Trust in automation and the consequences of reliance on monitoring checks

Jenna E. Cotter

Follow this and additional works at: https://louis.uah.edu/uah-theses

Recommended Citation
https://louis.uah.edu/uah-theses/610

This Thesis is brought to you for free and open access by the UAH Electronic Theses and Dissertations at LOUIS. It has been accepted for inclusion in Theses by an authorized administrator of LOUIS.
TRUST IN AUTOMATION AND THE CONSEQUENCES OF RELIANCE ON MONITORING CHECKS

Jenna E. Cotter

A THESIS

Submitted in partial fulfillment of the requirements for the degree of Master of Arts in Psychology in The Department of Psychology to The Graduate School of The University of Alabama in Huntsville December 2023

Approved by:

Dr. Nathan Tenhundfeld, Research Advisor
Dr. Kristin Weger, Committee Member
Dr. Jodi Price, Committee Member
Dr. Nathan Tenhundfeld, Department Chair
Dr. Sean Lane, College Dean
Dr. Jon Hakkila, Graduate Dean
Abstract

TRUST IN AUTOMATION AND THE CONSEQUENCES OF RELIANCE ON MONITORING CHECKS

Jenna E. Cotter

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Arts in Psychology

The Department of Psychology

The University of Alabama in Huntsville

December 2023

As automation becomes increasingly prevalent across society, it is crucial to further understand the benefits or consequences of the solutions implemented in automated systems that support user interaction and the overall system design. While previous literature has provided valuable knowledge to the understanding of automation and trust, there is still uncertainty about how individuals perform and respond to automated systems that provide monitoring checks to maintain users’ attention. To address this gap, this study assessed the effects of varied reliability and monitoring checks on individuals interacting with an automated system. This study is among the first to examine the consequences of monitoring checks to further understand the impact on individuals when interacting with an automated system. Our study provides mixed findings that warrant further exploration into the design of automated systems and helps to expand upon existing theories regarding trust and automation in the field.
Acknowledgments

During my time at The University of Alabama in Huntsville, I have had an incredible support system, and this project would not have been possible without it. I would first like to extend thanks to Dr. Nathan Tenhundfeld for his mentorship and guidance throughout this project and to other Advanced Teaming, Technology, Automation, and Cognition (ATTAC) lab members. I would also like to thank Dr. Kristin Weger and Dr. Jodi Price for being members of my committee and for the guidance they both provided throughout this project. The ATTAC lab has been instrumental to this project, as they provided tremendous support and assisted with data collection. Additionally, I would like to thank my peers at Sam Houston State University and my cohort at The University of Alabama in Huntsville. Lastly, I would like to thank my parents, Susan and Paul Cotter; my siblings, Ian, Blake, and Kayla Cotter; my extended family; and my close friends for their unconditional support and understanding they have given me along this adventure.
# Table of Contents

Abstract .......................................................................................................................... 2  
Acknowledgments ......................................................................................................... 4  
Table of Contents ......................................................................................................... 5  
List of Figures .................................................................................................................. 10  
List of Tables ................................................................................................................... 12  

Chapter 1. Introduction .................................................................................................. 1  

1.1 Automation .............................................................................................................. 1  
1.2 Levels of Automation ............................................................................................. 3  
1.3 Stages of Automation ............................................................................................. 4  
1.4 Degrees of Automation ........................................................................................... 5  
1.5 Reliance ................................................................................................................... 6  
1.6 Trust ........................................................................................................................ 6  
1.7 Situational Awareness .............................................................................................. 9  

Chapter 2. Method ........................................................................................................ 14  

2.1 Participants .............................................................................................................. 14  
2.2 Design .................................................................................................................... 14  
2.3 Apparatus ............................................................................................................... 15  
2.4 Tasks ....................................................................................................................... 15  
2.4.1 Primary Task ...................................................................................................... 15  
2.4.2 Secondary Task ................................................................................................. 16  
2.4.3 Monitoring Check .............................................................................................. 16  

v
2.5 Instruments ........................................................................................................... 17

2.6 Measures .............................................................................................................. 19

2.6.1 Subjective Measures ......................................................................................... 19

2.6.2 Behavioral Measures ......................................................................................... 19

2.6.3 Eye-tracking Measures ....................................................................................... 19

2.7 Procedure .............................................................................................................. 20

Chapter 3. Results ...................................................................................................... 23

3.1 Manipulation Checks ........................................................................................... 24

3.2 Subjective Assessments ......................................................................................... 25

3.2.1 Trust (B1) ......................................................................................................... 25

3.2.2 Trust (B2) ......................................................................................................... 27

3.2.3 Complacency Potential (AICP-R) ..................................................................... 28

3.2.4 Correlations: Subjective Assessments ............................................................... 29

3.3 Behavioral Assessments ......................................................................................... 30

3.3.1 Performance (B1) ............................................................................................ 30

3.3.2 Performance (B2) ............................................................................................ 30

3.3.3 Performance: Omission Errors (B1) ................................................................. 30

3.3.4 Performance: Omission Errors (B2) ................................................................. 31

3.3.5 Performance: Commission Errors (B1) ............................................................ 31

3.3.6 Performance: Commission Errors (B2) ............................................................ 32
3.3.7 Performance: Additional Analyses .............................................. 33

3.3.8 Reaction Time to Monitoring Checks (B1) .................................. 34

3.3.9 Reaction Time to Monitoring Checks (B2) .................................. 34

3.3.10 Correlations: Subjective and Behavioral Assessments ................. 34

3.4 Eye-tracking Assessments .................................................................. 35

3.4.1 Proportion of Fixations within the Primary Task AOI (B1) .......... 35

3.4.2 Proportion of Fixations within the Primary Task AOI (B2) ........ 35

3.4.3 Proportion of Fixations within the Secondary Task AOI (B1) .... 38

3.4.4 Proportion of Fixations within the Secondary Task AOI (B2) .... 38

3.4.5 Proportion of Fixations within the Monitoring Check AOI (B1) ... 38

3.4.6 Proportion of Fixations within the Monitoring Check AOI (B2) ... 40

3.4.7 Dwell Time within the Primary Task AOI (B1) ......................... 42

3.4.8 Dwell Time within the Primary Task AOI (B2) ......................... 44

3.4.9 Dwell Time within the Secondary Task AOI (B1) ...................... 44

3.4.10 Dwell Time within the Secondary Task AOI (B2) ................. 45

3.4.11 Dwell Time within the Monitoring Check AOI (B1) .............. 45

3.4.12 Dwell Time within the Monitoring Check AOI (B2) ............. 47

3.4.13 Task Switching from Primary to Secondary AOI (B1) .......... 47

3.4.14 Task Switching from Primary to Secondary AOI (B2) .......... 47

3.4.15 Task Switching from Primary to Monitoring AOI (B1) .......... 49
List of Figures

**Figure 1.1** Degrees of automation (Adapted from Wickens *et al.*, 2010).......................... 5

**Figure 2.1** Desktop view seen by the participants. The monitoring check is black when first presented and turns blue if not dismissed by the participant.............................. 16

**Figure 2.2** Monitoring check seen by the participants..................................................... 17

**Figure 2.3** Instructions seen by participants before starting the simulation..................... 22

**Figure 3.1** Visualization of AOI placement................................................................. 23

**Figure 3.2** Frequency of the average seconds between errors for blocks......................... 25

**Figure 3.3** The time between errors (B1) and checks predicting individuals’ trust.
Error ribbons represent 95% confidence intervals.................................................... 26

**Figure 3.4** The time between errors (B1 & B2) and checks predicting individuals’ trust.
Error ribbons represent 95% confidence intervals.................................................... 28

**Figure 3.5** Scatterplot of alleviating workload and trust. Error ribbons represent 95%
confidence intervals .................................................................................................. 29

**Figure 3.6** The time between errors (B1) and checks predicting individuals’
performance on commission errors. Error ribbons represent 95% confidence intervals.. 32

**Figure 3.7** The time between errors (B1 & B2) and checks predicting primary
proportion fixation. Error ribbons represent 95% confidence intervals. ....................... 37

**Figure 3.8** The time between errors (B1) and checks predicting monitoring check
proportion fixation. Error ribbons represent 95% confidence intervals. ....................... 40

**Figure 3.9** The time between errors (B1 & B2) and checks predicting monitoring
check proportion fixations. Error ribbons represent 95% confidence intervals............ 42

**Figure 3.10** The time between errors (B1) and checks predicting individuals’ primary
task dwell time. Error ribbons represent 95% confidence intervals............................ 44
Figure 3.11 The time between errors (B1) and checks predicting individuals’ monitoring check dwell time. Error ribbons represent 95% confidence intervals........... 46

Figure 3.12 The time between errors (B1 & B2) and checks predicting switching between primary to secondary. Error ribbons represent 95% confidence intervals. ........ 49

Figure 3.13 The time between errors (B1) and checks predicting switching between primary to monitoring. Error ribbons represent 95% confidence intervals. ..................... 51

Figure 3.14 The time between errors (B1 & B2) and checks predicting switching between primary to monitoring. Error ribbons represent 95% confidence intervals....... 53

Figure 3.15 The time between errors (B1 & B2) and checks predicting switching between secondary to primary. Error ribbons represent 95% confidence intervals. ....... 56

Figure 3.16 The time between errors (B1) and checks predicting switching between monitoring to primary. Error ribbons represent 95% confidence intervals. ..................... 58

Figure 3.17 The time between errors (B1 & B2) and checks predicting switching between monitoring to primary. Error ribbons represent 95% confidence intervals....... 60

Figure 3.18 Block comparisons. Error bars represent 95% confidence intervals. ............ 63
List of Tables

Table 1.1 Sheridan & Verplank – Ten LOA for human-machine decision making. ........ 4

Table 1.2 Research questions (RQ) ............................................................................................................. 13

Table 2.1 The Automation-Induced Complacency Potential Rating (AICP-R)
(Merritt et al., 2019). 5-point Likert scale, 1 (strongly disagree) and 5 (strongly agree).
Reverse coded * .................................................................................................................................................. 18

Table 2.2 The Trust of Automated Systems Test (TOAST) (Wojiton et al., 2020). 7-point Likert scale, 1 (strongly disagree) and 7 (strongly agree). .......................................................... 18

Table 3.1 The time between errors (B1) and checks predicting individuals’ trust. ........ 26

Table 3.2 The time between errors (B1 & B2) and checks predicting individuals’ trust. 27

Table 3.3 The time between errors (B1) and checks predicting individuals’
performance on commission errors.................................................................................................................. 32

Table 3.4 The time between errors (B1 & B2) and checks predicting primary
proportion fixation. ........................................................................................................................................... 36

Table 3.5 The time between errors (B1) and checks predicting monitoring check
proportion fixation. ........................................................................................................................................... 39

Table 3.6 The time between errors (B1 & B2) and checks predicting monitoring check
proportion fixation. ........................................................................................................................................... 41

Table 3.7 The time between errors (B1) and checks predicting individuals’ primary
task dwell time. ............................................................................................................................................... 43

Table 3.8 The time between errors (B1) and checks predicting individuals’ monitoring
check dwell time. ............................................................................................................................................... 46

Table 3.9 The time between errors (B1 & B2) and checks predicting switching
between primary to secondary. ......................................................................................................................... 48
Table 3.10 The time between errors (B1) and checks predicting switching between primary to monitoring. .......................................................... 50

Table 3.11 The time between errors (B1 & B2) and checks predicting switching between primary to monitoring. .......................................................... 52

Table 3.12 The time between errors (B1 & B2) and checks predicting switching between secondary to primary. .......................................................... 55

Table 3.13 The time between errors (B1) and checks predicting switching between monitoring to primary. .......................................................... 57

Table 3.14 The time between errors (B1 & B2) and checks predicting switching between monitoring to primary. .......................................................... 59

Table 3.15 Block comparisons of paired samples t-tests. ........................................... 62

Table 3.16 Representative statements from participants explaining their preference to disable after block one. .......................................................... 67

Table 3.17 Representative statements from participants explaining their preference to disable after block two. .......................................................... 67

Table 3.18 Representative statements from participants explaining their preference to not disable after block one. .......................................................... 68

Table 3.19 Representative statements from participants explaining their preference to not disable after block two. .......................................................... 69
Chapter 1. Introduction

1.1 Automation

Automation is defined from a human-machine comparison as a system that executes a function, partially or fully, that was previously carried out by a human operator (Parasuraman & Riley, 1997). Automated systems have become widely available in society; they have increased performance (David, 2000; Smith et al., 2015), safety (Donovan et al., 2014; Ralston et al., 2014; Sheridan, 1992), enjoyment (Frison et al., 2017), productivity and efficiency (Anagnoste, 2017; Wang & Siau, 2019) while reducing the workload (David, 2000; Wickens et al., 2006). Automation can be found in vehicles such as advanced driver assistance systems (Merat & Lee, 2012), household devices like coffee machines and vacuums (Hassenzahl & Klapperich, 2014; Klapperich et al., 2020; H. Lee & Banerjee, 2016), military applications such as drones (David, 2000), computer-controlled displays (Parasuraman et al., 1992; Singh et al., 1997), and healthcare applications that reduce the workload of healthcare workers (Ahn et al., 2015; Bhosekar et al., 2021; Broadbent et al., 2009; Holden et al., 2013).

Somewhat paradoxically, automation is often implemented because of a desire to reduce workload, even though automation can increase workload because of monitoring requirements (Sheridan, 1992). In any human-system interaction wherein the system is not fully autonomous, the user is not obsolete and must remain attentive. The user needs to remain vigilant to accurately monitor the system and regain control in the event of a failure (Endsley & Kaber, 1999; Molloy & Parasuraman, 1996; Sebok & Wickens, 2017).
To promote the safe and efficient transfer of control, humans need to remain vigilant for long periods. However, sustained vigilance is effortful and thus very difficult to maintain (Warm et al., 2008). As a result of reduced vigilance, users can experience a decrease in attention and an increase in monitoring failures. Tasks such as driving demand sustained vigilance from users to remain aware of their surroundings in the event of failures (Greenlee et al., 2018). Because of the workload associated with sustained vigilance and potential issues with individuals over-trusting systems, there is a documented propensity for users to fail to adequately monitor a system or to begin engaging in secondary tasks (Naujoks et al., 2016). Examples of secondary tasks include the use of cell phones, eating, mind wandering, etc. By engaging in secondary tasks, the user further fails to adequately monitor the environment and surroundings, which can lead to greater degradation of an individual’s awareness.

The aforementioned issues could explain high-profile accidents seen with self-driving vehicles (e.g., Joshua Brown) (Abrams & Kurtz, 2016). As a result, companies like Chevrolet and Comma.AI have engineered solutions to try and ensure the user is paying attention. One example of this includes eye-tracking systems, which detect when the user is distracted or inattentive toward their surroundings (Said et al., 2018). Eye-tracking systems have previously shown high reliability in predicting constructs such as attention (de Winter et al., 2021; Diaz-Piedra et al., 2019; Hasanzadeh et al., 2017) and trust (Gold et al., 2015; Hergeth et al., 2016; Walker et al., 2018) of individuals using automation. However, there are serious concerns associated with the privacy of using eye-tracking systems (Abrams & Kurtz, 2016). Privacy issues aside, there can also be logistical issues, such as changes in luminance, which can affect the ability of an eye-
tracker to work properly and may produce noisy data when analyzing data (Tenhundfeld et al., 2019). Some companies, like Tesla, have implemented a behavioral attention check; every so often, the vehicle will require the human to apply a slight force to the steering wheel. If ignored, this request becomes more blatant until the vehicle eventually pulls to the side of the road and does not permit the driver to use the self-driving feature until they turn off and back on the vehicle.

As companies implement solutions for attention checks to promote greater awareness and performance, there has not been much research evaluating the consequences of repeated attention checks. This is important to understand because the consequences of having a system tell you to remain engaged may unknowingly affect the user. To further evaluate the consequences of monitoring checks being implemented in autonomous systems, this study seeks to establish a comprehensive understanding of automation, reliance, trust in automation, and situational awareness (SA).

1.2 Levels of Automation

Levels of automation (LOA) are defined as the resources that are allocated between humans and machines when performing tasks (Endsley & Kaber, 1999). Sheridan and Verplanck (1978) defined ten LOAs for an automated decision aid, which has subsequently been applied to various automated systems. The 10-point scale establishes that when automation is low (level 1), the human will manually perform tasks; when automation is at its highest (level 10), the system performs the tasks entirely (see Table 1.1) (Kaber & Endsley, 2004; Sheridan & Verplanck, 1978). The highest level in their model describes a clear distinction of the difference between automation and autonomy, which is when the system executes the function without the human operator
(de Visser et al., 2018). An example of autonomy is an elevator that functions (i.e., moving floors) without help from a user (apart from the selection of the floor).

Table 1.1 Sheridan & Verplank – Ten LOA for human-machine decision making.

<table>
<thead>
<tr>
<th>Degree</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human does the whole job up to the point of turning it over to the computer to implement</td>
</tr>
<tr>
<td>2</td>
<td>Computer helps by determining the options</td>
</tr>
<tr>
<td>3</td>
<td>Computer helps determine options and suggests one, which humans need not follow</td>
</tr>
<tr>
<td>4</td>
<td>Computer selects action and human may or may not do it</td>
</tr>
<tr>
<td>5</td>
<td>Computer selects action and implements it if human approves</td>
</tr>
<tr>
<td>6</td>
<td>Computer selects action, informs human in plenty of time to stop it</td>
</tr>
<tr>
<td>7</td>
<td>Computer does whole job and necessarily tells human what it did</td>
</tr>
<tr>
<td>8</td>
<td>Computer does whole job and tells human what it did only if human explicitly asks</td>
</tr>
<tr>
<td>9</td>
<td>Computer does whole job and tells human what it did and it, the computer, decides he should be told</td>
</tr>
<tr>
<td>10</td>
<td>Computer does whole job if it decides it should be done, and if so tells human, if it decides he should be told</td>
</tr>
</tbody>
</table>

1.3 Stages of Automation

As shown in Table 1.1, these LOAs can take place across four distinct stages. Derived from the human information processing model, these four stages are information acquisition, information analysis, decision selection, and action implementation, which can be automated (Parasuraman et al., 2000). Information acquisition refers to the process of decoding data to be processed and then documented (e.g., vehicles have cameras that pick up the raw data; Parasuraman et al., 2000). Information analysis is the processing of information and inferential details (e.g., when the system makes sense of the raw data like image detection and distances; Parasuraman et al., 2000). Decision selection occurs when decisions are selected based on the processed information (e.g., the vehicle will decide to brake or the decision to change lanes). Finally, action implementation occurs when the selected decision is implemented (e.g., when the system
brakes or changes lanes). Within the human information processing model, the four stages can individually be automated independently of each other. For example, a traditional dishwasher will not have information acquisition, information processing, or decision selection automation but will be highly automated in the action implementation. In this case, the dishwasher is carrying out a preselected action.

1.4 Degrees of Automation

The stages can be combined with the LOAs for any given system to create the framework for degrees of automation (DOA), defined as the level and capability of automation at a given moment (Wickens et al., 2010). Higher DOAs are represented by higher LOAs and later stages, such as decision selection and action implementation (Manzey et al., 2012; Wickens et al., 2010). This is indicated by the dotted arrow in Figure 1.1. For example, a vehicle that takes over if the user drifts lanes or that brakes in time for pedestrians/other vehicles would have a higher DOA than a system that simply alerts drivers to a vehicle in their blind spot.

<table>
<thead>
<tr>
<th>Levels of Automation</th>
<th>Stages of Automation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Information Acquisition</td>
</tr>
<tr>
<td></td>
<td>Information Analysis</td>
</tr>
<tr>
<td></td>
<td>Decision Selection</td>
</tr>
<tr>
<td></td>
<td>Action Implementation</td>
</tr>
</tbody>
</table>

Figure 1.1 Degrees of automation (Adapted from Wickens et al., 2010).
1.5 Reliance

Humans are increasingly reliant on automation to provide a range of benefits. Reliance is when users use a machine or system to perform tasks (de Visser et al., 2020). This reliance falls along a continuum that can be represented by under-reliance (which is to say less use than may be warranted by system performance) and over-reliance (which is to say using the system for more than it is capable of or failure to provide sufficient supervisory control). Individuals use automation if viewed as dependable, while rejecting automation that users view as undependable (Dzindolet et al., 2003). For example, a coffee machine that has scheduling capabilities to brew but often commits errors will likely decrease the usage (i.e., under-reliance) of the coffee maker. Reliance on automation often relates to the user’s trust in the system's capabilities (Muir & Moray, 1996). For example, Muir and Moray (1996) examined operators’ trust in automation by having users participate in a simulation of a milk pasteurizing plant with pumps that are automated or manually controlled. Reliance was evaluated by users' increased/decreased use (or interaction) of the automation. Results indicated a high positive correlation between operator trust and the use of automation (Muir & Moray, 1996), indicating that when users trusted automation, they used automated pumps (i.e., over-reliance), and when users distrusted automation, they controlled the pumps manually (i.e., under-reliance).

1.6 Trust

Trust is “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability” (J. D. Lee & See, 2004, pg 54).
Trust in automation includes three major categories: dispositional, situational, and learned trust (Hoff & Bashir, 2015; Marsh et al., 2003). Dispositional trust reflects when an individual has the predisposition to trust or distrust automation (Hoff & Bashir, 2015; Merritt & Ilgen, 2008). These individual differences in dispositional trust include factors like culture (Chien et al., 2016; Niedober et al., 2014), age (Pak et al., 2012), gender (Gallimore et al., 2019; Haselhuhn et al., 2015), and personality (Chien et al., 2017).

Situational trust depends on both external (i.e., type of system, system complexity, task difficulty, workload, perceived risks/benefits, organizational setting, framing of task; Bailey & Scerbo, 2007; Molloy & Parasuraman, 1996) and internal variables (i.e., self-confidence, subject matter expertise, mood, and attentional capacity; Akgun et al., 2010; de Vries et al., 2003; Tzeng, 2004).

The category of trust associated with the users' experience with a system differs from the dispositional and situational trust categorization. This experience results in learned trust, which includes two categories: initial and dynamic learned. Trust gained by previous interactions with the automated system, or ones like it, is defined as initially learned trust (Hoff & Bashir, 2015). Factors influencing the pre-existing knowledge that affects initial trust include attitudes/expectations, reputation, experience, and understanding of the system (Hoff & Bashir, 2015; Merritt et al., 2013). Once a user begins interacting with a system, a new type of trust arises, known as dynamically learned trust. Dynamically learned trust is gained from ongoing interactions with an automated system. These dynamic factors which influence trust include performance (i.e., reliability, validity, predictability, dependability, usefulness, the timing of error(s), the severity of error(s), type of error(s)), and design (i.e., appearance, ease-of-use,
communication style, transparency/feedback, and level of control) of the system (de Visser et al., 2016; de Visser et al., 2012; Weinstock et al., 2012).

When designing systems, engineers should promote calibrated trust rather than simply focusing on elevating trust (de Visser et al., 2020; Merritt, Lee et al., 2015; Pop et al., 2015). Trustworthiness is defined as the ability to execute tasks, show integrity, and be benevolent toward others (Mayer et al., 1995). Trust calibration is the process of aligning the perceived and actual trustworthiness of a system so that there is appropriate reliance on the system (de Visser et al., 2020). Miscalibrated trust can result in either over-reliance on a faulty system or under-reliance on a system that would aid performance and safety. These miscalibrations can be a function of individual differences (e.g., personality traits such as extraversion) which result in the user perceiving the machine’s actual abilities (i.e., competence, responsibility, predictability, and dependability) differently (Merritt & Ilgen, 2008), differences in initially learned trust (Dzindolet et al., 2002; Merritt, Unnerstall et al., 2015), or even the result of system performance during use (Alarcon et al., 2021; Kohn et al., 2018). Miscalibrated trust can fall into one of two categories: under- and over-trust.

Under-trust refers to circumstances in which the user’s perception of the trustworthiness of the system is lower than the machine’s actual capabilities (de Visser et al., 2020). Under-trusting the machine leads to disuse and under-reliance (e.g., ignoring or disabling alarms) of the system (de Visser et al., 2020; J. D. Lee & See, 2004; Parasuraman & Riley, 1997). Users that under-trust systems often silence alarms due to the high rate of false alarms (Dixon et al., 2007). Under-trusting automated systems can produce consequences for users in that they disuse systems that were designed to increase
safety and performance (Wiczorek & Meyer, 2019). While under-trusting aids led to users relying on themselves (perhaps to their detriment), over-trusting aids evoked the opposite (also perhaps to their detriment).

Over-trust refers to the user’s trust in the system being higher than the machine’s actual capabilities warrant (de Visser et al., 2020). When the user over-trusts the machine, it can lead to users becoming complacent in the monitoring of the system, or becoming over-reliant on the system (e.g., using the system in situations where the system should not be used). Users insufficiently monitoring the system/machine performance can result in detrimental consequences, such as a pilot utilizing autopilot but not monitoring it properly (Reichenbach et al., 2010). Individuals who experience complacency when operating automation are often provoked by biases towards automation (Bahner, Elepfandt et al., 2008; Bahner, Hüper et al., 2008; Mosier et al., 1998; Parasuraman et al., 1993). For example, the perfect automation schema is the expectation that individuals have about machines performing at perfect rates (Lyons & Guznov, 2019; Merritt, Unnerstall et al., 2015). When individuals expect perfection, they tend to over-trust the machine’s capabilities.

1.7 Situational Awareness

Over-trusting automation and the resultant complacency can lead to the degradation of SA (Chen et al., 2014). SA is the users’ mental model of the dynamic state of the environment. SA is frequently discussed as having three levels: perception of the environment (level 1), comprehension of content (level 2), and projection of future conditions (level 3; Endsley, 1995). SA is necessary for high levels of performance and safety in both complex and simple environments. SA can be incredibly important in a
variety of occupational environments and for a variety of job types: aviation (Jones & Endsley, 1996), industrial facilities (i.e., power plants), police officers, firefighters, and military personnel (Endsley, 1995). For example, general SA requirements for aviation include geographical (i.e., location of own aircraft, other aircraft), spatial/temporal (i.e., altitude, heading, velocity), system (i.e., system status, functioning, and settings), environmental (i.e., weather formations, temperature), and tactical SA (i.e., identification, location, threats) (Endsley, 1997). SA is imperative even in non-occupational environments since degradations can lead to insufficient task performance, detection or interpretation of surroundings, and the safety of others (Endsley & Kiris, 1995; Kaber & Endsley, 1998). Multiple factors can facilitate the loss of SA, such as diverted attention, as well as task-related and task-unrelated mind wandering (Casner & Schooler, 2014; Kidwell et al., 2012; Kilingaru et al., 2013).

Becoming knowledgeable of the factors that can facilitate a loss of SA can help users be more aware of the consequences associated with higher DOAs. One such consequence relates to what is known as the lumberjack analogy (Onnasch et al., 2014; Sebok & Wickens, 2017). The analogy stems from the tree-felling notion that “the higher the trees, the harder they fall.” When machines have higher DOAs, an automation failure is more consequential than when the machines have lower DOAs. When an automated system provides increased support, users become more complacent and experience degraded SA (Onnasch et al., 2014). This complacency and degraded SA translate into significantly worse human performance following a system failure. Another consequence of using imperfect automation is a ‘black swan’ event, which is a rare failure of automation that is unique and unanticipated and, subsequently, results in a user either
missing a system failure or failing to respond appropriately (Sebok & Wickens, 2017; Taleb, 2007; Wickens et al., 2015). The rare occurrences of black swan events paired with a decrease in SA may result in catastrophic performance in identifying and responding to failures (as was seen in the 3 Mile Island incident; NRC, 1979).

One consequence of a loss of SA that is a primary concern to designers of automated systems is the ability to successfully transfer control from the user to the system or from the system to the user (Eriksson & Stanton, 2017). This transfer of control can be both human-initiated and system-initiated for both transfers to or from the system. An example of human-initiated transfer to a system would be the engagement of adaptive cruise control, while the human-initiated transfer from the system would be the disengagement of that adaptive cruise control. System-initiated transfer to the system would be emergency braking in which the vehicle overrides the input from the user and applies the brakes because of a detected safety concern. System-initiated transfer to the human would be the disengagement of the self-driving mode because of an inability to navigate a given scenario (but not having been requested from the human), as is seen in Teslas, for example (Gillmore & Tenhundfeld, 2020).

While companies have implemented solutions to promote greater SA and performance, there has not been much research evaluating the consequences of repeated monitoring checks on trust, especially as a function of DOA. This is important to understand because the consequences of having a system tell you to remain engaged may unknowingly affect the user’s trust. These constant attention checks could result in users trusting the system less since it (repeatedly) reveals to the user the imperfections of the system. Previous literature has suggested that regular false alarms can result in a “cry-
wolf” effect, resulting in under-trust of a system (Madhavan et al., 2006; Wickens, Hooey et al., 2009). However, it is unclear to what degree such monitoring checks, such as those incorporated into Tesla, result in something analogous to the cry-wolf effect. On the one hand, perhaps these attention checks could help to calibrate trust and result in increases in SA, appropriate monitoring behavior, and performance (both before and after a failure/transfer of control). This would mean that there was appropriate reliance on a system, thus increasing safety and overall performance in the way that the automation was designed. On the other hand, perhaps these attention checks result in under-trust and subsequent disuse of the system. It is critically important to understand the ramifications of these types of attention checks on human interaction with automated systems to best promote design practices that enhance safety, performance, and satisfaction of use with these systems (J. D. Lee et al., 2017).

As such, this study will aim to evaluate the use of monitoring checks when users are tasked with performing a primary and secondary task. The literature detailed herein has led to the following research questions (see Table 1.2).
Table 1.2 Research questions (RQ).

<table>
<thead>
<tr>
<th>Number</th>
<th>Research Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ.1</td>
<td>Does reliability affect trust?</td>
</tr>
<tr>
<td>RQ.2</td>
<td>Does the frequency of a monitoring check affect trust?</td>
</tr>
<tr>
<td>RQ.3</td>
<td>Does reliability affect performance?</td>
</tr>
<tr>
<td>RQ.4</td>
<td>Does the frequency of a monitoring check affect performance?</td>
</tr>
<tr>
<td>RQ.5</td>
<td>How do carryover effects impact subjective, behavioral, and eye-tracking metrics?</td>
</tr>
</tbody>
</table>
| RQ.6   | a. What is the relationship between complacency potential and users’ preference to interact with a system?  
b. What is the relationship between complacency potential and users’ likelihood of disabling monitoring checks? |
| RQ.7   | Does the frequency of monitoring checks or reliability impact users’ preference to disable monitoring checks? |
Chapter 2. Method

2.1 Participants

A convenience sample of participants (N = 134) was selected from the available participant pool of undergraduate Psychology students at The University of Alabama in Huntsville. Students received course credit for participating in this study. The sample size (N = 175) was determined by running an a priori analysis with a medium effect size, using G*power (Faul et al., 2007). The number of participants in this sample was not sufficient for the design.

The average age of participants was 20 years old (SD = 3.092), with 61% female, 34% male, 3% non-binary, and 2% preferred not to share that information. The ethnicity of the sample was 72% White, 11% Black/African American, 8% Other, 6% Asian. Four participants preferred not to share their ethnicity. Before collecting data for this study, we received approval from the Institutional Review Board at the University of Alabama in Huntsville (See Appendix A.1).

2.2 Design

It was intended for the current study to utilize a 2 (Block: one, two) x 3 (System Reliability: 90%, 70%, 50%) x 4 (Frequency of Monitoring Checks: no checks, 30, 60, and 90 seconds) mixed-factor design. Between subject factors included System Reliability and Frequency of Monitoring Checks with Block being within subjects. However, the reliability of the system performed differently than originally intended and
is treated (in the results and discussion) as a continuous variable because of limitations discovered when analyzing the data.

2.3 Apparatus

This study utilized an automated simulation built with the Tkinter library in Python. A Tobii pro fusion eye-tracker was set to sample at 250 Hz to measure metrics such as dwell time and fixation counts of specific areas of interest (AOIs). Three AOIs were specified around the dials in the primary task, the phone number entry pad in the secondary task, and the monitoring check alert (all of which are explained in the 2.4 section below).

2.4 Tasks

2.4.1 Primary Task

The simulation environment of the primary task (see Figure 2.1) included four circular dials with an arrow and a specified range (i.e., light grey region) inside the dial. The arrows within the dials spin, and when they land within the specified range, the automation will count them and increment the count on the screen. The arrows in the dials can spin in both directions and do so at a speed less than or equal to 24 degrees per second (the speed was variable for each frame update). The direction varied randomly within each dial. Below the dials is a text box with the count and a button with a plus or minus one symbol for participants to adjust errors (i.e., commission and omission errors). Commission errors are when the automation counts when it should not have, and an error
of omission is when the automation fails to count an instance of the arrow in the specified region.

### 2.4.2 Secondary Task

The secondary task (Figure 2.1) was a telephone number entry task modeled after the Defined Intensity Stressor Simulation (Wetherell & Sidgreaves, 2005). The secondary task presented a random phone number that participants relayed using the mouse to click on the corresponding and correct numbers. Every time they hit “Enter,” a new number appeared after a 10-second delay.

![Figure 2.1 Desktop view seen by the participants. The monitoring check is black when first presented and turns blue if not dismissed by the participant.](image)

### 2.4.3 Monitoring Check

The monitoring check shown in Figure 2.2 was displayed above the dials and presented the message (*i.e.*, “Press the spacebar”), which was accompanied by an auditory “beep” simultaneously to participants. Both occurred at the selected frequency of checks (*i.e.*, no checks, 30, 60, or 90 seconds). The border and message (*i.e.*, “Press the
spacebar") of the monitoring check flashed blue if participants did not dismiss the alert after five seconds. If participants did not dismiss the alert once the border flashed, the alert would “beep” after five seconds and remained flashing blue until it was dismissed. If participants did not dismiss the alert after the border flashed and the next alert was queued, the alert disappeared, and the next monitoring check was presented. Participants were able to dismiss the monitoring check by pressing the spacebar.

![Press the Spacebar](image)

**Figure 2.2** Monitoring check seen by the participants.

### 2.5 Instruments

A Qualtrics survey was used to administer the questionnaires to participants. The Automation-Induced Complacency Potential Rating (AICP-R) questionnaire consists of 10 questions assessing the participants’ overall attitudes towards automation with two subscales of alleviating workload (i.e., when I have a lot to do, it makes sense to delegate a task to automation) and monitoring (i.e., Even if an automated aid can help me with a task, I should pay attention to its performance) (Table 2.1) (Merritt *et al.*, 2019). The 10 questions utilized a 5-point Likert scale (1 “strongly disagree” to 5 “strongly agree”). The AICP-R was administered once before the simulation. The AICP-R has sufficiently good test-retest reliability for alleviating workload (0.84) and monitoring (0.77), which allows for individual differences comparisons (Merritt *et al.*, 2019). The Trust of Automated Systems Test (TOAST) consists of nine questions that were used to assess the participants' overall trust in the understanding and performance of the system automation.
(Table 2.2) (Wojton et al., 2020). The nine questions are divided into system understanding (i.e., “I understand what the system should do”) and performance (i.e., “the system performs consistently”) (Wojton et al., 2020). The questions are graded on a 7-point Likert scale (1 “strongly disagree” to 7 “strongly agree”). The TOAST was administered after each block. TOAST indicates high criterion validity of trust between the two factors: system understanding and performance (Wojton et al., 2020).

### Table 2.1 The Automation-Induced Complacency Potential Rating (AICP-R) (Merritt et al., 2019). 5-point Likert scale, 1 (strongly disagree) and 5 (strongly agree). Reverse coded *

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. When I have a lot to do, it makes sense to delegate a task to automation.</td>
</tr>
<tr>
<td>2. If life were busy, I would let an automated system handle some tasks for me.</td>
</tr>
<tr>
<td>3. Automation should be used to ease people’s workload.</td>
</tr>
<tr>
<td>4. If automation is available to help me with something, it makes sense for me to pay even more attention to my other tasks.</td>
</tr>
<tr>
<td>5. Even if an automated aid can help me with a task, I should pay attention to its performance. *</td>
</tr>
<tr>
<td>6. Distractions and interruptions are less of a problem for me when I have an automated system to cover some of the work.</td>
</tr>
<tr>
<td>7. Constantly monitoring an automated system’s performance is a waste of time.</td>
</tr>
<tr>
<td>8. Even when I have a lot to do, I am likely to watch automation carefully for errors. *</td>
</tr>
<tr>
<td>9. It’s not usually necessary to pay much attention to automation when it is running.</td>
</tr>
<tr>
<td>10. Carefully watching automation takes time away from more important or interesting things.</td>
</tr>
</tbody>
</table>

### Table 2.2 The Trust of Automated Systems Test (TOAST) (Wojiton et al., 2020). 7-point Likert scale, 1 (strongly disagree) and 7 (strongly agree).

#### System Understanding

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I understand what the system should do.</td>
</tr>
<tr>
<td>2. I understand the limitations of the system.</td>
</tr>
<tr>
<td>3. I understand the capabilities of the system.</td>
</tr>
<tr>
<td>4. I understand how the system executes tasks.</td>
</tr>
</tbody>
</table>

#### System Performance

<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The system helps me achieve my goals.</td>
</tr>
<tr>
<td>2. The system performs consistently.</td>
</tr>
<tr>
<td>3. The system performs the way it should.</td>
</tr>
<tr>
<td>4. I am rarely surprised by how the system responds.</td>
</tr>
<tr>
<td>5. I feel comfortable relying on the information provided by the system.</td>
</tr>
</tbody>
</table>
2.6 Measures

2.6.1 Subjective Measures

Subjective trust was measured by the TOAST questionnaire (Wojton et al., 2020) and complacency potential was measured by AICP-R questionnaire (Merritt et al., 2019).

2.6.2 Behavioral Measures

Behavioral measures included participants’ performance and reaction time to monitoring checks. Participants’ performance was measured by calculating the total deviation of the final count from what the count should have been and dividing that by the errors combined to standardize the raw errors. Participants’ performance was further broken down into errors of omission and commission. Detection of omission errors was measured by calculating the deviations of commission/omission errors and corrections implemented by participants and dividing that by the errors to standardize them. Detection of commission errors was measured by calculating the deviations of commission/omission errors and corrections implemented by participants and dividing that by the errors to standardize them.

2.6.3 Eye-tracking Measures

Objective trust was measured by eye-tracking metrics, including proportion of fixation counts, dwell time, and task switching within the areas of interest (AOIs). Fixation counts were calculated by collecting the number of fixations within each AOI and dividing it by the total number of fixations during the experiment. Dwell time was calculated by getting a running total and each time participants moved into an AOI, the
time was marked. When participants left the AOI, the time was subtracted from the time they entered the AOI, and it was added to the running total. Task switching was calculated by the number of times a participant switched from one AOI to another. Instances where participants moved from an AOI to anything other than another AOI were not included in the analysis.

2.7 Procedure

Participants were asked to present identification to ensure that they were at least 18 years of age. They were asked to read an informed voluntary consent form and sign the form if they agreed to participate. Participants were informed that the study included three questionnaires and a simulation environment which would be shown twice. The experiment took approximately 30 minutes. After participants were briefed on the experimental procedures, they were encouraged to ask any questions about the process. They were instructed to fill out the AICP-R questionnaire on Qualtrics. Following, the experimenter performed the eye-tracking calibration for the Tobii pro fusion (bar attached to computer screen) before being shown the instructions and beginning the simulation (see Figure 2.3). The participants were instructed to make their primary focus the accuracy of the count for the dial task, and a secondary focus should be the phone number task. During the simulation, participants interacted with the system, and he/she monitored the dials, which the automation would count when the arrows landed in the specified range. The participant monitored the dials and established whether an error occurred. During the simulation, errors occurred at random intervals and were programmed to be distributed between errors of omission and commission. An error of commission was to be corrected by the participant clicking -1. An error of omission was
to be corrected by the participant clicking +1. While monitoring the dials, participants were presented with a phone number that reset following a short interval (i.e., 10 seconds) after they submitted each full phone number. During the simulation, participants were presented with the monitoring check visually (i.e., “Press the spacebar”) and auditorily (i.e., “beep”) that participants could dismiss by pressing the spacebar. The check was visual/auditory to mimic the real-world application. The frequency of this check was determined by their condition (i.e., either no check or a check every 30, 60, or 90 seconds). The monitoring check was presented at the frequency regardless of whether the participant correctly dismissed the check. There were two blocks of trials in the simulation. The monitoring check and reliability (as a continuous variable) condition was presented to participants throughout both blocks. After completion of the first and second blocks of the simulation, participants were instructed to fill out the TOAST questionnaire on Qualtrics. Lastly, they were debriefed about the purpose of the study and thanked for their participation before being dismissed.
You will now be shown a simulation with two tasks. Your primary focus are the dials and automated count, while the secondary focus is the phone number task. Your job is to monitor the automated dials and when the arrow enters the range (per dial) the count adjusts. The count should not add one if the arrow is not in the greyed out range. Conversely, it should count each time the arrow does enter the greyed out range. If errors occur use the -1 and +1 buttons to correct, in order to ensure the running count is an accurate reflection of the number of times that the arrow entered the greyed out range.

Your job for the phone number task is to relay the phone number when shown by using the buttons. After selecting enter, a new phone number will be shown after 10 seconds. Remember that the phone number task is your secondary focus.

You will potentially be presented an alert informing you to "press the spacebar". To make the alert go away you can press the spacebar on keyboard or click the alert.

If you have any questions, ask the researcher for clarification before moving on.

---

**Figure 2.3** Instructions seen by participants before starting the simulation.
Chapter 3. Results

Statistical analyses were conducted in Rstudio using functions from the dyplyr (Wickham et al., 2023), tidyverse (Wickham et al., 2019), emmeans (Lenth, 2022), effects (Fox & Weisberg, 2019), ggplot2 (Wickham, 2016), and base packages (R Core Team, 2022).

Before analyzing the eye-tracking data, AOIs were created in Tobii to evaluate metrics as they relate to the primary task, secondary task, and monitoring check, as shown in Figure 3.1. Four participants were omitted from the analyses because of errors in the data collection process, with 130 remaining participants to include in analyses.

![Figure 3.1 Visualization of AOI placement.](image)
3.1 Manipulation Checks

It was intended for reliability conditions to fit within three groups: 50%, 70%, and 90%. However, when performing a manipulation check, it was discovered that the program was not coded correctly, and as such, reliability covered a much larger spectrum than intended. As such, and for the remainder of the analyses, reliability was included as a continuous average time to error variable rather than a categorical variable (Figure 3.2).

A paired samples t-test indicated that the average seconds between errors were not significantly different from block one (\(M = 22.079, SD = 9.475, \text{Min} = 10.169 \text{ s}, \text{Max} = 60 \text{ s}\)) to block two (\(M = 20.515, SD = 4.610, \text{Min} = 13.636 \text{ s}, \text{Max} = 40 \text{ s}\)) for the average seconds between errors, \(t(123) = 1.592, p = .114, d = .143\). There was greater variability in block one than in block two. Additionally, analyses used the average time between monitoring checks for the regressions throughout, unless otherwise specified. This allowed for linear regressions to be run with the data. Those in the “no check” condition were given a value of 600 which represented the amount of time for each block.
3.2 Subjective Assessments

3.2.1 Trust (B1)

A linear regression with the average time between monitoring checks and the average seconds between errors as the predictors, on trust following block one was performed to better understand whether individuals' trust was affected by the frequency
of monitoring checks and reliability (as quantified by average seconds between errors) for blocks one/two. The average reported trust (1 strongly disagree to 7 strongly agree) in the system for block one was 4.236 ($SD = 0.934$). The regression was a significant predictor of trust, $F (2, 120) = 7.245, p < .001, R^2 = .109, R^2_{adj} = .094$ (see Table 3.1). The time to error and time between monitoring checks were significant predictors of trust. Figure 3.3 illustrates the linear regression with predictors.

**Table 3.1** The time between errors (B1) and checks predicting individuals’ trust.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>$SE$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.902</td>
<td>0.222</td>
<td>17.795</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors</td>
<td>0.229</td>
<td>0.009</td>
<td>2.637</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Average time between checks</td>
<td>-0.240</td>
<td>0.0003</td>
<td>-2.765</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>

**Figure 3.3** The time between errors (B1) and checks predicting individuals’ trust. Error ribbons represent 95% confidence intervals.
3.2.2 Trust (B2)

A linear regression with the average time between monitoring checks (including no check condition) for block two along with the average seconds between errors for both blocks as the predictors was run on trust following block two. The analysis included the average seconds between errors for block one so carryover effects could be analyzed. The average reported trust in the system on TOAST for block two was 4.280 ($SD = 0.878$). The regression was a significant predictor of trust, $F (3, 113) = 4.093$, $p = .009$, $R^2 = .100$, $R^2_{adj} = .076$ (Table 3.2). While the average time between monitoring checks was a significant predictor of trust, the time between errors for blocks one and two were not significant predictors (Table 3.2). The average time to error for block two was not a significant predictor of trust (Figure 3.4). Figure 3.4 illustrates the linear regression for block one.

Table 3.2 The time between errors (B1 & B2) and checks predicting individuals’ trust.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>$SE$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.683</td>
<td>0.401</td>
<td>9.181</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors (B1)</td>
<td>0.082</td>
<td>0.009</td>
<td>0.902</td>
<td>0.369</td>
</tr>
<tr>
<td>Average time between errors (B2)</td>
<td>0.155</td>
<td>0.017</td>
<td>1.704</td>
<td>0.091</td>
</tr>
<tr>
<td>Average time between checks (B2)</td>
<td>-0.265</td>
<td>0.003</td>
<td>-2.924</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>
Figure 3.4 The time between errors (B1 & B2) and checks predicting individuals’ trust. Error ribbons represent 95% confidence intervals.

3.2.3 Complacency Potential (AICP-R)

Overall, individuals had an average AICP-R score of 3.047 ($SD = 0.445$) (out of 5), with an average of 3.609 ($SD = 0.569$) for alleviating workload subscale (delegating tasks to automation) and a 2.485 ($SD = 0.577$) for monitoring subscale (need to monitor
Individuals’ with higher alleviating workload subscale scores should have greater propensity to delegate tasks to automation.

### 3.2.4 Correlations: Subjective Assessments

The correlation between complacency potential (AICP-R) and trust indicated no correlation for block one or block two. Since the AICP-R has subscales for alleviating workload and monitoring, correlations were analyzed between subscales and trust. The correlation between alleviating workload and trust for block one suggested a positive correlation, $r(125) = .254$, $p = .004$, and no correlation for block two (Figure 3.5). There was no correlation found between the monitoring subscale and trust for block one or block two.

![Figure 3.5 Scatterplot of alleviating workload and trust. Error ribbons represent 95% confidence intervals.](image-url)
3.3 Behavioral Assessments

3.3.1 Performance (B1)

A linear regression was performed with the average time between monitoring checks and the average seconds between errors as the predictors, on performance following block one. Individuals’ performance (measured by accuracy) average was 0.538 ($SD = 0.741$) for block one (perfect performance was 1.0). The regression was not a significant predictor of performance, $F (2, 120) = 2.496, p = .087, R^2 = .041, R^2_{adj} = .024$.

3.3.2 Performance (B2)

A linear regression was performed with the average time between monitoring checks for block two along with the average seconds between errors for both blocks as the predictors, on performance following block two. The average reported performance for block two was 0.466 ($SD = 0.969$). The regression was not a significant predictor of performance, $F (3, 115) = .544, p = .653, R^2 = .014, R^2_{adj} = -.012$.

3.3.3 Performance: Omission Errors (B1)

Additional linear regressions were analyzed to investigate whether the different types of errors (commission or omission) were impacted by the average time between errors and the average time between monitoring checks. Overall, across both blocks’ the simulation had more omission errors than commission errors. Omission errors occurred on average ($M = 0.473, SD = 0.439$) for block one. The regression for block one was not
a significant predictor of performance for omission errors, $F(2, 120) = .664$, $p = .517$, $R^2 = .011$, $R^2_{adj} = -.005$.

### 3.3.4 Performance: Omission Errors (B2)

A linear regression was run with the average time between monitoring checks for block two and the average seconds between errors for both blocks as the predictors, on performance regarding omission errors following block two. The average reported performance regarding omission errors for block two was $0.502 \ (SD = 1.032)$. The regression was not a significant predictor of performance for omission errors, $F(3, 115) = .352$, $p = .788$, $R^2 = .009$, $R^2_{adj} = -.017$.

### 3.3.5 Performance: Commission Errors (B1)

Commission errors occurred an average of $1.396 \ (SD = 3.845)$ for block one. The regression for block one was a significant predictor of performance for commission errors, $F(2, 120) = 4.917$, $p = .009$, $R^2 = .077$, $R^2_{adj} = .061$. The time to error was a significant predictor of performance regarding commission errors (Table 3.3). The time to monitoring check was not a significant predictor of performance regarding commission errors for block one (Table 3.3). Figure 3.6 illustrates the linear regression to demonstrate more clearly the performance regarding commission errors.
Table 3.3 The time between errors (B1) and checks predicting individuals’ performance on commission errors.

<table>
<thead>
<tr>
<th></th>
<th>( \beta )</th>
<th>( SE )</th>
<th>( t )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.336</td>
<td>0.965</td>
<td>-1.384</td>
<td>0.169</td>
</tr>
<tr>
<td>Average time between errors</td>
<td>0.266</td>
<td>0.039</td>
<td>3.004</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between checks</td>
<td>0.077</td>
<td>0.001</td>
<td>0.876</td>
<td>0.383</td>
</tr>
</tbody>
</table>

Figure 3.6 The time between errors (B1) and checks predicting individuals’ performance on commission errors. Error ribbons represent 95% confidence intervals.

3.3.6 Performance: Commission Errors (B2)

A linear regression was run with the average time between monitoring checks for block two along with the average seconds between errors for both blocks as the predictors, on performance regarding commission errors following block two. The average reported performance regarding commission errors for block two was 0.834 (SD
The regression was not a significant predictor of performance regarding commission errors, \( F(3, 115) = .231, p = .875, R^2 = .006, R^2_{adj} = -.020. \)

### 3.3.7 Performance: Additional Analyses

One-sample \( t \)-tests were conducted to evaluate the participants’ performance of over- or under-counting errors of commission and omission (See Appendix B.1). The one-sample \( t \)-test indicated that individuals under-counted omission errors for block one, \( t(128) = -24.202, p < .001 \), and block two, \( t(128) = -11.574, p < .001 \) (See Appendix B.2). Additionally, a one-sample \( t \)-test indicated that individuals over-counted commission errors for block two, \( t(128) = -5.013, p < .001 \), but block one was not significant, \( t(128) = -0.964, p = .337 \) (See Appendix B.2).

Variance ratio tests were conducted to evaluate whether there were differences in variances of the omission and commission errors. The variance ratio test to compare commission and omission errors for block one indicated a statistically significant difference between variances, \( F(128, 128) = 46.770, p < .001 \), and block two, \( F(128, 128) = 2.943, p < .001 \). Additionally, the variance ratio test to compare errors for block one/two indicated a statistically significant difference between variances for commission errors, \( F(128, 128) = 4.291, p < .001 \), and omission errors for blocks, \( F(128, 128) = .270, p < .001 \).

Paired sample \( t \)-tests were conducted to evaluate the participants’ accurate count (what the count should have been at the end of the block) and the final count (what the count was at the end of the block). The accurate count for block one \( (M = 64.471, SD = 10.462) \) to block two \( (M = 62.843, SD = 9.889) \) while the final count for block one \( (M = 57.386, SD = 16.170) \) to block two \( (M = 57.339, SD = 19.934) \). The paired sample \( t \)-
tests to compare accurate and final count indicated a significant difference between counts in block one, \( t(120) = 4.729, p < .001, M_{\text{diff}} = 6.537 \), and block two, \( t(120) = 2.738, p = .007, M_{\text{diff}} = 4.901 \) (See Appendix B.3).

### 3.3.8 Reaction Time to Monitoring Checks (B1)

Individuals’ reaction time to monitoring checks was 2.942 seconds (SD = 4.224) for block one. There was no statistically significant regression, \( F(2, 80) = 1.384, p = .257, R^2 = .034, R^2_{\text{adj}} = .010 \).

### 3.3.9 Reaction Time to Monitoring Checks (B2)

Next, a linear regression was run with the average time between monitoring checks for block two along with the average seconds between errors for both blocks as the predictors, on reaction time to monitoring checks following block two. Individuals’ reaction time to monitoring checks was 3.626 seconds (SD = 6.835) for block two. The regression was not a significant predictor of reaction time to checks, \( F(3, 77) = .007, p = .999, R^2 = .007, R^2_{\text{adj}} = -.040 \).

### 3.3.10 Correlations: Subjective and Behavioral Assessments

Previous research has investigated the relationship between complacency potential and performance/error detection (Bahner, Hüper, et al., 2008; Merritt et al., 2013). As such, correlations were analyzed for block one and block two. Since multiple comparisons were analyzed, a Bonferroni correction indicated the \( p \)-value needs to be less than .001 to be statistically significant. Pearson’s correlations were run for both blocks to investigate complacency potential and behavioral assessments (participants’
performance, participants’ performance of omission/commission, and reaction time) which resulted in no significance. Additionally, correlations were also run between behavioral assessments (participants’ performance, participants’ performance of omission/commission, and reaction time) and trust (as measured by TOAST) and resulted in no significance for blocks.

3.4 Eye-tracking Assessments

3.4.1 Proportion of Fixations within the Primary Task AOI (B1)

A linear regression was used to understand whether individuals’ proportion of fixations within the primary task AOI (See Figure 3.1) was affected by the frequency of monitoring checks and average seconds between errors for blocks one/two. To address this relationship, a series of linear regressions were run with the average time between monitoring checks and the average seconds between errors as the predictors, on proportion of fixation counts within the primary AOI following blocks. Individuals’ primary task proportion of fixations was 0.639 ($SD = 0.139$) for block one. The regression was not a significant predictor of the primary task proportion of fixation counts, $F (2, 111) = 1.921, p = .151, R^2 = .034, R^2_{adj} = .016$.

3.4.2 Proportion of Fixations within the Primary Task AOI (B2)

Next, a linear regression was run with the average time between monitoring checks for block two and the average time between errors for both blocks as the predictors, on the primary task proportion of fixations in block two. Individuals’ primary task proportion of fixations was 0.667 ($SD = 0.146$) for block two. The regression was
not a significant predictor of primary task proportion fixations, \( F (3, 106) = 2.509, p = 0.063, R^2 = 0.068, R^2_{adj} = 0.041 \) (Table 3.4). The average time between monitoring checks was not a significant predictor, however, the average time to error for block one was a significant predictor (Table 3.4). This significant predictor indicates that carryover effects may have occurred across blocks suggesting that participants’ fixations might have been partially informed by the reliability of the system in block one. The average time to error for block two was not a significant predictor of individuals' primary task proportion of fixation counts. Figure 3.7 illustrates the linear regression for block two.

<table>
<thead>
<tr>
<th></th>
<th>( \beta )</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>8.218e-01</td>
<td>6.766e-02</td>
<td>12.146</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors (B1)</td>
<td>-2.157e-01</td>
<td>1.444e-03</td>
<td>-2.258</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Average time between errors (B2)</td>
<td>-1.284e-01</td>
<td>2.929e-03</td>
<td>-1.343</td>
<td>0.182</td>
</tr>
<tr>
<td>Average time between checks (B2)</td>
<td>9.498e-03</td>
<td>5.486e-05</td>
<td>0.100</td>
<td>0.921</td>
</tr>
</tbody>
</table>

Table 3.4 The time between errors (B1 & B2) and checks predicting primary proportion fixation.
Figure 3.7 The time between errors (B1 & B2) and checks predicting primary proportion fixation. Error ribbons represent 95% confidence intervals.
3.4.3 Proportion of Fixations within the Secondary Task AOI (B1)

A linear regression was used to understand whether individuals’ proportion of fixations within the secondary task AOI (See Figure 3.1) was affected by the frequency of monitoring checks and average seconds between errors for blocks one/two. To address this relationship, a series of linear regressions were run with the average time between monitoring checks and the average seconds between errors as the predictors, on proportion of fixation counts within the secondary AOI following blocks. Individuals’ secondary task proportion of fixations counts was 0.263 (SD = 0.131) for block one. The regression was not a significant predictor of the secondary task proportion of fixation counts, \( F(2, 111) = 1.247, p = .292, R^2 = .022, R^2_{adj} = .004 \).

3.4.4 Proportion of Fixations within the Secondary Task AOI (B2)

Next, a linear regression was run with the average time between monitoring checks for block two along with the average seconds between errors for both blocks as the predictors, on secondary task proportion fixation counts following block two. Individuals’ secondary task proportion of fixation counts was 0.291 (SD = 0.142) for block two. The regression was not a significant predictor of secondary task proportion of fixation counts, \( F(3, 106) = 2.65, p = .053, R^2 = .072, R^2_{adj} = .045 \).

3.4.5 Proportion of Fixations within the Monitoring Check AOI (B1)

A linear regression was used to understand whether individuals’ proportion of fixations within the monitoring check AOI (See Figure 3.1) was affected by the frequency of monitoring checks and average seconds between errors for blocks one/two.
To address this relationship, a series of linear regressions were run with the average time between monitoring checks and the average seconds between errors as the predictors, on proportion of fixation counts within the monitoring check AOI following blocks.

Individuals’ monitoring check proportion of fixation counts was 0.005 (SD = 0.005) for block one. The regression for block one was a significant predictor of monitoring check proportion of fixation counts, $F(2, 76) = 9.079, p < .0001, R^2 = .197, R_{adj}^2 = .175$ (Table 3.5). The average time between checks (omitting condition with no checks) was a significant predictor of monitoring check proportion of fixation counts (Table 3.5). The time to error was not a significant predictor of monitoring check proportion of fixation counts. Figure 3.8 illustrates the linear regression with predictors.

**Table 3.5** The time between errors (B1) and checks predicting monitoring check proportion fixation.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.081e-02</td>
<td>1.312e-03</td>
<td>8.242</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors</td>
<td>-1.120e-02</td>
<td>4.742e-05</td>
<td>-0.106</td>
<td>0.916</td>
</tr>
<tr>
<td>Average time between checks</td>
<td>-4.419e-01</td>
<td>1.404e-05</td>
<td>-4.182</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Next, a linear regression was run with the average time between monitoring checks for block two along with the average seconds between errors for both blocks as the predictors, on monitoring check proportion fixation counts following block two. Individuals’ monitoring check proportion fixation counts was 0.006 ($SD = 0.009$) for block two. The regression was not a significant predictor of monitoring check proportion fixation counts, $F (3, 73) = 1.588, p = .199$, $R^2 = .064$, $R^2_{adj} = .024$. The average time between monitoring checks (omitting condition with no checks) was a significant predictor (Table 3.6). The average time to error for block one was not a significant predictor, indicating that no carryover effects occurred across blocks meaning that individuals treated the blocks separately (Table 3.6). The average time to error for block
two was not a significant predictor of individuals’ monitoring check proportion fixation counts. Figure 3.9 illustrates the linear regression for block two.

Table 3.6 The time between errors (B1 & B2) and checks predicting monitoring check proportion fixation.

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.626e-02</td>
<td>6.404e-03</td>
<td>2.539</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Average time between errors (B1)</td>
<td>2.408e-02</td>
<td>1.325e-04</td>
<td>0.204</td>
<td>0.839</td>
</tr>
<tr>
<td>Average time between errors (B2)</td>
<td>-4.934e-02</td>
<td>2.625e-04</td>
<td>-0.426</td>
<td>0.671</td>
</tr>
<tr>
<td>Average time between checks (B2)</td>
<td>-2.494e-01</td>
<td>4.177e-05</td>
<td>-2.113</td>
<td>0.04</td>
</tr>
</tbody>
</table>
3.4.7 Dwell Time within the Primary Task AOI (B1)

A linear regression was used to understand whether individuals’ dwell time within the primary task AOI (See Figure 3.1) was affected by the frequency of monitoring.
checks and average seconds between errors for blocks one/two. To address this relationship, a series of linear regressions were run with the average time between monitoring checks and the average seconds between errors as the predictors, on individuals’ dwell time within the primary task AOI following blocks. Individuals’ primary task dwell time was 381.187 seconds \((SD = 82.510)\) for block one. The regression was not a significant predictor of primary task dwell time, \(F(2, 111) = 2.173, p = .119, R^2 = .038, R^2_{adj} = .021\). The time to error was a significant predictor of primary task dwell time (Table 3.7). However, the average time between monitoring checks was not a significant predictor. Figure 3.10 illustrates the linear regression with predictors.

Table 3.7 The time between errors (B1) and checks predicting individuals’ primary task dwell time.

<table>
<thead>
<tr>
<th></th>
<th>(\beta)</th>
<th>SE</th>
<th>(t)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>420.784</td>
<td>20.826</td>
<td>20.205</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors</td>
<td>-0.195</td>
<td>0.842</td>
<td>-2.079</td>
<td>0.040</td>
</tr>
<tr>
<td>Average time between checks</td>
<td>-0.010</td>
<td>0.031</td>
<td>-0.102</td>
<td>0.919</td>
</tr>
</tbody>
</table>
3.4.8 Dwell Time within the Primary Task AOI (B2)

Next, a linear regression was run with the average time between monitoring checks for block two along with the average seconds between errors for both blocks as the predictors, on primary task dwell time following block two. Individuals’ primary task dwell time was 391.973 seconds ($SD = 93.501$) for block two. The regression was not a significant predictor of primary task dwell time, $F (3, 101) = .665, p = .576, R^2 = .020, R^2_{adj} = -.010$.

3.4.9 Dwell Time within the Secondary Task AOI (B1)

A linear regression was used to understand whether individuals’ dwell time within the secondary task AOI (See Figure 3.1) was affected by the frequency of monitoring checks and average seconds between errors for blocks one/two. To address this relationship, a series of linear regressions were run with the average time between
monitoring checks and the average seconds between errors as the predictors, on individuals’ dwell time within the secondary task AOI following blocks. Individuals’ secondary task dwell time was 149.170 seconds ($SD = 76.797$) for block one. The regression was not a significant predictor of secondary task dwell time, $F (2, 111) = 1.839, p = .164, R^2 = .033, R^2_{adj} = .015$.

3.4.10 Dwell Time within the Secondary Task AOI (B2)

Next, a linear regression was run with the average time between monitoring checks for block two along with the average seconds between errors for both blocks as the predictors, on secondary task dwell time following block two. Individuals’ secondary task dwell time was 163.936 seconds ($SD = 85.434$) for block two. The regression was not a significant predictor of secondary task dwell time, $F (3, 101) = 1.992, p = .120, R^2 = .057, R^2_{adj} = .029$.

3.4.11 Dwell Time within the Monitoring Check AOI (B1)

A linear regression was used to understand whether individuals’ dwell time within the monitoring check AOI (See Figure 3.1) was affected by the frequency of monitoring checks and average seconds between errors for blocks one/two. To address this relationship, a series of linear regressions were run with the average time between monitoring checks and the average seconds between errors as the predictors, on individuals’ dwell time within the monitoring check AOI following blocks. Individuals’ monitoring check dwell time was 4.305 seconds ($SD = 2.802$) for block one. The regression was a significant predictor of monitoring check dwell time, $F (2, 72) = 7.315, p = .001, R^2 = .173, R^2_{adj} = .149$. The time to error was not a significant predictor
of monitoring check dwell time (Table 3.8). The average time between monitoring checks (omitting condition with no checks) was a significant predictor (Table 3.8). Figure 3.11 illustrates the linear regression with predictors.

**Table 3.8** The time between errors (B1) and checks predicting individuals’ monitoring check dwell time.

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.837</td>
<td>0.946</td>
<td>7.231</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors</td>
<td>0.033</td>
<td>0.034</td>
<td>0.299</td>
<td>0.765</td>
</tr>
<tr>
<td>Average time between checks</td>
<td>-0.420</td>
<td>0.010</td>
<td>-3.810</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

**Figure 3.11** The time between errors (B1) and checks predicting individuals’ monitoring check dwell time. Error ribbons represent 95% confidence intervals.
3.4.12 Dwell Time within the Monitoring Check AOI (B2)

Next, a linear regression was run with the average time between monitoring checks (omitting condition with no checks) for block two along with the average seconds between errors for both blocks as the predictors, on monitoring check dwell time following block two. Individuals’ monitoring check dwell time was 4.605 seconds ($SD = 3.523$) for block two. The regression was not a significant predictor of monitoring check dwell time, $F(3, 67) = .982, p = .407, R^2 = .044, R^2_{adj} = -.001$.

3.4.13 Task Switching from Primary to Secondary AOI (B1)

A linear regression was used to understand whether individuals’ task switching from primary to secondary AOIs (See Figure 3.1) was affected by the frequency of monitoring checks and average seconds between errors for blocks one/two. To address this relationship, a series of linear regressions were run with the average time between monitoring checks and the average seconds between errors as the predictors, on task switching from primary to secondary AOIs following blocks. The average task switching (number of times switching) from primary to secondary AOIs was 80.537 ($SD = 46.528$) for block one. The regression was not a significant predictor of task switching from primary to secondary AOIs, $F(2, 114) = 1.185, p = .310, R^2 = .021, R^2_{adj} = .003$.

3.4.14 Task Switching from Primary to Secondary AOI (B2)

Next, a linear regression was run with the average time between monitoring checks for block two along with the average seconds between errors for both blocks as the predictors, on task switching of primary to secondary following block two.
Individuals' task switching from primary to secondary was 79.446 (SD = 46.353) for block two. The regression was a significant predictor of task switching from primary to secondary AOI, $F (3, 109) = 2.786, p = .044, R^2 = .073, R^2_{adj} = .047$. The average time between monitoring checks was not a significant predictor (Table 3.9). The average time to error for block one was not a significant predictor. The average time to error for block two was a significant predictor. Figure 3.12 illustrates the linear regression for block two.

**Table 3.9** The time between errors (B1 & B2) and checks predicting switching between primary to secondary.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>$SE$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>32.539</td>
<td>22.119</td>
<td>1.471</td>
<td>0.144</td>
</tr>
<tr>
<td>Average time between errors (B1)</td>
<td>0.001</td>
<td>0.471</td>
<td>0.013</td>
<td>0.990</td>
</tr>
<tr>
<td>Average time between errors (B2)</td>
<td>0.253</td>
<td>0.954</td>
<td>2.695</td>
<td>0.008</td>
</tr>
<tr>
<td>Average time between checks (B2)</td>
<td>-0.096</td>
<td>0.018</td>
<td>-1.030</td>
<td>0.306</td>
</tr>
</tbody>
</table>
3.4.15 Task Switching from Primary to Monitoring AOI (B1)

A linear regression was used to understand whether individuals’ task switching from primary to monitoring check AOIs (See Figure 3.1) was affected by the frequency
of monitoring checks and average seconds between errors for blocks one/two. To address this relationship, a series of linear regressions were run with the average time between monitoring checks and the average seconds between errors as the predictors, on task switching from primary to monitoring check AOIs following blocks. The average task switching from primary to monitoring AOIs was 4.645 ($SD = 4.685$) for block one. The regression was not a significant predictor of task switching from primary to monitoring AOIs, $F(2, 78) = 29.5, p < .0001, R^2 = .437, R^2_{adj} = .422$. The time to error was not a significant predictor of task switching from primary to monitoring. However, the average time between monitoring checks (omitting conditions with no checks) was a significant predictor (Table 3.10). Figure 3.13 illustrates the linear regression with predictors.

Table 3.10 The time between errors (B1) and checks predicting switching between primary to monitoring.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>13.239</td>
<td>1.215</td>
<td>10.898</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors</td>
<td>-0.024</td>
<td>0.044</td>
<td>-0.277</td>
<td>0.782</td>
</tr>
<tr>
<td>Average time between checks</td>
<td>-0.656</td>
<td>0.013</td>
<td>-7.494</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Figure 3.13 The time between errors (B1) and checks predicting switching between primary to monitoring. Error ribbons represent 95% confidence intervals.

3.4.16 Task Switching from Primary to Monitoring AOI (B2)

Next, a linear regression was run with the average time between monitoring checks (omitting conditions with no checks) for block two along with the average seconds between errors for both blocks as the predictors, on task switching of primary to monitoring following block two. Individuals’ task switching from primary to monitoring was 4.537 (SD = 4.738) for block two. The regression was a significant predictor of task switching of primary to monitoring AOIs, $F(3, 74) = 15.57, p < .0001$, $R^2 = .397$, $R^2_{adj} = .371$. The average time between monitoring checks was a significant predictor (Table 3.11). However, the average time to error for both blocks were not significant predictors. Figure 3.14 illustrates the linear regression for block two.
Table 3.11 The time between errors (B1 & B2) and checks predicting switching between primary to monitoring.

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>9.085</td>
<td>2.267</td>
<td>4.008</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors (B1)</td>
<td>0.073</td>
<td>0.047</td>
<td>0.773</td>
<td>0.442</td>
</tr>
<tr>
<td>Average time between errors (B2)</td>
<td>0.161</td>
<td>0.093</td>
<td>1.746</td>
<td>0.085</td>
</tr>
<tr>
<td>Average time between checks (B2)</td>
<td>-0.627</td>
<td>0.015</td>
<td>-6.643</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
Figure 3.14 The time between errors (B1 & B2) and checks predicting switching between primary to monitoring. Error ribbons represent 95% confidence intervals.

3.4.17 Task Switching from Secondary to Primary AOI (B1)

A linear regression was used to understand whether individuals’ task switching from secondary to primary AOIs (See Figure 3.1) was affected by the frequency of
monitoring checks and average seconds between errors for blocks one/two. To address this relationship, a series of linear regressions were run with the average time between monitoring checks and the average seconds between errors as the predictors, on task switching from secondary to primary AOIs following blocks. The average task switching from secondary to primary AOI was 80.322 (SD = 46.505) for block one. The regression was not a significant predictor of task switching from secondary to primary, \( F(2, 114) = 1.188, p = .309, R^2 = .021, R^2_{adj} = .003. \)

### 3.4.18 Task Switching from Secondary to Primary AOI (B2)

A linear regression was run with the average time between monitoring checks for block two along with the average seconds between errors for both blocks as the predictors, on task switching of secondary to primary following block two. Individuals’ task switching from secondary to primary was 78.942 (SD = 46.307) for block two. The regression was a significant predictor of task switching of secondary to primary, \( F(3, 109) = 2.811, p = .043, R^2 = .074, R^2_{adj} = .047. \) The average time between monitoring checks was not a significant predictor. However, the average time to error for block two was significant (Table 3.12). The average time to error for block one was not a significant predictor, indicating that no carryover effects occurred across blocks. Figure 3.15 illustrates the predictors from the analysis for block two.
Table 3.12 The time between errors (B1 & B2) and checks predicting switching between secondary to primary.

<table>
<thead>
<tr>
<th></th>
<th>β</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>31.699</td>
<td>22.075</td>
<td>1.436</td>
<td>0.153</td>
</tr>
<tr>
<td>Average time between errors (B1)</td>
<td>0.003</td>
<td>0.470</td>
<td>0.029</td>
<td>0.977</td>
</tr>
<tr>
<td>Average time between errors (B2)</td>
<td>0.254</td>
<td>0.952</td>
<td>2.709</td>
<td>0.008</td>
</tr>
<tr>
<td>Average time between checks (B2)</td>
<td>-0.096</td>
<td>0.018</td>
<td>-1.024</td>
<td>0.308</td>
</tr>
</tbody>
</table>
Figure 3.15 The time between errors (B1 & B2) and checks predicting switching between secondary to primary. Error ribbons represent 95% confidence intervals.

3.4.19 Task Switching from Monitoring to Primary AOI (B1)

A linear regression was used to understand whether individuals’ task switching from monitoring to primary AOIs (See Figure 3.1) was affected by the frequency of monitoring checks and average seconds between errors for blocks one/two. To address this relationship, a series of linear regressions were run with the average time between
monitoring checks and the average seconds between errors as the predictors, on task switching from monitoring to primary AOs following blocks. The average task switching from monitoring to primary AOI was 4.810 (SD = 4.836) for block one. The regression was a significant predictor of task switching from monitoring to primary, $F(2, 78) = 24.94, p < .0001, R^2 = .396, R^2_{adj} = .380$. The time to error was not a significant predictor of task switching from monitoring to primary. However, the average time between monitoring checks (omitting conditions with no checks) was a significant predictor (Table 3.13). Figure 3.16 illustrates the linear regression with predictors.

Table 3.13 The time between errors (B1) and checks predicting switching between monitoring to primary.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>13.285</td>
<td>1.290</td>
<td>10.297</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors</td>
<td>-0.019</td>
<td>0.047</td>
<td>-0.209</td>
<td>0.835</td>
</tr>
<tr>
<td>Average time between checks</td>
<td>-0.626</td>
<td>0.014</td>
<td>-6.900</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
3.4.20 Task Switching from Monitoring to Primary AOI (B2)

Next, a linear regression was run with the average time between monitoring checks (omitting condition with no checks) for block two along with the average seconds between errors for both blocks as the predictors, on task switching of monitoring to primary following block two. Individuals’ task switching from monitoring to primary was 4.851 ($SD = 4.926$) for block two. The regression was a significant predictor of task switching from monitoring to primary AOIs, $F (3, 101) = 16.24, p < .0001, R^2 = .407$, $R^2_{adj} = .382$ (Table 3.14). The average time between monitoring checks was a significant predictor. The average time to error for either block were not significant predictors. The average time to error for block one, was not a significant predictor, indicating that no carryover effects occurred across blocks. Figure 3.17 illustrates the linear regression for block two.
Table 3.14 The time between errors (B1 & B2) and checks predicting switching between monitoring to primary.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>$SE$</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>10.655</td>
<td>2.329</td>
<td>4.575</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Average time between errors (B1)</td>
<td>0.028</td>
<td>0.048</td>
<td>0.299</td>
<td>0.766</td>
</tr>
<tr>
<td>Average time between errors (B2)</td>
<td>0.136</td>
<td>0.096</td>
<td>1.482</td>
<td>0.143</td>
</tr>
<tr>
<td>Average time between checks (B2)</td>
<td>-0.635</td>
<td>0.015</td>
<td>-6.783</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>
The time between errors (B1 & B2) and checks predicting switching between monitoring to primary. Error ribbons represent 95% confidence intervals.

**Figure 3.17**

3.4.21 Block Comparisons

To compare blocks a series of paired sample *t*-tests was performed to decipher whether individuals treated blocks differently within eye-tracking metrics (Table 3.15).

The paired samples $t$-test to compare the proportion of fixations within the primary task indicated a significant difference between the proportion of fixations within the primary task in block one ($M = 0.639, SD = 0.139$) and block two ($M = 0.667, SD = 0.146$), $t(114) = -2.324, p = .022$ (Figure 3.18). The paired samples $t$-test to compare the proportion of fixations within the secondary task indicated a significant difference between the proportion of fixations within the secondary task in block one ($M = 0.263, SD = 0.131$) and block two ($M = 0.291, SD = 0.142$) conditions, $t(114) = -3.009, p = .003$ (Figure 3.18). The paired samples $t$-test performed to compare secondary task dwell time between blocks was statistically significant between the secondary task dwell time in block one ($M = 149.170, SD = 76.797$) and block two ($M = 163.936, SD = 85.434$), $t(109) = -2.104, p = .038$ (Figure 3.18).
Table 3.15 Block comparisons of paired samples \( t \)-tests.

<table>
<thead>
<tr>
<th></th>
<th>( t )</th>
<th>( df )</th>
<th>( d )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Fixations Proportions</td>
<td>-2.324</td>
<td>114</td>
<td>-0.217</td>
<td>0.022</td>
</tr>
<tr>
<td>Secondary Fixations Proportions</td>
<td>-3.009</td>
<td>114</td>
<td>-0.281</td>
<td>0.003</td>
</tr>
<tr>
<td>Monitoring Fixations Proportions</td>
<td>-1.541</td>
<td>113</td>
<td>-0.144</td>
<td>0.126</td>
</tr>
<tr>
<td>Primary Dwell Time</td>
<td>-1.159</td>
<td>109</td>
<td>-0.111</td>
<td>0.249</td>
</tr>
<tr>
<td>Secondary Dwell Time</td>
<td>-2.104</td>
<td>109</td>
<td>-0.201</td>
<td>0.038</td>
</tr>
<tr>
<td>Monitoring Dwell Time</td>
<td>-1.009</td>
<td>74</td>
<td>-0.117</td>
<td>0.316</td>
</tr>
<tr>
<td>Primary to Secondary</td>
<td>-0.157</td>
<td>118</td>
<td>-0.014</td>
<td>0.875</td>
</tr>
<tr>
<td>Primary to Monitoring</td>
<td>0.353</td>
<td>118</td>
<td>0.032</td>
<td>0.725</td>
</tr>
<tr>
<td>Secondary to Primary</td>
<td>-0.053</td>
<td>118</td>
<td>-0.005</td>
<td>0.958</td>
</tr>
<tr>
<td>Monitoring to Primary</td>
<td>-0.105</td>
<td>118</td>
<td>-0.010</td>
<td>0.917</td>
</tr>
</tbody>
</table>
Figure 3.18 Block comparisons. Error bars represent 95% confidence intervals.
3.4.22 Correlations: Subjective Assessments and Eye-tracking

To examine the relationship between complacency potential and preference to interact with a system, correlations for eye-tracking metrics and complacency potential were completed between various variables. These correlations included complacency potential scores and proportion fixation of primary/secondary AOI resulted in no significant relationships for blocks. Additionally, correlations were analyzed between eye-tracking metrics and trust. These correlations included TOAST scores and each AOI with proportion fixation counts and resulted in no significance for blocks.

3.4.23 Correlations: Eye-tracking and Time between Monitoring Checks

Correlations were analyzed between the time between monitoring checks (including the no-check condition) and ratio of primary/secondary proportion of fixation counts. Pearson’s correlations for both blocks investigated time between monitoring checks and ratio of primary/secondary proportion of fixation counts and resulted in no significant correlations.

3.5 Preference to Disable Monitoring Check

To evaluate individuals’ preference to disable monitoring checks, participants were asked about their preference to disable the monitoring check after each block. For block one, a logistic regression was performed to assess the relationship between the frequency of monitoring checks and the preference to disable. There was not a statistically significant difference between how frequently the monitoring checks appeared and whether participants indicated that they wanted to disable the monitoring
check. For block two, a logistic regression was performed to assess the relationship between frequency in monitoring checks and preference to disable. There was not a significant relationship between how frequently the monitoring checks appeared on the frequency with which participants indicated that they wanted to disable the monitoring check. Additionally, reliability and preference to disable were not statistically significant for block one or two.

The relationship between complacency potential and preference to disable correlations resulted in no significant relationships between individuals' AICP-R score and preference to disable for block one or two. In addition to the AICP-R, the alleviating workload subscale also indicated no significant relationships between individuals’ alleviating workload subscale and preference to disable for block one or two. The monitoring subscale also indicated no significant relationships between individuals’ monitoring subscale and preference to disable for block one or two.

3.6 Qualitative Analysis

3.6.1 Thematic Analysis Description

A thematic analysis is used to identify and contextualize themes within data (Braun & Clarke, 2006). Our thematic analysis was conducted utilizing an inductive approach, meaning that themes were identified when analyzing the data.

3.6.2 Thematic Analysis Procedure

Participants were asked their preference to disable the monitoring check, specifically, whether they would have disable or turn off the “press the spacebar” check if
given the option. Those that said they wanted to were asked why they would disable the check for both blocks. A thematic analysis was coded into four categories: (1) distracting, (2) annoying, (3) other, and (4) no comment. One of the most common reasons individuals provided for wanting to disable the monitoring check was that it was distracting from the other tasks. Additionally, participants indicated that the alert was a disruption of the other tasks. The theme/categorization for the responses after block one can be found in Table 3.16. Table 3.17 provides the responses of participants after block two. Conversely, when participants were asked their preference to disable the monitoring check those that said they would not disable, were also asked why they would not disable the check for both blocks. This thematic analysis was coded into five additional categories: (1) sound helpful, (2) not bothered, (3) beneficial, (4) other, and (5) no comment. One of the most common reasons individuals provided for not wanting to disable the monitoring check after blocks were that they were not bothered by the monitoring check. The theme/categorization for the responses for block one can be found in Table 3.18. Table 3.19 provides the participants' responses after block two.
Table 3.16 Representative statements from participants explaining their preference to disable after block one.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Statement</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distracting</td>
<td><em>It was distracting.</em></td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td><em>Distracted me from typing in the phone numbers and checking on the dials.</em></td>
<td></td>
</tr>
<tr>
<td>Annoying</td>
<td><em>It was unnecessary and annoying to hear.</em></td>
<td>21%</td>
</tr>
<tr>
<td></td>
<td><em>The sound was annoying.</em></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td><em>It was not that distracting and did not impede my work, but I still would rather it not be there.</em></td>
<td>12%</td>
</tr>
<tr>
<td>No Comment</td>
<td></td>
<td>21%</td>
</tr>
</tbody>
</table>

Table 3.17 Representative statements from participants explaining their preference to disable after block two.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Statement</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distracting</td>
<td><em>It was distracting.</em></td>
<td>56%</td>
</tr>
<tr>
<td></td>
<td><em>Although the sound alerted me, I still lost focus.</em></td>
<td></td>
</tr>
<tr>
<td>Annoying</td>
<td><em>The sound was annoying.</em></td>
<td>17%</td>
</tr>
<tr>
<td></td>
<td><em>The space bar was very annoying.</em></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td><em>Gives me one less thing to do.</em></td>
<td>10%</td>
</tr>
<tr>
<td>No Comment</td>
<td></td>
<td>17%</td>
</tr>
</tbody>
</table>
Table 3.18 Representative statements from participants explaining their preference to not disable after block one.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Statement</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound Helpful</td>
<td><em>The alert was audible, so while I kept my eye on the numbers, I could use my other hand to press the spacebar.</em></td>
<td>22%</td>
</tr>
<tr>
<td></td>
<td><em>The noise cued me in, so it did not require my attention to be diverted.</em></td>
<td></td>
</tr>
<tr>
<td>Not Bothered</td>
<td><em>It didn't really have an impact because it was easy to just press the spacebar and continue the simulation.</em></td>
<td>43%</td>
</tr>
<tr>
<td></td>
<td><em>I was not bothered with the extra pop-up task.</em></td>
<td></td>
</tr>
<tr>
<td>Beneficial</td>
<td><em>It made me feel like I was actually keeping up. It was hard to focus on if there were mistakes made or not but pressing the space bar when needed made it seem like I was keeping up with it right.</em></td>
<td>14%</td>
</tr>
<tr>
<td></td>
<td><em>I don't know what purpose it serves.</em></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td><em>I feel that it had no effect on how well I performed the tasks.</em></td>
<td>13%</td>
</tr>
<tr>
<td>No Comment</td>
<td></td>
<td>8%</td>
</tr>
</tbody>
</table>
Table 3.19 Representative statements from participants explaining their preference to not disable after block two.

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Statement</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sound Helpful</td>
<td><em>It was a simple task, and I could perform it without looking away from the dials and I was alerted by sound and a visual which I could glance at.</em></td>
<td>23%</td>
</tr>
<tr>
<td>Not Bothered</td>
<td><em>The pressing the space bar did not really bug me in the task that I was doing.</em></td>
<td>51%</td>
</tr>
<tr>
<td>Beneficial</td>
<td><em>The alert helped me to stay focused on the primary task.</em></td>
<td>10%</td>
</tr>
<tr>
<td></td>
<td><em>Serves as a reminder to check the other task if I've focused too hard on one.</em></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td><em>Don’t know what it does</em></td>
<td>6%</td>
</tr>
<tr>
<td>No Comment</td>
<td></td>
<td>10%</td>
</tr>
</tbody>
</table>
Chapter 4. Discussion

As self-driving vehicles become increasingly popular, it is essential to understand whether there are consequences regarding the design integrations to recapture users’ attention when utilizing automated features. Previous research has found evidence to support the cry-wolf effect regarding users’ tendency to under-trust systems prone to false alarms (Madhavan et al., 2006; Wickens, Hooey, et al., 2009). This theory suggests that implementing monitoring solutions with higher rates of false alarms may lead users to trust the system less because of these imperfections revealed throughout the human-machine interaction. When considering the effects of reliability on automated systems, researchers have found evidence that imperfect automation influences trust and performance when reliability is above 70% (Wickens & Dixon, 2007; Rovira et al., 2007; Lee & See, 2004). Conversely, researchers have found evidence that suggests low levels of reliability yield some benefits (Wickens & Dixon, 2007); however, they are limited and suggest individuals transition to manual control rather than using under-reliable systems (Moray et al., 2000; Muir & Moray, 1996). Typically, reliability has been evaluated as a categorical variable (discrete thresholds such as 50%, 70%, or 90%) (Chavaillaz et al., 2016; de Visser & Parasuraman, 2011; Pop et al., 2015). Our study evaluates reliability as a continuous variable because of the limitations discovered when analyzing the data.

This study aimed to evaluate the effects of varying reliability and frequency of monitoring checks on trust and performance when interacting with an automated
system. A secondary purpose was to evaluate the degree to which users’ trust was a function of the adjustment of system reliability or whether it was an averaging over longer periods (i.e., do the participants calibrate to each block, or are they using the totality of the interaction to form their trust).

4.1 The Effect of Reliability on Trust

When evaluating how reliability differences affect trust, results were mixed when dealing with low levels of reliability. Some subjective results, including TOAST metrics, provided some support (RQ.1), while other physiological results, including eye-tracking metrics, seemed not to provide support (RQ.1). However, there was no consistency among these results. The subjective self-report results convey that reliability had limited effects on the participants’ trust in the system, depending on how the data were analyzed. Where significant effects were found, the results indicated that when individuals experienced a system with longer durations between errors, they tended to have higher trust than those with shorter durations between errors. This is in line with previous work (Hoff & Bashir, 2015), which has suggested that trust is dynamically learned when users interact with a system and can be influenced by the system’s performance and design.

The reliability of the system also had limited effects on the participants’ fixation counts and dwell time. This could mean that the low reliability influenced users such that they had to calibrate to the system's capabilities, impacting their trust in the system and thus causing them to check on the system constantly. This aligns with the notion that users’ proportion of fixations in the primary task can be viewed as adjusting to the system's reliability, resulting in them switching between tasks. Another possibility could be that users inaccurately gauged the reliability of the system, leading the measurements
to show no difference in participants’ fixation counts and dwell time. This is supported by no differences between the proportion of fixations in the primary task and users’ subjective trust. The results could also be explained by users experiencing a system with more errors than the average study (Chavaillaz et al., 2016; de Visser & Parasuraman, 2011) that investigates reliability and trust. This could made users to be more cautious when interacting with the system, leading to no differences in fixation patterns.

Historically, there have been inconsistent results in the eye-tracking literature about the measure being a useful way to measure trust. Our study additionally did not elicit trust differences in the subjective measures, like are found in these other studies (Foroughi et al., 2023; Gold et al., 2015; Hergeth et al., 2016; Victor et al., 2018; Walker et al., 2018).

The system’s reliability affected the participants’ task switching in block two between the primary and secondary tasks, which could be a function of users adapting to the environment during block one. Our data supports this idea as the proportion of fixations within the primary/secondary tasks was higher in block two, suggesting the user’s fixations were exploring the simulation environment in block one. This is supported by looking at the proportions of fixations within the primary/secondary task which indicate that fixations are outside of that AOI. This aligns with previous work (Warm et al., 2008) that found that sustained vigilance is difficult to maintain, resulting in decreased attention and the propensity to engage in secondary tasks, leading to further degradations of attention.

Companies strive to create automated systems, such as self-driving vehicles, to be highly reliable; however, the actual levels of reliability of automated systems have recently come under scrutiny by society, researchers, and governmental agencies like the
NHTSA. The reliability of a system has traditionally shown individuals’ trust to be influenced by the performance and design of systems. Our results were mixed to show that reliability influenced individuals’ trust and eye-tracking metrics when interacting with the system. However, our study utilized lower system reliabilities than others (Chavaillaz et al., 2016; de Visser & Parasuraman, 2011), which could have influenced our results. Overall, this needs to be investigated further to understand how reliability influences individuals’ trust when measuring trust in various ways.

4.2 The Effect of Frequency of a Monitoring Check on Trust

Results were again mixed concerning the impact of monitoring check frequency on trust when assessed with the TOAST. The data supported that the frequency of monitoring checks affects trust (RQ.2). Where significant effects were found, the results indicated that when individuals experienced a system with longer durations between checks, they tended to have lower subjective trust than those with shorter durations between checks. One potential explanation could be that the performance of the monitoring check was factored into the trust. Thus more interactions resulted in higher trust as they learned the performance of the monitoring checks. The data supports this with significant effects of the higher proportion of fixations in the monitoring check and greater task switching when attending to the monitoring check as there was an increase in frequency for the monitoring check. This explanation aligns with previous research suggesting that eye-tracking systems can predict constructs such as attention (de Winter et al., 2021; Diaz-Piedra et al., 2019; Hasanzadeh et al., 2017). Additionally, the monitoring checks could have distracted users from the primary/secondary task so that
they would not have noticed the failures; however, users’ performance did not become worse as expected if they were a distraction.

Eye-tracking results provided no support for the idea that monitoring checks affect trust (RQ.2). This null result could be explained by differences between our paradigm and previous eye-tracking literature on trust. Previous literature has primarily focused on movements from different points in a display. However, because the paradigm in this study provided other stimuli to attend to with the more frequent monitoring checks, the results do not provide evidence that monitoring checks affect trust. Additionally, the results do not indicate that users had decreases in fixations between the ratio of primary/secondary tasks, meaning that monitoring checks did not influence individuals’ trust.

Although there was no support that monitoring checks affect trust because of the design of the paradigm, the frequency of monitoring checks did affect the individuals’ fixations and dwell time within the monitoring check AOI (dwell time was inconsistent across blocks, meaning results were not consistent), and task switching within the primary/monitoring task. Individuals tended to have more fixations in the monitoring check AOI as the monitoring checks became more frequent due to the participant being cued or bringing attention to the simulation. This indicates that individuals are responding to the system because of the cue they are receiving, causing their attention to shift between tasks. The findings could be supported by the participants’ increased proportions of fixations within the monitoring checks when interacting with the system. Previous research has indicated that attention guidance or cueing can cause users to be less attentive toward other display areas than necessary (Yeh & Wickens, 2000).
As companies have engineered solutions such as eye-tracking systems and behavioral attention checks to assist users in paying attention, it has become necessary to evaluate and understand the consequences of these strategies. Our results were mixed, showing an incomplete picture of how the frequency of monitoring checks influenced individuals’ trust when interacting with the system. As one of the first studies to investigate these attention-keeping strategies, further research is needed to understand the consequences of their use.

4.3 The Effect of Reliability on Performance

The results of this study provided no support that system reliability affected individuals’ performance, detection of omission errors, or reaction time to monitoring checks; however, results did suggest that participants were worse at detecting commission errors when the system reliability was low (RQ.3). This null result could be explained by calibrated trust in the system such that participants allocated attention appropriately to maintain supervisory control over the system; thus even though the system made more errors in the low-reliability condition, participants detected them and corrected them accordingly. While it was thought that reliability would affect individuals’ reaction time to monitoring checks because participants' attention would be taken from the monitoring check to the primary/secondary task, the results provided limited support that system reliability affected eye-tracking metrics, which explains the null results. While results indicate a significant effect for reliability on the detection of commission errors, it is not entirely clear why this would have been the case. One possible explanation pertains to the law of small numbers, such that smaller numbers are more likely to have more extreme values, which could have resulted in more extreme values
(Tversky & Kahneman, 1971). Said another way, if one were to flip a coin, it is more likely that the coin would come up heads 100% of the time if the coin was flipped 1 time, than if it were flipped 50 times. In this study, there were inadvertently fewer commission errors than omission errors, which may have led to more extreme values. Our data supports this as the variances significantly differ between the omission and commission errors.

Performance is often investigated differently by researchers depending upon whether they are assessing human or system failures when utilizing an automated system. Traditionally, researchers have investigated paradigms with near-perfect automation, first failure effect, errors of commission, or errors of omission (Bowden et al., 2021; Foroughi et al., 2023; Johnson et al., 2004; Wickens et al., 2015). Our study assessed the human’s performance of noticing errors of omission/commission, which has not been researched extensively when systems are prone to both errors. Our results indicated that reliability did not affect individuals’ performance, detection of omission errors, or reaction time to monitoring checks but did influence users to be worse at detecting commission errors when the system reliability was low. As one of the first studies to investigate these errors at higher rates, further research is needed to build a complete understanding of how reliability alters how users’ respond to errors.

4.4 The Effect of Frequency of a Monitoring Check on Performance

The frequency of monitoring checks did not affect the participants’ performance, error detection, or reaction time to the monitoring checks for either block (RQ.4). The design of the monitoring check could have influenced the null results, as users were not provided with information behind the reasoning for the check. In another way,
participants were not told that the monitoring check’s purpose was to remind them to attend to the primary task, but the monitoring checks were designed to elicit a response for participants to press the spacebar when shown the monitoring check. This could have impacted the response towards the checks, making responses more automatic than initially intended. This is supported by the qualitative analysis of individuals’ decision not to disable monitoring checks in which participants said they were “cued” by the noise and that the reaction to press the spacebar was an automatic response as users interacted with the system.

Companies that engineer solutions to ensure the user is paying attention do not openly explain these solutions to end-users (and if they do, users potentially subvert these solutions). These behavioral attention checks have not been evaluated regarding their influence on individuals’ performance or overall efficiency. Our results indicated that the frequency of monitoring checks did not affect individuals’ performance. This could have been influenced by the design of the monitoring check, suggesting that future monitoring check design and the purpose of the check should be relayed to users and further investigated by researchers.

4.5 Carryover Effects

When evaluating whether users’ trust was a function of the adjustments of reliability over time, the results from the subjective and eye-tracking measures did not support that there were carryover effects (RQ.5). This suggests that participants ignored the performance in block one while establishing their trust, attention/task switching strategies, and monitoring according to the system performance during block two. Said another way, participants could treat each block as a new system instead of a continuation
of the first block. This would mean their trust in the system during block one would be irrelevant to their trust formation in block two (Tenhundfeld et al., 2021). This indicates that users are treating them as two different systems, or at least that their dynamically learned trust discounted earlier interactions and was primarily concerned with the most recent interactions to formulate trust. Overall, further research needs to be done to understand carryover effects when there are deviations in system performance to expand the existing theory of factors (i.e., situational, dispositional, and learned trust) that influence individuals' trust in automation (Hoff & Bashir, 2015). This would help expand upon the field’s understanding of the residual impacts on trust, as a function of the earlier performance.

4.6 Block Comparisons

The comparison between blocks indicated that individuals had greater fixations within the primary/secondary tasks and dwelling within the secondary tasks in block two compared to block one. One potential reason for the greater proportions of fixations and dwell time within the secondary task could be that participants' overall understanding of the secondary task was greater in block two than in block one. This may be because participants did not understand that the secondary task provided them with numbers continuously, rather than exclusively once at the beginning of the block, influencing the fixations in the primary task rather than the secondary task in blocks. Additionally, researchers noted that some participants stated that they noticed the secondary task had provided them with new information after some time had passed during the block. The increase in fixations within the primary/secondary task in block two could indicate that users were trying to understand the system more in block one (fixations were exploring
other aspects of the interface). Users adapted to the system, influencing their fixations to be higher block two.

4.7 Complacency Potential (AICP-R)

The results did not support the idea that individuals’ complacency potential (measured by AICP-R) would have an impact on their interactions with the system (RQ.6a). Additionally, the results suggested that users’ score on the AICP-R did not predict their preference to disable the monitoring checks (RQ.6b). Results suggested that neither scores on the AICP-R nor the monitoring subscale were related to their preference to disable the monitoring checks. However, this is not entirely surprising as these two measures are distinct, and the AICP-R does not measure factors that could contribute to the preference for monitoring checks.

4.8 Preference to Disable Monitoring Check

The results did not provide support for the prediction that one’s preference to disable monitoring checks was a function of system reliability or frequency of monitoring checks; however, the thematic analysis did provide valuable insights into the reasons behind this preference (RQ.7). A potential reason for the system’s reliability not influencing an individual’s preference to disable the checks was that the checks were independent of the reliability condition. This means that the monitoring checks’ design characteristics (auditory and visual) had nothing to do with the system’s reliability. While the results suggest that frequency of monitoring checks and reliability are not associated with users’ preference to disable, further research is necessary to identify factors that may predict a preference to disable.
The qualitative analysis suggests that participants who wanted to disable the monitoring checks indicated that the checks were distracting from the primary and secondary tasks. Conversely, participants who preferred to keep the monitoring checks indicated that the checks were not bothering them from the tasks. Future research is needed to understand better the individual differences that may be contributing to differences between users in preference to disable the monitoring check. This is a potentially fruitful line of research as companies implement these monitoring checks to promote greater SA and performance. While the consequences of repeated monitoring checks remain unknown, previous research indicates detrimental consequences to the user, given that repeated false alarms can lead to users disabling the alarms (i.e., the cry-wolf effect; Wickens et al., 2009). However, given the mixed results of the current study and the design limitations, more research is needed to establish whether or not there could be negative (or positive, based on some of the results here) effects of monitoring checks on trust.

4.9 Limitations and Future Research

As with all experiments, several limitations were present in the study that should be expanded upon by future researchers. The monitoring check used was designed to mimic the attentional check found in self-driving vehicles like those made by Tesla, which simply requires participants to activate torque sensors on the steering wheel to indicate attention. However, this approach may not be the best representation of the types of real-world monitoring checks that are being applied. As research expands on the consequences of monitoring checks, the design should be evaluated to align more closely with self-driving vehicles. In addition to the limitations of monitoring checks, our study
evaluated reliability as a continuous variable, and the overall system reliability was unintentionally very low, which could have influenced the overall study (i.e., floor effects). While it is recommended that future studies look at more of a continuum of reliability, the reliability should be more in line with actual commercially viable systems. Lastly, this study was designed to apply to the monitoring strategies companies implement in self-driving vehicles; future research should evaluate this in a real-world environment rather than the laboratory.
Chapter 5. Conclusion

The main purpose of this study was to help understand the benefits or consequences that occur when monitoring checks are implemented for use in the supervisory control over automation, as can be found in self-driving vehicles like those offered by Tesla. Additionally, this study evaluated this in the context of fluctuations in a system's reliability. Overall, our results provide mixed findings that warrant further exploration. Furthermore, our research findings suggest that monitoring solutions should be investigated further. One area of consistent findings in this study is that users are either treating each block as two distinct systems, or their dynamically learned trust is continuously updating and discounting previous information which influences users to establish their trust and attentional allocation strategies to the current system. This study is among the first to evaluate the consequences of monitoring checks on user trust, reliance strategies, and performance when teaming with an automated system. This study helps expand upon existing theory, but more research is needed in order to better understand the consequences of monitoring checks on trust in systems.
References


86


Appendix A. IRB Approval

Date: 26 July 2022

PI: Jenna Cotter
PI Department: Psychology
The University of Alabama in Huntsville

Dear Jenna,

The UAH Institutional Review Board of Human Subjects Committee has reviewed your proposal titled: *Monitoring Automated Tasks* and found it meets the necessary criteria for approval. Your proposal seems to be in compliance with these institutions Federal Wide Assurance (FWA) 00019998 and the DHHS Regulations for the Protection of Human Subjects (45 CFR 46). Please note that this approval is good for one year from the date on this letter. If data collection continues past this period, you are responsible for processing a renewal application a minimum of 60 days prior to the expiration date.

No changes are to be made to the approved protocol without prior review and approval from the UAH IRB. All changes (e.g., a change in procedure, number of subjects, personnel, study locations, new recruitment materials, study instruments, etc.) must be prospectively reviewed and approved by the IRB before they are implemented. You should report any unanticipated problems involving risks to the participants or others to the IRB Chair. If you have any questions regarding the IRB’s decision, please contact me.

Sincerely,

Ann L. Bianchi
IRB Chair
Associate Professor, College of Nursing

Figure A.1 IRB Approval
Appendix B. Additional Performance Analyses

Table B.1 One-sample $t$-tests were conducted to evaluate the participants’ performance of over- or under-counting errors of commission and omission.

<table>
<thead>
<tr>
<th></th>
<th>$t$</th>
<th>df</th>
<th>$d$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commission errors (B1)</td>
<td>-0.964</td>
<td>128</td>
<td>-0.085</td>
<td>0.337</td>
</tr>
<tr>
<td>Commission errors (B2)</td>
<td>-5.013</td>
<td>128</td>
<td>-0.441</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Omission errors (B1)</td>
<td>-24.202</td>
<td>128</td>
<td>-2.131</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Omission errors (B2)</td>
<td>-11.574</td>
<td>128</td>
<td>-1.019</td>
<td>&lt; .001</td>
</tr>
</tbody>
</table>
Appendix B. Additional Performance Analyses

Figure B.2 Participants’ performance of over- or under-counting errors of commission and omission.
Appendix B. Additional Performance Analyses

Figure B.3 Compare participants accurate and final count.