Woods, D. D. (1996). Decomposing automation: Apparent simplicity, real complexity. *Automation and Human Performance: Theory and Applications*, 3-17.

Yuan Zhang, M., & Jessie Yang, X. (2017, September). Evaluating effects of workload on trust in automation, attention allocation and dual-task performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1799-1803). Sage CA: Los Angeles, CA: SAGE Publications.

Zhao, J., Zhang, Z., Yu, W., & Truong, T. K. (2018). A cascade coupled convolutional neural network guided visual attention method for ship detection from SAR images. *IEEE Access*, *6*, 50693-50708.