Navigating Imperfect Automation: Automation's Impact on Operator Dependence Behaviors, Response Strategies, and Adaptations

by

Patrik T. Schuler

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Industrial and Operations Engineering)
in the University of Michigan
2024

Doctoral Committee:

Associate Professor Jessie X. Yang, Chair
Professor Yili Liu
Professor Lionel Robert
Professor Nadine Sarter

Patrik T. Schuler

patriks@umich.edu

ORCID iD: 0000-0003-2082-4421

DEDICATION

to my family and friends.

ACKNOWLEDGEMENTS

This PhD journey has been an experience marked by both triumphs and challenges, and I am grateful for the support I've received along the way. It is important to acknowledge the collaborative efforts shaping this endeavor.

First, I would like to thank my advisor, Jessie. Thank you for guiding me on this research apprenticeship and for trusting in me. Your curiosity, kindness, and dedication to the craft inspire me.

I extend gratitude to my esteemed committee members: Dr. Nadine Sarter, Dr. Yili Liu, and Dr. Leonel Robert. Your support, insightful perspectives, and dedicated time have been instrumental to my progress.

There are countless student peers who helped me on this journey - we are all growing together. Special mention goes to my friend Taa. The times we have had together were very influential on me. I also want to express my thanks to Jin, Kam, Hannah, Yao, Hyesun, Doo Won, Shreyas, Jennifer, Yaunchen, Caleb, Jacqueline, and Drew. Our discussions, often with coffee or snacks, have been immensely helpful.

I am indebted to those who paved the way when I started the program — Kevin, Na, Yadrianna, Alex, KP, Julia, Daniel, Ruikun, and other students in IOE and the ICRL. The HFES student chapter deserves appreciation for helping in my professional development and creating connections with researchers in our field.

Gratitude extends to the supportive members of IOE who played a crucial role in my graduate studies — Marina, Valerie, Matt, Tina, Cornelius, Chris, Mint, Sheryl, Rod, Sherie, Paul, and others who have contributed to my academic journey.

I am thankful for the financial support provided by the NSF, Rackham Graduate School, GSI funding from the IOE department, GSRA funding from Jessie, and the remainder of my G.I. Bill.

Finally, a special acknowledgment goes to my ambitious, beautiful, and intelligent wife, Lauren. Your unwavering support has been a cornerstone during my journey and this milestone would be impossible without you. You make me a better person, and for that, I will always be grateful.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ACRONYMS

SDT Signal Detection Theory

AVI Alarm Validity Information

PPV Positive Predictive Value

NPV Negative Predictive Value

OSL Overall Success Likelihood

ABSTRACT

Automation has become an integral aspect of modern work environments, promising enhanced efficiency, safety, and accuracy across various domains. Despite this, automation is still imperfect, and the human operator is ultimately responsible for outcomes. Operators have been inappropriately using automation, which has resulted in documentation of various incidents and accidents. Researchers have extensively explored the influence of automation reliability on human dependence behaviors and collaborative performance in human-automation interactions. A limited body of research has explored the impact of automation errors on operator cross-checking behaviors or strategies. While existing trust research (i.e., attitudes toward automation) explores operator adaptations, there is a notable gap in the literature regarding how operators are adjusting their dependence behaviors and strategies. The projects in this dissertation contribute knowledge by (1) examining how operators' dependence behaviors (i.e., compliance and reliance rates) and cross-checking rates are affected by automation performance; (2) evaluating how operators' adapt their dependence behaviors, cross-checking rates, and response strategies to varying degrees of imperfect automation; (3) investigating a design intervention focusing on the incorporation of likelihood information, specifically, to compare the effects of predictive values with a frequency format, additionally, a baseline condition where no *a priori* information was examined. A meta-analysis was utilized to address aim 1, where we systematically extracted dependence and cross-checking behavior data. We found that the human operators not only varied their compliance and reliance behaviors to the automation, but also varied how often they used additional information to verify the automation's recommendation. Human operators' blind compliance ($\beta_1 = .74$) and reliance ($\beta_1 = .89$) rates increased as automation's Positive

Predictive Value (PPV) and Negative Predictive Value (NPV) increased. Alternatively, the operators were more likely to cross-check automation's recommendation when automation performed worse. Operators' cross-checking behaviors were marginally more sensitive ($p = 0.08$) to non-alarm errors ($\beta_1 = -.90$) than alarm errors ($\beta_1 = -.52$). To address aim 2, we utilized a dual-task laboratory experiment to evaluate how operators adapt their dependence behaviors, cross-checking rates, and response. Automation performance influenced dependence behaviors and response strategies. More specifically, operators adapted using trial-by-trial feedback during alarms and non-alarms; their behaviors and strategies were independently adapted to the automation's PPV and NPV. We introduced a novel optimal decision-making strategy that considers operator access to alarm validity information. In the experiment, adjustments in behaviors converged towards the theoretical optimal behavior. However, more research is needed to empirically validate the proposed optimal strategy. Aim 3 was addressed through a data reanalysis and a human-subject study. Results indicated that automation performance influenced operator dependence behaviors, response strategies, and adaptations. More specifically, operators used trial-by-trial feedback to adjust to alarms and non-alarms; their behaviors and strategies were independently adapted to the automation's PPV and NPV. Participants strategically changed their behaviors to improve performance; they accepted a loss in accuracy for to allocate more attentional resources to a compensatory tracking task. Participants with likelihood information made slightly faster behavioral adjustments than those without information. The findings of this dissertation enrich our understanding of how operators depend on, validate, or ignore automated systems during dual-task performance. The introduction of a theoretical optimal standard can serve as a benchmark, enabling operators to calibrate their dependence behaviors. Insights into how information affects changes to operator behaviors can facilitate an accelerated learning process for operators and support more effective solution implementation.

CHAPTER 1

Introduction

Automation has become an integral aspect of modern work environments, promising enhanced efficiency, safety, and accuracy across various domains. Research addressing how humans can work with automation extends far back (Rasmussen, 1983). A seminal paper by Parasuraman and Riley (1997) on the human-automation relationship has lead to a substantial body of research (Lee and See, 2004; Wickens and Dixon, 2007; Lee, 2008; Hancock et al., 2011, 2021; Schaefer et al., 2016; Guo et al., 2021; Hutchinson et al., 2023). Broadly speaking, automation can be defined as a system or device that partially or fully completes tasks that was previously completed by a human (Parasuraman and Riley, 1997). The application of automated systems can be found in a wide array of domains and applications, including unmanned aerial vehicles (UAV) (Hocraffer and Nam, 2017), aviation (Sarter and Woods, 1997; Wickens, 2002b), manufacturing (Elghoneimy and Gruver, 2012), healthcare (Loftus et al., 2020), and military (Wang et al., 2009).

Despite technological advances, automation is still imperfect, and can be unreliable when completing tasks that are accompanied by uncertainty. Often, the human operator is the responsible for decision-making and outcomes for human-automation collaborations. Situations can become disastrous when humans are not appropriately using imperfect automation; incidents and accidents associated with inappropriate automation use include the grounding of the cruise ship Royal Majesty (NTSB, 1995; Degani, 2001) and the fatal Tesla crash in Mountain View, California (NTSB, 2020; Laris, 2020).

Parasuraman and Riley (1997) defined two types of inappropriate automation use: disuse and misuse. Disuse refers to when the human discards or neglects the automation and misuse refers to an over-reliance on the automation. Another view of inappropriate automation use was discussed by Dzindolet et al. (2003), where comparisons between an operator's performance and their trust in the automation determine if use is considered appropriate. This view considers it inappropriate to distrust automation that is more accurate than manual operator accuracy (disuse) or to trust automation that is less accurate than the operator (misuse).

1.1 Single- and Dual-Task Scenarios

Research in human-automation interaction can be broadly categorized into two paradigms: single-task and dual-task. In the single-task paradigm, studies involve participants engaging in a single task, which may include activities involving system monitoring, decision-making, or operating specific equipment. In the dual-task paradigm, the human is tasked with simultaneously completing two tasks. Automation assists in various tasks, allowing the human operator to concurrently participate in a secondary task. The dual-task paradigm presents the advantage that, if the automation is assigned the responsibility for n specific tasks, the collaborative efforts of the human and automation can result in the successful completion of a total of n + 1 tasks, as demonstrated by Dixon et al. (2005).

However, the introduction of a secondary task has concerned researchers, specifically, with how automation use affects the human's cognitive resources (Rice, 2009). Monitoring an imperfect automated aid is an additional task demand, which increases the situational complexity placed on the human. As automation reduces the manual task workload from the operator, the human's role changes from being an active controller to a supervisor. This role change has been observed in various empirical dual-task studies. In numerous studies, participants take on the role of simulated operator and are assisted by imperfect automated aids in decision-making tasks (Du et al., 2020a; Sanchez et al., 2006; McBride et al., 2014). To impose cognitive demands, researchers assign participants an additional task, such as a word search (Chancey et al., 2013), a compass positioning task (Yeh, 2000), or a compensatory tracking task (Sanchez et al., 2014; Yang et al., 2017; Wickens and Colcombe, 2007). In some studies, the scoring system was manipulated by researchers to establish a primary task and a concurrent task (Dixon and Wickens, 2006; Gérard and Manzey, 2010; Du et al., 2020a), while other researchers required participants to complete two primary tasks simultaneously (Mayer et al., 2006; Sanchez et al., 2014). Analyzing the performance of individual tasks and overall dual-task provides insights into how participants prioritize attentional resources (Wiczorek and Manzey, 2014; Du et al., 2020a). Previous research has identified that dual tasks can compete for operator cognitive resources across information processing stages (perceptual vs. response) or sensory modalities (visual vs. auditory), leading to a bottleneck effect (Wickens, 2002a; Wickens et al., 2000). Design considerations can alleviate the operator's cognitive load (Sarter, 2006). When proper design considerations are implemented, and automation is appropriately used, the collaborative human-automation performance can extend beyond the capabilities of the operator or automation alone. This dissertation operates within the framework of the dual-task paradigm, with a specific focus on diagnostic automation.

1.2 Diagnostic Automation and Signal Detection Theory

Diagnostic automation, a form of automation that analyzes raw information and infers the status of the world, falls under stage two in the four-stage taxonomy of automation proposed by Parasuraman et al. (2000). In the first stage, the automation acquires information and has the ability to register data from multiple sources. The second stage (i.e., diagnostic) focuses on information analysis and involves the perception and processing of retrieved information. Automation in this stage diagnoses on the state of the world and generates a recommendation for the human, who is often considered an operator of the automation. In the third stage, automation makes the decision and selects an action; in the fourth, the automation implements action without requiring direct operator involvement. An operator using diagnostic automation is responsible for decision-making and performance outcomes; they consider the automation's recommendation, integrate the recommendation against their knowledge and experience, and make the final decision. The binary nature of diagnostic automation makes it useful when quantifying performance, which is often done through Signal Detection Theory (SDT).

SDT has been used methodologically and theoretically to quantify decision-making under uncertainty (Wickens et al., 2021; Tanner and Swets, 1954; Macmillan and Creelman, 2005). SDT applies to situations where an individual or automated aid, the 'detector', is presented with ambiguous information about the state of the world, which is often full of noisy information. The detector is then responsible to classify the event to either "signal present" or "signal absent". A "signal" denotes a relevant stimulus, while "noise" denotes background/environmental interference or irrelevant stimuli. SDT quantifies the how often a signal is *actually* present in the world, referred to as the base rate parameter. Hazard events (i.e., true signals) are typically low in the real world and increased for laboratory experiments (Parasuraman and Riley, 1997). SDT allows for the categorization of diagnostic automation's performance outcomes into four distinct states: correct detection of a signal (termed "hit"), failure to detect a signal (termed "miss"), incorrect perception of signals in the presence of noise (termed "false alarm" (FA)), and correct identification of signal absent (termed "correct rejection" (CR)).

Using SDT, we can calculate metrics of discrimination performance (Sorkin and Woods, 1985; Meyer and Bitan, 2002). The hit rate, $P(alert|signal)$, is the conditional probability of automation detecting the signal when there is in fact a signal in the world, and the CR rate, $P(no\ alert|no\ signal)$, the conditional probability of not issuing an alert when there is no signal in the world. The hit rate and CR rate are intrinsic characteristics of the detector, however, they do not alone accurately capture a detector's ability to discriminate signals and noise. Another measure of behavioral efficiency, which are closely related to the hit and CR rates, are predictive values (Getty et al., 1995; Meyer and Bitan, 2002). The positive predictive value (PPV), $P(signal|alert)$, is the conditional probability of having a true signal given an automation alert; the negative predictive value (NPV), $P(no\ signal|no\ alert)$, is the conditional probability of not having a signal given the automation is silent. The overall success likelihood (OSL), $P(success)$, is a weighted average of PPV and NPV, indicating percentage of time that automation is correct, regardless of automation's diagnosis. Mathematically speaking, hit rate and PPV, and CR rate and NPV, are inverse conditional probabilities. If the base rate (i.e., *a priori* information on the prevalence of signals) is known, PPV and NPV can be derived from the hit rate and CR rate using the Bayes' Theorem, and vice versa.

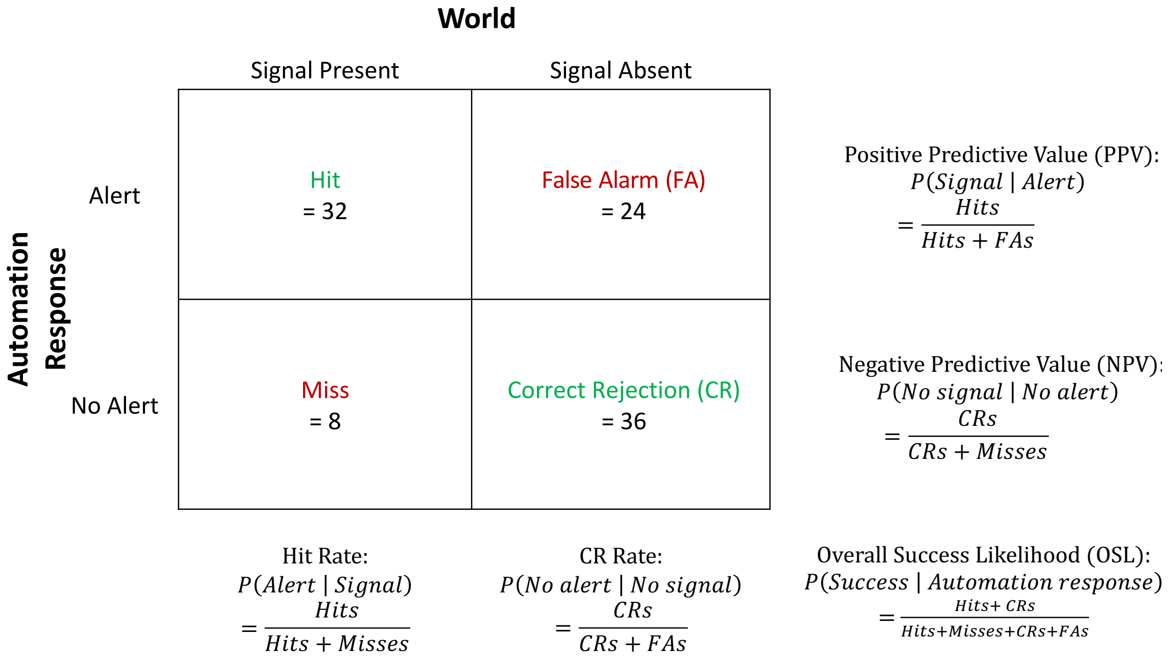

World
Signal Present
Signal Absent
Automation Response
Alert
No Alert
Hit = 32
False Alarm (FA) = 24
Miss = 8
Correct Rejection (CR) = 36
Positive Predictive Value (PPV):
$P(Signal \mid Alert) = \frac{Hits}{Hits + FAs}$
Negative Predictive Value (NPV):
$P(No\ signal \mid No\ alert) = \frac{CRs}{CRs + Misses}$
Hit Rate:
$P(Alert \mid Signal) = \frac{Hits}{Hits + Misses}$
CR Rate:
$P(No\ alert \mid No\ signal) = \frac{CRs}{CRs + FAs}$
Overall Success Likelihood (OSL):
$P(Success \mid Automation\ response) = \frac{Hits + CRs}{Hits + Misses + CRs + FAs}$

Figure 1.1: Signal detection theory formulae for hit rate, correct rejection rate, positive predictive value, negative predictive value, and overall success likelihood.

Using the numeric values shown in Figure 1.1, where the base rate is 0.4 (40/100), we can calculate the hit rate, CR rate, PPV, and NPV as follows:

$Hit\ rate = P(alert|signal) = \frac{Hits}{Hits\ +\ Misses} = \frac{32}{32+8} = 0.8$

$CR\ rate = P(no\ alert|no\ signal) = \frac{CRs}{FAs\ +\ CRs} = \frac{36}{36+24} = 0.6$

$PPV = P(signal|alert) = \frac{Hits}{Hits\ +\ FAs} = \frac{32}{32+24} = 0.57$

$NPV = P(no\ signal|no\ alert) = \frac{CRs}{Misses\ +\ CRs} = \frac{36}{36+8} = 0.82$

$OSL = P(success) = \frac{Hits\ +\ CRs}{Hits\ +\ Misses\ +\ FAs\ +\ CRs} = \frac{32+24}{32+8+24+36} = 0.68$

Other measures have been used to model detector properties, namely, response bias (*c*) and sensitivity (*d'*). The response bias is a threshold set to determine how likely a detector will classify an event as alert or silent. Response biases range from conservative (positive *c* values), where there is a tendency to classify the event as noise, to liberal (negative *c* values), where there is a tendency to classify as signal. A *c* value of 0 indicates an unbiased detector. Sensitivity (*d'*) refers to the detector's ability to discriminate between signal and noise events. A *d'* of 0 indicates a chance performance and a 5 would indicate near perfect performance. Sensitivity is determined by the reliability and quality of the algorithms used to process raw data.

1.3 Operator Dependence and Alarm Validity Information

When a human operator receives a recommendation from diagnostic automation, they must choose to follow or ignore that specific recommendation. Their behavioral response to the automation is referred to as dependence, as they can depend on the automated aid. Dependence behaviors are often separated into compliance and reliance, which refer to how an operator responds to automation alarms or non-alarms (Meyer, 2001, 2004). Compliance refers to the human operator's tendency to perform an action when the automation diagnoses a signal in the world; and reliance is the human operator's tendency to refrain from performing an action when the automation is silent, indicating "all is well". The notion of compliance and reliance as two distinctive manifestations of automation dependence have been supported by empirical studies (Dixon et al., 2007; Yang et al., 2017). Inappropriate use (i.e., over- or under- dependence) may result in the total human-automation performance to be less than possible achievable levels (Meyer, 2001; Robinson and Sorkin, 1985; Rice and McCarley, 2011). Numerous factors influencing operator dependence behaviors have been investigated, including trust (Yang et al., 2022), workload (McBride et al., 2011), task demands (Lin et al., 2020), and domain knowledge (Sanchez et al., 2014).

Challenges to making decisions under uncertainty depend on if human operator can access additional information regarding the validity of an alarm, referred to as Alarm Validity Information (AVI). AVI can be presented in multiple forms, such as with a photo, video feed, or manual in-person verification. Allendoerfer et al. (2008) describes how AVI use extends decision making to 2-stage process. The first stage is similar to situations where the operator can only follow or discard an alarm. The second stage is entered when the operator has access to AVI and considers whether to cross-check the automation's recommendation. The second stage is more cognitively effortful and delays operator response time. If the decision is made to cross-check alarm, more time time is required to manually validate the alarm in order to remove uncertainty. The probabilistic nature of detecting signal from noise make perfect performance impossible (Wickens, 2001). A human using automation is considered a type of decision-making under uncertainty task and the responsibility of the outcome is placed on the human operator (Sorkin and Woods, 1985; Meyer, 2004; Manzey et al., 2014; Yang et al., 2017; Wang and Yang, 2022).

1.4 Influential Factors in Human-Automation Interactions

While prior research has focused on a variety of factors influencing automation use (for early models, refer to:Riley (1996) or Dzindolet et al. (2001)), this dissertation centers on the effects of automation reliability, alarm validity information (AVI), and likelihood information on operator dependence behaviors and strategies. The following subsections each identify a research gap that this dissertation will address, followed by the listed aims in Section 1.5.

1.4.1 Related Research on Automation Error Type and AVI

1.4.1.1 Prior Meta-analysis Study on the Effect of Reliability on Performance

The benefits of a decision aid have been found to decline as the aid's reliability decreases (Rovira et al., 2007; Madhavan and Wiegmann, 2007; Skitka et al., 1999). Until the work by Wickens and Dixon (2007), it was unclear at what point unreliable automation was considered more beneficial or costly. To explore how automation reliability influences the net sum of performance costs/benefits, the authors conducted a literature synthesis to compare manual performance against the performance of the operator using diagnostic automation. They extracted 35 data points from 20 different studies and completed a regression analysis. The manual task performance was treated as a baseline. Automation reliability accounted for 41% of the variance in their model. As reliability was reduced, the total human-automation performance decreased. The cost/benefit analysis identified a reliability cross-over point of 0.7. Above this threshold, the benefits of automation outweighed the costs; below this point, the diagnostic accuracy was worse than if the human manually conducted the task. The authors also found that operators compensated for poor automation reliability to protect the concurrent task performance. The synthesis was insightful regarding the impact of imperfect automation's reliability on performance outcomes, however, there are further opportunities for in-depth exploration of automation reliability across alarm and non-alarm states.

1.4.1.2 Error Type

Within SDT, automation performance can be separated into PPV and NPV, which separately indicate how well the automation performs in alarms or non-alarms. Since the proposal of compliance and reliance as independent cognitive states that are associated with diagnostic automation (Meyer, 2001, 2004), there has been mixed discussion in the literature regarding the independence of the effects of different SDT error types (i.e., false alarms or misses). Some empirical findings support the notion of separate effects (Meyer, 2001; Dixon and Wickens, 2006; Wickens and Colcombe, 2007; Chancey et al., 2017). Conversely, alternative studies indicate that the effects of false alarms are not selective across the types of errors (Meyer et al., 2014; Dixon et al., 2007). Frameworks have been created to organize factors contributing how operators detect, understand, and correct errors (collectively referred to as error management) (McBride et al., 2014)

An early study investigating the independence of compliance and reliance was conducted by Dixon and Wickens (2006). Participants completed a simulated flight task through 10 missions and, as a secondary task, were required to monitor system gauges for potential system failures. There were 10 system failures (i.e., hits) and researchers added either 5 false alarms or 5 misses to introduce imperfections to automation's performance. Automation reliability was manipulated as a between-subjects variable, with four conditions total: baseline of no aid, perfect automation, miss prone (67% reliability), and false alarm prone (67% reliability). Their results indicated that the effects of error type were relatively independent of each other. Regarding performance, they found that lower automation clearly inhibited performance, but in different ways based on error type; the FA prone automation was negatively correlated with the monitoring task and miss prone automation was negatively correlated with concurrent tasks.

Dixon et al. (2007) conducted a study to further investigate trends of operator compliance and reliance across error types. Participants were separated into 4 experimental conditions: baseline of no automation, perfect automation, false alarm prone automation (60% reliability), and miss prone automation (60% reliability). The authors in this study isolated error type across condition, meaning that participants in the miss prone condition experienced no false alarms and vice versa. Results indicated that miss prone automation affected reliance behaviors, while false alarm prone automation affected compliance and reliance behaviors. Additionally, the authors noted how, in the FA prone condition, tracking task performance was not only worse than miss prone, but also worse than the baseline condition.

A study published around the same time compared effects of balanced errors (equal misses and false alarms) with effects of false alarm prone automation on dual-task performance (Wickens and Colcombe, 2007). This study consisted of 2 experiments with student pilots. Participants completed a tracking task while monitoring for air traffic conflicts, the monitoring task was supported with an imperfect automated aid. In the first experiment, the automation was 75% reliable with an equal number of false alarms and misses. In the second experiment, the researchers manipulated the automation's response bias (c became more liberal) to induce a 4:1 false alarm to miss ratio. Their results found that automation with a high proportion of false-alarms improved performance when compared to an equal number of false alarms and misses, this was shown through accuracy and concurrent task performance.

Also investigating the effects of error types, a study by Gérard and Manzey (2010) systematically manipulated the base rate to examine a wide range of imperfect automation. The authors explored effects of having access to alarm validity information and error type (false alarm and misses) onto operator dependence behaviors. Participants simulated a chemical plant operator in a dual-task scenario and the experimental session consisted of completing two blocks that lasted 800 seconds each. Automation reliability was treated as a between subjects variable and authors manipulated the experimental base rate to vary reliability, with PPV ranging from 0.10 to 0.90 and NPV ranging from 0.98 to 0.41. The authors examined the dependence (separated by alarm state) and cross-checking behaviors of participants. They found that participants were more sensitive to misses than false alarms, as there were strong behavioral shifts to a small difference in NPVs, which was not seen in alarm responses.

1.4.1.3 Alarm Validity Information

Additionally, the synthesis by Wickens and Dixon (2007) did not take into account the utilization of alarm validity information (AVI), likely because AVI was outside the defined objectives of their study. AVI allows the human operator to access additional information and cross-check an alarm against the ground truth. Limited research attention has been given to the use of AVI (Manzey et al., 2014; Bliss, 2003).

Manzey et al. (2014) examined the impact of AVI access through a series of 4 controlled laboratory experiments. In addition to AVI impact, the authors also examined the effort required to validate alarms and the workload on response rates, strategies, and dual task performance. Participants conducted a dual-task experiment simulating a chemical plant control room. The authors manipulated the experimental base rate to examine a variety of reliability levels, with PPV ranging from 0.10 to 0.90 and NPV ranging from 0.98 to 0.41. The sensitivity of the alarm (1.09) and response criterion (0.29) remained the same across reliability conditions. In the first experiment, the authors denied participants AVI access and found that low PPV lead to participants utilizing a strategy with a significant amount of alarm non-responses, referred to as the 'cry wolf' effect (coined from Breznitz (1984)). In the second experiment, they provided participants with AVI and found a dramatic reduction in the amount of alarms ignored (i.e., reduction in cry wolf effect). In addition, the authors found that AVI directly impacted operator response strategies and having AVI access increased operator sensitivity to non-alarms. Experiment three noted that the reduction in cry wolf effect still occurred when increasing efforts required to cross-check. When the authors manipulated the workload in the fourth experiment, they found that increasing the concurrent task workload resulted in participants ignoring alarms. While they examined how operators dependence behaviors and strategies across varying PPV and NPV conditions, they were limited by one set of alarm characteristics used, with one sensitivity level ($d' = 1.09$) and response bias ($c = -0.29$). The study was informative for the specific scenario and alarm characteristics, but there is an opportunity to expand by synthesizing data from multiple studies and scenarios.

While individual studies have explored the impact of automation performance on the dependence and/or cross-checking rates of human operators, and there exists literature synthesis that compares the costs and benefits of automation; there is not a literature synthesis to provide an overview of operator dependence and cross-checking behaviors across studies.

1.4.2 Related Research on Adaptive Operators and Strategies

1.4.2.1 Adaptive Operators

Although the notion of operators adapting to automation is not new (Riley, 1996), there is a limited body of research investigating the evolution of operator responses to automation, with few exceptions (Manzey et al., 2012; Sanchez et al., 2014).

A study by Manzey et al. (2012) contained two experiments that explored the impact of task distribution and operator experience on behavioral responses and performance. The first experiment focused on levels of automation and compared manual, diagnostic (they termed it IA support), and two higher levels of automation. Participants completed 5 blocks, which began and ended in manual tasks. The authors found that, when compared to a manual baseline condition, the primary and secondary task performance improved and the effects depended on the automation level. In their second experiment, they examined how automation bias developed over time by manipulating how long the operators worked with the automation before the automation eventually failed. The verification behaviors and performance consequences of the automation bias were quantified. Results indicated the development of operator response biases over time and the operators are influenced by experience with the decision aid. Automation errors had stronger effects than experiences with a more reliable aid.

In a related vein, Sanchez et al. (2014) conducted two laboratory experiments, one focusing on participant age and another on domain experience, simulating a multitask scenario in an agricultural vehicle simulator. Participants were responsible for a tracking task and a collision avoidance task. The session consisted of 240 events events that each lasted 15 seconds. A highly reliable (95.4%), but imperfect automated aid recommended on the status of possible collisions. Participants could follow the recommendation, ignore it, or manually validate the alarm with alarm validity information. The authors divided the experiment into halves and reported the proportion of participants that cross-checked for each of the 15 second events, illustrating the adaptations of operator dependence and cross-checking behaviors. They found that operators logically adapted their dependence behaviors in a manner that enhanced collaboration with automation, indicating that operators acquired patterns of behavior and adapted while striving for optimal performance.

There have also been recent studies examining how operators adapt to automation in single-task paradigms. While participants were not conducting a dual task scenario, the results are still informative to this dissertation, as they provide indications about how operators are adapting their automation usage. In a study by Hutchinson et al. (2022), the authors conducted two maritime simulation experiments examining how fluctuations of automation performance impacted the operator's perceptions of reliability. Participants were required to classify underwater vessels (with 6 options) and could do it manually before the automated aid would classify the vessel, then participants could accept or ignore the aid's recommendation. The experiment consisted of 20 trial blocks and the automation's reliability either was constant (75%) or varied across low (55%) and high (95%) conditions. Data regarding the participant's perception of automation reliability was compared to the actual reliability. Additionally, the overall response accuracy was examined. The authors found that participants were adapting their expectations and use of the automation as the reliability changed. In the constant reliability condition, participants increased their automation use (termed acceptance) as the experiment progressed. Regarding the changing reliability conditions, participants were adapting to the automation slightly differently based on reliability increases or decreases. If the automation started in a low reliability condition, automation use would slowly increase; whereas when starting with high reliability, the following reduction in automation use was greater. In a subsequent study by Hutchinson et al. (2023), the authors explored how quickly operators are adapting their automation use. Participants completed a maritime simulation task and were randomly assigned to low (60%) or high (90%) reliability conditions. The authors found that operators were adapting their automation use based on their experiences with the aid. Initially, participants tended to under-use the aid and were sluggish in response adjustments.

1.4.2.2 Operator Strategies

As logical automation operators repeatedly predict one of two alternatives, they often employ a strategy. In several decision-making models within the literature on SDT, the operator chooses a strategy in a single-task scenario. In such scenarios, the operator often has information about the automation's accuracy and can allocate cognitive resources to select a strategy suitable to them. The probability matching model was noted by Bartlett and McCarley (2017), where the operator defers to the automation's judgement with a probability that is equal to the accuracy of the automated aid. In the optimal weighting model, the operator averages their judgement with the automation's and weighs the average by the operator's sensitivity (Bahrami et al., 2010; Sorkin et al., 2001). The uniform weighting model is similar, but instead the human and automation's judgements are averaged without weights (Bahrami et al., 2010; Sorkin et al., 2001). The operator can randomly defer to the aid half the time, termed coin flip (Bahrami et al., 2010). Finally, operators have been found to use a strategy termed best decides, where they defer to the more accurate detector (human or aid) (Denkiewicz et al., 2013).

In a study examining how automation reliability affected the operator response strategy, Bartlett and McCarley (2021) required participants to make binary decisions for a signal detection task, with a manual baseline condition. In three conditions, participants would use automated aids with reliability ranging from 60% to 96%. The results indicated that while the high reliability automation improved discrimination sensitivity, participants were more efficient when using lower performing automation. The authors attributed this to the participants preference to disuse imperfect automation over use of highly reliable automation.

When making decisions and judgements under uncertainty, humans are known for rarely following purely logical considerations, instead they often employ heuristics (Tversky and

Kahneman, 1974; Gigerenzer and Todd, 1999). Heuristics are strategies that do not prioritize performance outcomes. Additionally, heuristics prove to be valuable shortcuts for offsetting a variety of factors, such as coping with uncertainty, addressing challenges in optimization, or when cognitive resources are exceeded by task demands (Gigerenzer, 2008; Gigerenzer et al., 2011). A heuristic has the benefits of minimizing operator cognitive load, improving reaction time, and often improving performance. One study attributed effort as an influential aspect of the strategy selection process, where the decision is made based on the combination of the anticipated accuracy of the alarm and the cognitive effort of the strategy employed (Bettman et al., 1990). Other researchers have found that heuristics can outperform an uncertainty optimization model (Gigerenzer et al., 2011).

The decision making under uncertainty literature generally focuses on single-task scenarios, few studies explore operator strategies in dual-task scenarios. One strategy noted in both single and dual-task research is the the probabilistic matching strategy, which occurs when the human's compliance and reliance rate roughly mirrors the alarm/non-alarm reliability (Dorfman, 1969; Bartlett and McCarley, 2017). An example of the probabilistic matching would when operator is aware that automation alarms are 70% accurate, leading them to comply with alarms approximately 70% of the time. This strategy frequently results in suboptimal detection performance and has sparked debate regarding whether it represents a sophisticated response to uncertainty or a human cognitive limitation (Koehler and James, 2014).

An extreme response strategy occurs when the operator decides to consistently use or discard the automated aid (Bliss et al., 1995). Theoretically, the rational operator could use information about automation accuracy and situational specific consequences (costs and benefits) to model the outcome. They could then calculate a threshold and decide to always use or ignore the alarm to maximize their performance. Early recognition of the extreme response strategy dates back to a dissertation by Bliss (1993) and published in Bliss et al. (1995). The study found that a small number of participants (just under 10%) resorted to an extreme response strategy. Participants appeared to be sensitive to alarm performance; they applied an extreme disuse strategy to automation that was 25% reliable and an extreme use strategy to automation that was 75% reliable.

In a subsequent study by Bliss (2003), the authors compared extreme response strategies for participants informed about collective alarm performance to participants that could access AVI (trial specific information to validate each alarm). They gathered empirical data from 7 previous studies and categorized participants in each study as over-responders (followed every alarm) or under-responders (ignored every alarm). They found that the participants who could not validate alarms were much more likely to resort to an extreme use response strategy. They found a tendency of participants to over-respond to the automation and some participants ignored all alarms if the alarm's reliability was at or below 50%.

To explore the effects of alarm reliability and AVI on operator response strategies, Manzey et al. (2014) conducted 4 controlled laboratory experiments. Participants conducted a dual-task experiment simulating a chemical plant control room. The authors manipulated the experimental base rate to examine a variety of reliability levels, with PPV ranging from 0.10 to 0.90 and NPV ranging from 0.98 to 0.41. Extreme response strategies were categorized when participant followed or ignored 90% or more of automation's recommendations; strategies with frequencies between these were categorized as mixed. In conditions without AVI, the majority of participants utilized an extreme strategy during alarms, with more extreme use found in high PPV conditions and more extreme disuse in low PPV conditions. The extreme disuse was only found in PPV conditions on 0.5 or below. A different pattern was noted during non-alarm strategies, where participants had extreme use during all non-alarm NPV conditions and extreme non-alarm disuse only occurred in the lowest (NPV = 0.41) condition. In the second experiment, where participants had access to AVI, the extreme response behaviors appeared to be mitigated, except in the lowest performance conditions. The authors attributed the reduction in extreme strategies to participants incorporating the AVI.

We suggest that, in addition to the impact of AVI access, the global perspective on strategies might have artificially magnified the decline in extreme strategies. It is plausible that information might have been omitted, especially considering the operationalization of strategies at 90%. If in the Manzey et al. (2014) experiment, participants initially employed an extreme use strategy and transitioned to extreme disuse after the first 10% of the experiment, they would be categorized as employing a mixed strategy. Operators often adjust their dependence behaviors as they gain experience with a particular automated aid. Finally, other prior work suggests, but does not focus on, that operators are changing their strategy to compensate for automation (McBride et al., 2011).

While recent trust research indicated that trust is dynamic ((Guo and Yang, 2021; Yang et al., 2017)), rather than the traditional 'snapshot' view (Merritt et al., 2013; Bailey and Scerbo, 2007). There is a gap surrounding the examination of changes in dependence behaviors. If attitudes toward automation are dynamic, the corresponding response behaviors may parallel the dynamic attitudes. Additionally, studies have categorized operator strategies (Bliss, 2003; Manzey et al., 2012; Bliss et al., 1995), however, there is a gap surrounding how operators are adapting their strategies.

1.4.3 Related Research on Disclosing Information

1.4.3.1 Likelihood Alarms

Likelihood alarms offer information to human operators beyond the automation's recommendation, as they not only provide the recommendation but also convey information about the confidence level associated with the recommendation. An early examination of likelihood information was conducted by Sorkin et al. (1988), where researchers examined the operator's sensitivity and concurrent task performance. Participants completed a dual task experiment using a joystick for a primary-tracking task and completed a numeric monitoring secondary task. Participants were separated across primary task difficulty (easy vs hard) and received an automated alert in either a binary or 4-state likelihood, separated by visual or speech alarm formats. Participants completed an immense amount of trials, roughly 8000, but the study sample size was limited (6 students). The authors only found improvements of likelihood alarms, when compared to binary alarms, when the tracking task was difficult.

In an alternative study conducted by Fletcher et al. (2017), researchers investigated whether a visual likelihood alarm would improve human performance on a radar simulation task. Participants utilized a visual likelihood alarm to distinguish objects, specifically potential enemy submarines. The findings revealed that while the likelihood alarm did not enhance operator sensitivity, it did influence response bias and spatial attention.

1.4.3.2 Incorrect Likelihood Information

In addition to likelihood alarms, researchers have observed how perceptions of the automation can play a mediating role in influencing automation use (Merritt and Ilgen, 2008; Sanchez, 2009).

In a study by Barg-Walkow and Rogers (2016), researchers investigated the effects of incorrect reliability information. Participants completed a dual task warehouse management simulation, where they completed a bar-code matching task and a truck-dispatching task. An imperfect (75% reliability) automated aid provided recommendations for the truck dispatching task. Prior to the experiment, participants were provided reliability statements regarding the automated aid. Researchers investigated dependence behaviors (i.e., compliance and reliance) and task performance over either 2 or 4 days. The reliability statements varied in accuracy (correct or incorrect), and statement type; certain statements manipulated perceptions to enhance perceived automation reliability, while other statements diminished perceptions. Their findings revealed a consistent improvement in both compliance and reliance as the experiment progressed. Additionally, the manipulation of expectations did not influence compliance or reliance behaviors.

1.4.3.3 Overall Success Likelihood

Several investigations have disclosed the overall success likelihood values to participants and found that participants demonstrated a more appropriate use of automated systems (Walliser et al., 2016; Dzindolet et al., 2002; Wiczorek and Manzey, 2014).

In a study conducted by Dzindolet et al. (2002), researchers investigated the influence of presenting participants with the number of automation errors on the operator's perception of automation, performance, and reliance. Participants were tasked with examining pictures of military terrain and deciding on the presence of a camouflaged enemy soldier. They received assistance from either a human or an automated aid. Half of the participants were provided with information about the number of automation errors. The findings revealed a preference for both aids when participants were informed about their reliability.

1.4.3.4 Hit and CR Rates

In the study conducted by Bagheri and Jamieson (2004), researchers explored an alternative form of likelihood information provided to decision-makers, specifically, the hit and correct rejection (CR) rates. The authors compared operator detection performance across different information conditions by contrasting data from a previous study, where participants received no reliability information, to conditions where participants were presented with data representing the hit rate of the automation. Participants engaged in a flight task simulation using the Multi-Attribute Task Battery (MAT-B; Comstock Jr and Arnegard (1992)), simultaneously managing fuel, tracking, and system monitoring tasks. The imperfect automation assisted in the system monitoring task. Despite expectations, the results indicated that revealing likelihood information in this format did not lead to performance improvements.

1.4.3.5 Predictive Values

Predictive values have been found to positively influence operator reliance on automation. In the study conducted by Wang et al. (2009), participants were tasked with utilizing an imperfect combat identification aid (CID) to determine whether a target was a friend or enemy. The experimental design employed a 3 (automation conditions: no aid, 67% reliable aid, and 80% reliable aid) x 2 (information disclosed: informed or uninformed) experimental design. In the informed conditions, the researchers disclosed the automation's PPV to participants. The specific aid employed for their experiment was always correct when it determined a target was a friend. However, when the aid identified a target as unknown, the target could be a friend, enemy, or neutral. The results indicated that participants who were informed about the PPV exhibited more appropriate reliance compared to those in the uninformed condition.

1.4.3.6 Frequency Formats

As highlighted above, the dual-task literature has explored the effects of providing a diverse variety of likelihood information, with hopes of enhancing the human's decision-making. However, one format that has not been extensively investigated in dual-task contexts is the frequency format. In the literature, two common formats are discussed: natural frequency formats and normalized frequency formats (Hoffrage et al., 2002). Natural frequency provides likelihood information in the form of joint frequencies collected from natural samples, while normalized frequency formats are similar, but the base rate is fixed *a priori*. Debate persists regarding the potential benefits of information presented in a frequency format for aiding decision-makers.

In seminal work by Gigerenzer and Hoffrage (1995), authors illustrated different ways to summarize statistical information, specifically in ways that are mathematically equal but computationally distinct. The authors noted how frequencies are computationally much simpler to use than probabilities format for decision makers in a SDT task. An example would be how there are two correct methods to calculate a positive predictive value. The first would be to utilize Bayes' Theorem to calculate the inverse of hit rate:

$$\frac{p(Alarm \mid Threat) * p(Threat)}{p(Alarm \mid Threat) * p(Threat) + p(Alarm \mid NoThreat) * p(NoThreat)} \tag{1.1}$$

The second, and much simpler method would be to calculate the PPV using the SDT measures of hits and false alarms:

$$\frac{Hits}{Hits + False\ Alarms} \tag{1.2}$$

Although there is a lack of studies in the dual-task domain investigating frequency formats, it has been widely discussed in single-task SDT areas. In a study by Hoffrage and Gigerenzer (1998), researchers provided 48 physicians with four diagnostic problems. Each physician decided on the presence of a signal, they were to diagnose on the status of 2 forms of cancers and 2 diseases. Likelihood information was presented to the physicians, in addition to a positive test result, in the form of probabilities and natural frequencies. The information type and question order was systematically varied. The physicians were required to calculate the predictive values for each case. When physicians were provided the probabilities, they were only able to successfully calculate the predictive values in 10% of cases. The outcomes with natural frequency formats exhibited much better performance, reaching a success rate of 46%. A meta-analysis conducted by McDowell and Jacobs (2017), synthesized 20 years of research to collected 35 articles with 226 performance estimates. They found that probabilistic decision making improved when people are provided with conditional information in a naturally sampled frequency. There are ongoing scientific discussions exploring ways to enhance information for decision-makers (Trevena et al., 2021).

While single-task research has indicated benefits of disclosing information to individuals in a frequency format, the benefits surrounding frequency formats have not been extensively explored in the dual-task paradigm.

1.5 Research Aims

A substantial amount of research has focused on automation use over the last 20 years, there are several gaps that this dissertation aims to address:

First, while the literature mentions specific accidents and incidents that are attributed to inappropriate automation use, it is still unclear how well automation must perform for operators to depend on recommendations, or how poorly automation performs for operators cross-check with additional information. Many studies focused on specific variables, including workload (Bliss and Dunn, 2000; Lin et al., 2020), age (Sanchez et al., 2014; McBride et al., 2011), and expectations (Mayer, 2008; Barg-Walkow, 2013). A previous literature synthesis examined the costs and benefits of imperfect automation; however, there is not a systematic overview of empirical data examining when operators are choosing to blindly follow or cross-check automation with AVI.

Second, the dependence measures examined are typically presented as a global measure (i.e., they examine all trials at the end of an experiment). However, perceptions of automation change while the human uses the automation, in turn the behaviors used are dynamic. Restricting examinations to a global measure of automation may lead to a loss of information. Additionally, when examining group dependence behaviors, individual automation use and strategy can be lost.

Finally, operators have been shown to use likelihood information to improve their automation use. We wanted to investigate this further, as previous studies did not include a baseline and there are additional likelihood information formats that may improve performance outcomes. We aimed to improve the knowledge around dependence behaviors and likelihood information.

To address the problem of inappropriate use of imperfect automation and contribute to the knowledge gaps surrounding alarm validity information, individual strategy, and the dynamic use of automation, the aims of this dissertation were to:

1. Systematically examine how operators' dependence behaviors (i.e., compliance and reliance) and cross-checking behaviors are affected by automation performance, separated across alarm states.

2. Evaluate how operators adapt their dependence behaviors, cross-checking rates, and response strategies to varying degrees of imperfect automation.

3. Investigate a design intervention focusing on the incorporation of likelihood information, specifically, to compare the effects of predictive values with a frequency format and a baseline condition, where no *a priori* information was provided.

1.6 Dissertation Structure

This dissertation work is structured as five chapters. Chapter one introduces the problem, the aim of this dissertation's research, and reviews relevant studies. Chapter two delves into a meta-analysis, where empirical data was systematically gathered to illustrate relationships between automation reliability and operator dependence. Chapter three presents a laboratory study, where automation performance was manipulated to observe and understand how individuals adapt their dependence behaviors and response strategies. Chapter four focuses on the influence of likelihood information on individual operators' response behaviors, strategies, and human-automation team performance. This chapter comprises of two distinct components: a reanalysis of an existing study and a novel laboratory experiment which builds off the reanalysis results. Chapter five will merge findings across the research, present the intellectual merit and broad impact, and propose avenues for future research.

CHAPTER 2

Meta-Analysis Examining the Impact of Automation Reliability on Dependence Behaviors

2.1 Introduction

The landscape of automation errors is well-documented, including incidents such as the Royal Majesty cruise grounding (Degani, 2001; NTSB, 1995) and the fatal Tesla crash in Mountain View, California (Laris, 2020; NTSB, 2020). Substantial research attention has been directed towards inappropriate automation use, with a focus on investigating instances where humans either over- or under-use automation (Parasuraman and Riley, 1997; Lee, 2008). In dual-task scenarios, the human operators are challenged with deciding when to use or discard automation and must bear responsibility for performance outcomes.

The benefits of a decision aid have been found to decline as the aid's reliability decreases (Rovira et al., 2007; Madhavan and Wiegmann, 2007; Skitka et al., 1999). It is evident that as reliability decreases, dependence tends to decrease as well. However, when examining performance across alarm and non-alarm conditions (i.e., PPV and NPV), the literature has presented a mixed discussion regarding the impact of different error types (i.e., false alarms or misses). Some empirical findings support the notion of separate effects (Meyer, 2001; Dixon and Wickens, 2006; Wickens and Colcombe, 2007; Chancey et al., 2017). Alternative studies indicate that the effects of false alarms are not selective across the types of errors (Meyer et al., 2014; Dixon et al., 2007). Frameworks have been created to organize factors contributing how operators detect, understand, and correct errors (collectively referred to as error management) (McBride et al., 2014).

Researchers have explored various factors influencing human operators use of automation, including expectations (Mayer, 2008; Barg-Walkow, 2013), workload (Bliss and Dunn, 2000; Lin et al., 2020), and age (Sanchez et al., 2014; McBride et al., 2011), among others. However, while individual studies have examined the impact of automation performance on the dependence and/or cross-checking rates of human operators, there is not a literature synthesis to provide an overview of operator dependence and cross-checking behaviors across studies. This chapter aims to fill in the research gap by conducting a meta-analysis to examine the impact of automation reliability on dependence behaviors.

2.1.1 Current Study

To address the gaps noted above, we conducted a meta-analysis, involving an extensive literature search and aggregating empirical data to further explore how automation performance influences operators' dependence and cross-checking rates across multiple studies and scenarios. It is important to note certain scope limitations: first, our focus is on human use of diagnostic automation, where the automation performance can be modeled using SDT (Tanner and Swets, 1954; Macmillan and Creelman, 2005). Second, we concentrate on studies utilizing the dual-task paradigm, where operators engage in both an automation-aided task and a manual task simultaneously. Third, our emphasis lies on compliance and reliance behaviors rather than response times, such as the speed of human operators' compliance with automated alarms.

2.1.2 Contributions

This study makes three primary contributions to the literature. First, we systematically assembled empirical data to illustrate the effects of imperfect automation on operator dependence. Secondly, we differentiated operators' dependence behaviors based on alarm state and their utilization of AVI. Finally, we highlighted an automation performance 'choice' threshold between when operators chose to blindly follow automation or cross-check with the AVI. This study contributes to our comprehension of decision-making under uncertainty, specifically how operators depend on, cross-check, or ignore imperfect automation's recommendation.

2.2 Methods

We conducted a large-scale literature search to identify relevant articles and visualized the search using the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) process (Moher et al., 2009). Data were extracted from these studies to examine the effects of automation performance on dependence behaviors, additionally our analyses considered use of alarm validity information.

2.2.1 Searching Relevant Literature

We guided our search process based on prior literature review articles (Kaplan et al., 2021; Hancock et al., 2011, 2021; Schaefer et al., 2016; Hoff and Bashir, 2015). The search was restricted to empirical research from articles published between January 1996 and June 2021. The lower bound of the time period was based on the review by Hancock et al. (2011), which is one of the earliest literature review articles about relationships between humans and robots. We searched 6 digital libraries: Science Direct, JSTOR, ACM Digital Library, EBSCOhost, ProQuest, and IEEE Xplore, using combinations of key terms such as *trust, compliance, reliance, dependence, automation, automated system, autonomy.* Results were restricted to

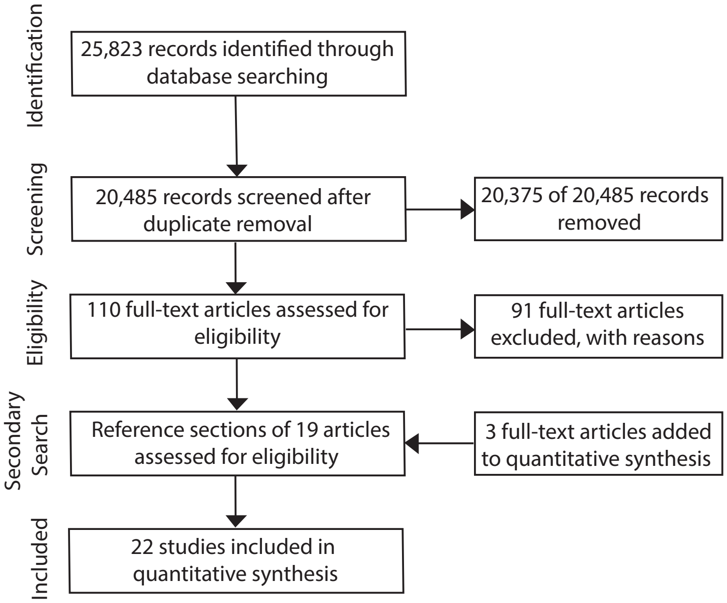

Identification
25,823 records identified through database searching
Screening
20,485 records screened after duplicate removal
20,375 of 20,485 records removed
Eligibility
110 full-text articles assessed for eligibility
91 full-text articles excluded, with reasons
Secondary Search
Reference sections of 19 articles assessed for eligibility
3 full-text articles added to quantitative synthesis
Included
22 studies included in quantitative synthesis

Figure 2.1: PRISMA overview of literature search process.

journal publications, conference proceedings, and theses. Two researchers independently conducted the article exclusion and secondary search process and results were merged to include all possible articles.

If data was published multiple times, such as in a conference proceedings paper, a thesis, and a journal article, we chose the thesis instead of the journal article and the journal article instead of the conference proceedings paper. This prioritization was selected because a thesis typically reported more detailed data than the journal article, and the journal article often reported more detailed data than the conference proceedings paper. Figure 2.1 visualizes our literature search process.

Our initial search yielded 25,823 records. There were a considerable amount of duplicates because multiple search engines and keywords were used. Removing the duplicates yielded 20,485 unique records.

2.2.1.1 Screening Criteria

The 20,485 records were screened with the following criteria:

1. Researchers conducted a human-subject experiment wherein participants interacted with automation;

2. The automation had to be imperfect for at least one experimental condition;

3. Automation that provided binary diagnoses regarding signal/noise (conducted functions of diagnostic automation);

4. The operator could choose how to respond to the automation: they could follow the diagnosis, ignore it, or cross-check the automation's diagnosis against the ground truth;

5. The participant was required conduct at least two tasks simultaneously.

The screening process excluded 20,375 articles. The majority of these articles were excluded because researchers did not conduct a human-subjects experiment.

2.2.1.2 Eligibility Criteria

The remaining 110 articles were read carefully to ensure that they met the following eligibility criteria:

1. The article contained sufficient data for quantifying the automation's performance (PPV and/or NPV and/or Overall Success Likelihood (OSL));

2. The article reported at least one human response behavior (i.e., compliance, reliance, or dependence).

2.2.1.3 Secondary Search

After removing ineligible articles, 19 articles remained. Google Scholar was used to seek additional articles by scanning the reference lists of the 19 articles. We identified 3 additional articles to expand our total to 22.

2.2.2 Data Extraction

From each of the 22 records, we extracted available data on automation performance, participants' response behaviors, and whether participants cross-checked the automation. We distinguished three types of response behaviors based on *if* the operator used the alarm validity information (AVI); which were blind behaviors (AVI not used), cross-checking behaviors (AVI used), and agreement (AVI usage unclear in article). The independent variables in articles we extracted from were treated as random noise in our meta-analyses. In addition, researchers may have conducted multiple experiments within one article and the resulting data was nested in our models accordingly.

When performing data extraction, we noted that some studies reported the PPV, NPV, and OSL values directly, whereas others did not. When not directly reported, we calculated the PPV and NPV based on the number of hits, misses, FAs, and CRs used in the experiment. Some studies reported the response behavior rates in numerical format whereas others plotted the values in figures. For the latter, we interpolated the figures to estimate the reported values.

2.2.3 Data Analyses

We built multilevel models (MLMs) using the 'nlme' package in R (version 4.3.1). We treated the automation performance (PPV, NPV, or OSL) as the predictor for each response behavior (i.e., blind dependence, cross-check, and agreement). In each MLM, the extracted data were nested within article. The analyses used an unstructured error structure and maximum likelihood estimation. We followed the standard procedure to build MLMs (Hofmann et al., 2000), by initially building simple models and then more complex models: from fix-effects only model, to random-intercept model, and finally to the random-slope random-intercept model. We compared the Akaike information criterion (AIC) scores between the simpler and more complex models to determine if the more complex models were needed. Whenever a non-significant comparison was found or a more complex model failed to converge, we reverted to the simpler model. An independent samples t-test was conducted to compare the model's slopes for blind and cross-checking behaviors, this was done for alarm and non-alarm states. Significance for α in statistical tests was set at 0.05.

2.3 Results

Our results are separated by the following relationships: (i) between alarm response behaviors and PPV, (ii) between non-alarm response behaviors and NPV, and (iii) response behaviors against OSL. It should be noted that we only used the OSL data when we could not separate it into alarm and non-alarm data. The counts of articles and data points for extracted data are below in Table 2.1 and are followed by Equations 2.1, where we define the variables.

Table 2.1: Articles and data point counts for the meta-analysis, separated by response behavior.

Data	Rate Type	# of Articles	# of Data Points
Alarm	Blind Compliance	12	93
	Cross-Check Rate	3	21
	Agreement With Alarms	3	15
Non-Alarm	Blind Reliance	9	82
	Cross-Check Rate	2	20
	Agreement With Non-Alarms	1	4
Alarm and Non-Alarm	Blind Dependence	12	105
	Cross-Check Rate	2	20
	Total Agreement	4	41

$$
\begin{aligned}
&\textit{Blind Compliance} = P(\textit{report and not cross-checking} \mid \textit{automation alert})\\
&\textit{Cross-Check Rate (Alarm)} = P(\textit{cross-check} \mid \textit{automation alert})\\
&\textit{Agreement With Alarm} = P(\textit{report} \mid \textit{automation alert})\\
&\textit{Blind Reliance} = P(\textit{not report and not cross-checking} \mid \textit{automation silence})\\
&\textit{Cross-Check Rate (Non-Alarm)} = P(\textit{cross-check} \mid \textit{automation silence})\\
&\textit{Agreement With Non-Alarm} = P(\textit{not report} \mid \textit{automation silence})\\
&\textit{Blind Dependence} = P(\textit{follow recommendation and not cross-checking})\\
&\textit{Cross-Check Rate} = P(\textit{cross-check})\\
&\textit{TotalAgreement} = P(\textit{follow recommendation})
\end{aligned}
\tag{2.1}
$$

2.3.1 Response Behaviors to Alarms

15 unique articles containing 129 data points reported data regarding alarm response rates, ranging from 1% to 99%. The average agreement with alarm rate ($M = 74\%,\ SD = 11\%$) was higher than the average cross-check rate ($M = 59\%,\ SD = 22\%$). These both were higher than the average blind compliance rate ($M = 44\%,\ SD = 24\%$).

PPV was a significant predictor of the 3 response behaviors and all 3 were best fitted with a random-slope fixed-intercept model. For blind compliance, there was a positive relationship with PPV, $t(80) = 9.16,\ p < 0.001$ and $\beta_1 = .74$. For cross-check rate, there was a negative relationship with PPV, $t(18) = -4.2,\ p < 0.001$ and $\beta_1 = -.52$. Finally, for agreement with alarm, there was a positive relationship with PPV, $t(11) = 5.52,\ p < 0.001$ and $\beta_1 = 0.81$. There was a significant difference found between the blind compliance slope ($\beta_1 = .74$) and the cross-checking rate ($\beta_1 = -.52$), $t(101) = -6.73, p < 0.001$. Figure 2.2 shows the MLM regression lines, where the data illustrates that both blind compliance and agreement increased with PPV, while the cross-check rate decreased as PPV increased.

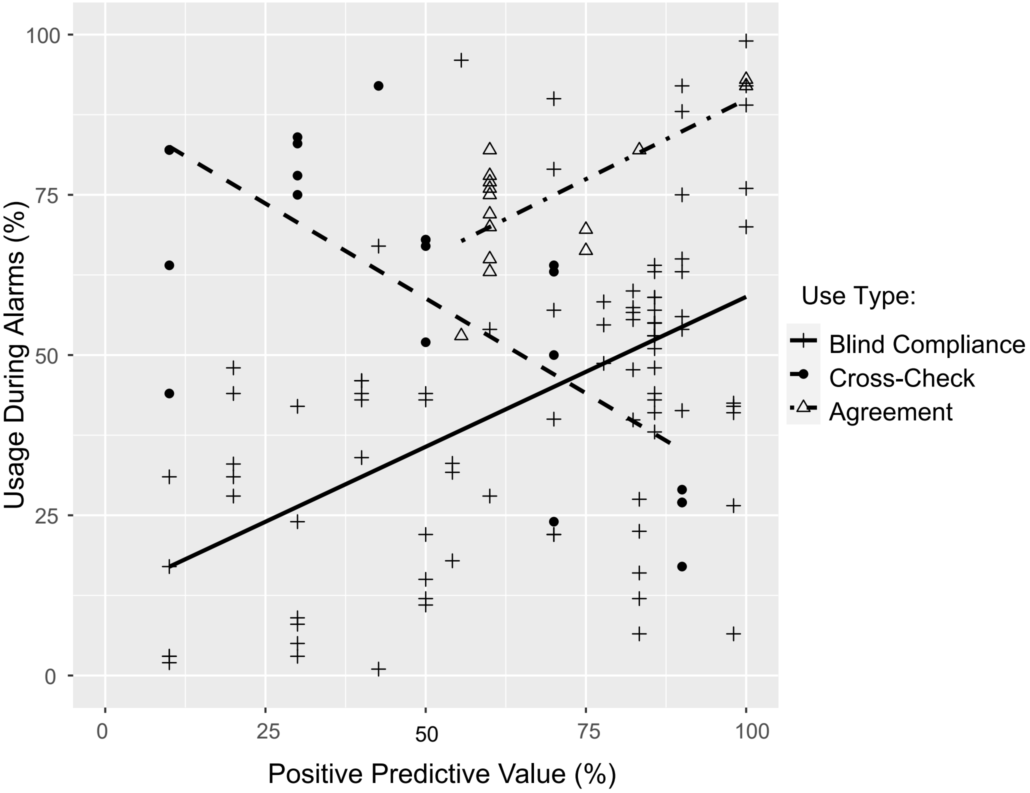


Figure 2.2: Alarm response behaviors are plotted against *PPVs*.

2.3.2 Response Behaviors to Non-Alarms

10 unique articles, containing 106 data points, reported data regarding non-alarm response rates, ranging from 2% to 99%. The average agreement with non-alarm rate ($M = 84\%,\ SD = 12\%$) was higher than the average blind reliance rate ($M = 60\%,\ SD = 22\%$). These both were higher than the average cross-check rate ($M = 35\%,\ SD = 23\%$).

NPV was a significant predictor for blind reliance and cross-check rate and the random-slope fixed intercept model was used for both. No meta-analyses for agreement with non-alarms were completed because this data was extracted from 1 article. For blind reliance, there was a positive relationship with NPV, $t(72) = 3.67, p < 0.001$ and $\beta_1 = .89$. For cross-check rate, there was a negative relationship with NPV, $t(17) = -6.25, p < 0.001$ and $\beta_1 = -.90$. There was a significant difference found between the blind reliance slope

($\beta_1 = .89$) and the cross-checking rate ($\beta_1 = -.90$), $t(99) = -7.84, p < 0.001$. Figure 2.3 shows the MLM regression lines, where the data illustrates that as NPV increased, blind reliance increased and the cross-check rate decreased.

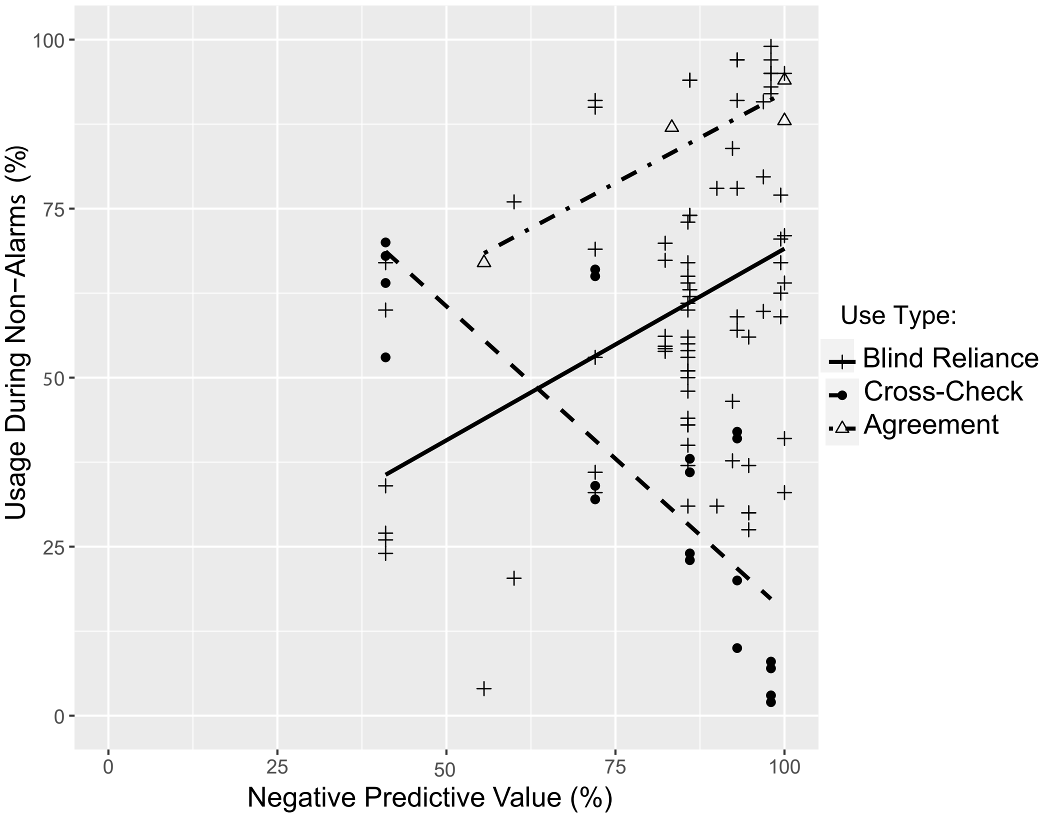


Figure 2.3: Operator responses to non-alarms are plotted against *NPVs*.

2.3.3 Response Behaviors Across Alarms and Non-Alarms

17 unique articles, containing 166 data points, reported data regarding automation usage across alarms and non-alarms, which ranged from 13.5% to 99.2%. The average agreement with automation ($M = 72\%,\ SD = 13\%$) was higher than the average blind dependence ($M = 57\%,\ SD = 18\%$). These both were higher than the average cross-check rate ($M = 44\%,\ SD = 13\%$).

OSL was a significant predictor for blind dependence and agreement with automation, both used a random-slope fixed intercept model. The *OSL* was not found to be a significant predictor of cross-check rate $t(17) = -.28,\ p = 0.78$, this is likely due to a small sample size (n=2). For blind dependence, there was a positive relationship with *OSL*, $t(91) = 4.41,\ p < 0.001$ and $\beta_1 = .57$. For agreement with automation, there was also a positive relationship with *OSL*, $t(36) = 9.38,\ p < 0.001$ and $\beta_1 = 0.79$. Figure 2.4 shows the MLM regression lines, where the data illustrates that as *OSL* increased, both blind dependence and agreement increased.

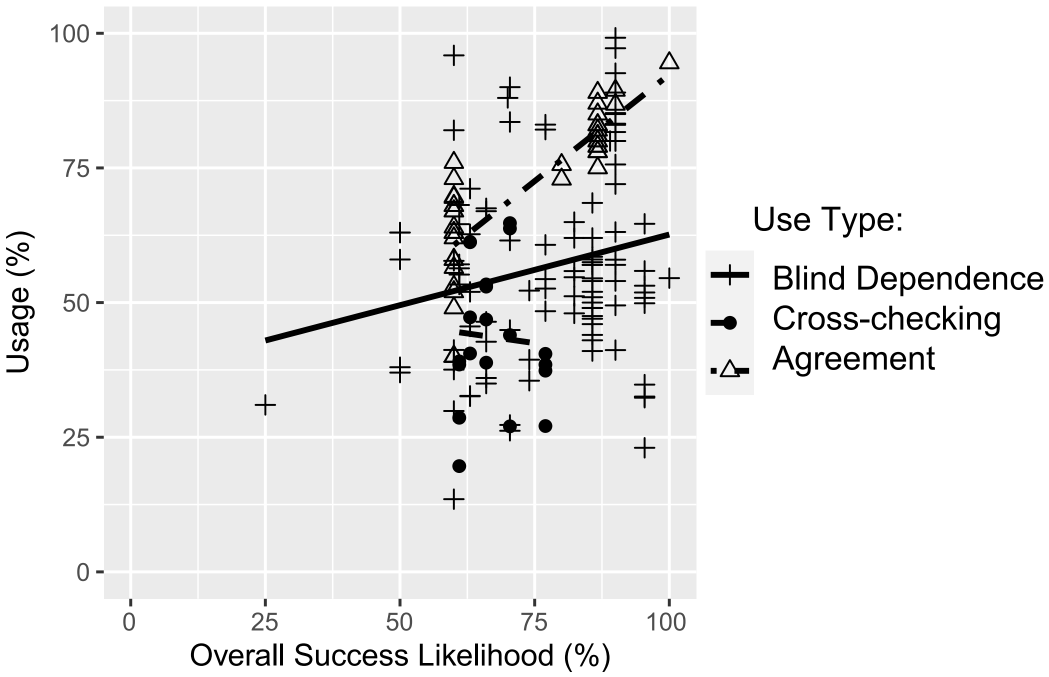


Figure 2.4: Operator responses to automation are plotted against *OSL*.

2.3.4 Comparing Alarm and Non-Alarm Response Behaviors

When comparing the slopes of blind compliance ($\beta_1 = .74$) and blind reliance ($\beta_1 = .89$), statistical differences were not found ($p = 0.37$). There were marginally significant differences found between the operator's cross-checking rates across alarms ($\beta_1 = -.52$) and non-alarms ($\beta_1 = -.90$), $t(38) = 1.78$, $p = 0.08$.

2.4 Discussion

The objective of this study was to examine the influence of automation performance onto human dependence behaviors, while considering use of AVI.

2.4.1 Operator Dependence and Cross-checking Behaviors

Our results show that operators' dependence behaviors increased as the automation's performance increased, which aligns with previous research (Skitka et al., 1999; Rovira et al., 2007; Madhavan and Wiegmann, 2007). As expected, the relationship between operators blindly following automation and automation performance is positive and linear; additionally this trend is illustrated across alarm and non-alarm states - indicating that operators are sensitive to how well automation is performing. When examining cross-checking response behaviors, we found that operators altered their use of alarm validity information based on automation's performance. Operators reduced their cross-checking as automation's performance increased, which aligns with prior research that considered AVI use (Manzey et al., 2014; Bustamante, 2005). The act of cross-checking in dual task scenarios requires redirecting the operator's attention, resulting in a higher cognitive demand than when an operator blindly follows automation. Each operator must evaluate the automated aid they are interacting with to decide whether they want to use additional information to validate the recommendation. Other researchers have mentioned how operators use the automated aid as an attention allocation tool to allocate attentional resources to a manual task (Du et al., 2020a; Manzey et al., 2014).

We compared operator responses across alarm and non-alarm states to determine if there were significant differences between states. Our results indicated no significant differences in the operator's sensitivity to errors across alarm and non-alarm states, with the blind reliance slope ($\beta_1 = .89$) being slightly higher than the blind compliance ($\beta_1 = .74$). Studies that specifically investigated the effects of false alarms on compliance and reliance behaviors yielded similar findings (Dixon et al., 2007; Rice, 2009). Our results also indicated that operator's cross-checking behaviors were marginally more sensitive to non-alarm errors ($\beta_1 = -.90$) than alarm errors ($\beta_1 = -.52$). However, this should be noted with caution as the sample size was small. Other researchers have utilized controlled experiments to determine if there are differences stemming from error type (i.e., false alarms or misses). Meyer (2004, 2001) postulated that misses and false alarms have different effects onto operator dependence. Rice (2009) suggested that an asymmetry may stem from the inherent cognitive salience of false alarms, as false alarms have greater salience than misses, leading to reduced sensitivity to true and false alarms. In a study by Gérard and Manzey (2010), the authors compared alarm's influence on compliance to non-alarm's influence on reliance. They manipulated the PPV from 90% to 70% and found that operators reduced their blind compliance by roughly 25%. A 5% reduction (from 98% to 93%) in NPV was required to reach the same 25% reduction in blind reliance, implying that operators were more sensitive to non-alarm errors. The authors attributed the disproportionately strong reliance shifts to the tendency of participants to avoid errors of omission, when possible. Additionally, reduced sensitivity to alarms has been attributed to frequently occurring false alarms; this has been described as the 'cry-wolf' phenomena (Breznitz, 1984). Researchers have investigated this phenomena and found inconsistencies (Dixon and Wickens, 2006; Gérard and Manzey, 2010; Bliss, 2003; Dixon et al., 2007).

2.4.2 Choice Threshold

By examining where the regression lines of cross-checking and blind dependence intercept, a reliability threshold is found that indicates an operator's behavioral shift between blind and cross-checking response behaviors. The data indicate that this point is where operators choose to switch from blindly following the automation to validating the automation with AVI. This point appears be near 60% for non-alarms and 70% for alarms. Interestingly, this threshold is close to a threshold from a previous meta-analysis. Wickens and Dixon (2007) synthesized literature to quantify the cost-benefit trade-off operators experienced when using imperfect automation. They found that around 70% was a threshold where the benefits of an automated aid would outweigh the costs of errors. When automation performed better than 70% correct, the benefits outweighed the cost of errors; when the automation performed worse than 70%, the costs outweighed the benefits. Our data suggest that operators are cross-checking appropriately; they tend to blindly follow automation when the costs of automation outweigh the benefits and cross-check otherwise.

2.4.3 Agreement

The measure of agreement is a combined measure of blind behaviors and cross-checking, where it was unclear if AVI was used. We collected this category of data when authors did not report and/or control for AVI use, likely since AVI was not one of their variables of interest. An example would be an experiment where there are 2 display windows on one screen, with one display showing information that would validate the automation. The participant would be able to rapidly cross-check the automation's recommendation without it being quantified. For these cases, we were unable to determine when operators *actually* used AVI. The slope in this data aligns with the blind behaviors, illustrating that agreement increased as automation performance increased. However, caution should be taken when examining agreement, if participants had continuous access to AVI, such as a secondary screen with the ground truth, agreement would match the automation performance. This is because they would agree more with better performing automation only because it is correct more often than poorly performing automation. If it is wrong, they would see the ground truth and disagree with the recommendation.

2.4.4 Limitations and Future Work

This review, however, disregards variables other than automation performance and dependence behaviors. We treated independent variables from extracted studies as statistical noise, this included age (Sanchez et al., 2014; McBride et al., 2011), workload (Bliss and Dunn, 2000; Lin et al., 2020), and alarm expectations (Mayer, 2008; Barg-Walkow, 2013). Future work could examine the relationship between dependence and cross-checking behaviors with other variables, such as human-automation team performance, attention allocation, or perceived effort. Finally, meta-analyses pool data from individual operators over the duration of each experiment; within each experiment, operator behaviors and strategies are likely to change as they gain experience with that specific automated aid. Researchers could request raw data from various authors, with the goal of an individual level analyses.

2.5 Conclusions

We systematically extracted data from the literature to examine how operators used imperfect automation in dual-task scenarios. The human operators not only varied their compliance and reliance behaviors to the automation, but also varied how often they used information to verify the automation's recommendation. In general, as automation performed better, human operators were more likely to blindly follow the recommendations. Alternatively, when automation performed worse, the operators were more likely to cross-check the ground truth and verify automation's recommendation. Our data suggest that operators are choosing to use alarm validity information appropriately; suggesting that operators cross-check with AVI when the costs of automation outweigh the benefits (Wickens and Dixon, 2007). Considerations of automation usage and alarm validity information can promote effective decision-making and improve safety and efficiency.

2.6 Acknowledgement

I wish to express gratitude for the valuable assistance provided by Yuxin Chen. Yuxin and I independently applied the screening and eligibility criteria to the articles, and extracted data (on operator behaviors and automation performance) from the articles. Her efforts helped

ensure the quality of the screening process.

CHAPTER 3

Changes in Operators' Dependence Behaviors, Response Strategies, and Performance with Imperfect Automation

3.1 Introduction

Extensive research has investigated operators' response behaviors across varying levels of automation reliability (Getty et al., 1995; Bliss et al., 1995; Manzey et al., 2014; Yang et al., 2017; Du et al., 2020a). These investigations largely aggregate behaviors over the experimental duration and present a 'snap-shot' view of operator dependence across conditions. Trust in automation, defined as the "attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004); has been conventionally treated as a 'snapshot' through end-of-experiment surveys (Merritt et al., 2013; Bailey and Scerbo, 2007; Chancey et al., 2017; Chavaillaz et al., 2016). However, recent research indicates the dynamic nature of human trust in automation (De Visser et al., 2020; Guo and Yang, 2021; Guo et al., 2023). This chapter is grounded in the conceptualization that, if attitudes toward automation are dynamic, the corresponding response behaviors may serve as behavioral analogs to these dynamic attitudes.

In addition to dynamic behaviors, this chapter will focus on operator strategies. Operators are known for rarely following purely logical considerations, instead they utilize heuristics and develop strategies to making decisions under uncertainty (Tversky and Kahneman, 1974; Gigerenzer and Todd, 1999). As logical automation operators repeatedly predict one of two alternatives, they often employ a strategy to compensate for automation errors. Heuristics are strategies that do not prioritize performance outcomes, instead, they are useful when dealing with uncertainty, optimization is challenging, and cognitive resources are exceeded by task demands (Gigerenzer, 2008; Gigerenzer et al., 2011). Prior research within the dual-task paradigm analyzed dependence behavior rates and identified two operator strategies: the probabilistic matching strategy and the extreme response strategy (Manzey et al., 2014; Bliss et al., 1995; Bliss, 2003). The probabilistic matching strategy occurs when the human's compliance and reliance rate roughly mirrors the alarm/non-alarm reliability (Dorfman, 1969; Koehler and James, 2014; Bliss et al., 1995). An extreme response strategy occurs when the operator decides to consistently use or discard the automated aid (Bliss, 2003; Manzey et al., 2014).

3.1.1 Current Study

The aim of this chapter is to evaluate how operators' adapt their dependence behaviors, cross-checking rates, and response strategies to varying degrees of imperfect automation. We wanted to clarify how operators adapt in response to imperfect automation; specifically, by manipulating the automation performance (i.e., PPV and NPV) and observing changes in operator dependence behaviors and strategies. This chapter serves as an extension of a prior investigation (Schuler and Yang, 2023), where the dependence behaviors and strategies of operators, in response to interactions with an automated aid, were explored with one reliability condition. The current study expands on the prior investigation by manipulating automation reliability across five conditions. In addition, we propose an ideal optimal strategy that incorporates AVI, as the literature had a gap surrounding optimal decision making when information is available to reduce uncertainty. Our current study concentrates on the extreme response strategy, given its relevance to the ideal optimal strategy proposed.

We hypothesized that participants would adjust their strategies as they increase experience with the automated aid, and the resulting performance would converge toward an optimal strategy. Additionally, we hypothesized that operators would vary their adaptation rates based on the automation's performance, with faster adaptations to the highest and lowest reliability conditions.

Limitations include the absence of a priori information on automation performance, necessitating participants to assess performance and form opinions on a trial-by-trial basis. The choice to divide the experiment into quarters was a deliberate trade-off, balancing considerations of alarm characteristics, experiment duration, and strategy operationalization. Furthermore, the study focused on operationalizing and tallying extreme strategies, which were optimal for the experiment and scenario.

3.1.2 Contributions

This chapter contributes to the literature in two primary ways. First, it introduces a novel optimal strategy incorporating alarm validity information. This strategy designates when participants should use, validate, or ignore an automated aid. Second, it addresses a literature gap by exploring how operators adapt their dependence behaviors and response strategies as they increase their experience with an automated aid. This knowledge holds significance for engineers and designers of autonomous systems and can facilitate the creation of individualized systems and refined training procedures.

3.2 Optimal Response Strategy with Fully Informative AVI

Yang (2024) proposes an ideal strategy that incorporates when participants can cross-check automation's recommendation against a ground truth. Given an automation alarm, we can calculate the expected utility of cross-checking the alarm or blindly complying with or ignoring the alarm, $U(cross\text{-}checking(cc)|alarm)$, $U(y|alarm)$, and $U(n|alarm)$, respectively. cc means the human operator chooses to cross-check, y means the human operator responds yes, e.g., evacuate after hearing a fire alarm, and n means the human operator responds no, e.g., not evacuate after hearing a fire alarm. $U(cc|alarm)$, $U(y|alarm)$, and $U(n|alarm)$ are calculated in Equations 3.1, 3.2, and 3.3.

Assume that after cross-checking, the human operator has full access to the true state of the world; i.e., the human operator responds yes when there is a signal and responds no when there is no signal. The expected utility of cross-checking:

$$\begin{aligned} U(cc|alarm) &= p(S|alarm) \times V_{hit} + p(N|alarm) \times V_{CR} - C_{cc} \\ &= PPV \times V_{hit} + (1 - PPV) \times V_{CR} - C_{cc} \end{aligned} \tag{3.1}$$

The expected utility of blindly complying with the alarm:

$$\begin{aligned} U(y|alarm) &= p(S|alarm) \times V_{hit} - p(N|alarm) \times C_{FA} \\ &= PPV \times V_{hit} - (1 - PPV) \times C_{FA} \end{aligned} \tag{3.2}$$

The expected utility of blindly ignoring the alarm:

$$\begin{aligned} U(n|alarm) &= p(N|alarm) \times V_{CR} - p(S|alarm) \times C_{Miss} \\ &= (1 - PPV) \times V_{CR} - PPV \times C_{Miss} \end{aligned} \tag{3.3}$$

where S presents a signal in the world, and N means there is only noise in the world.

To maximize the expected utility, a person should crosscheck if $EU(cc|alarm) > EU(y|alarm)$ and $EU(cc|alarm) > EU(n|alarm)$.

$$\begin{aligned} PPV \times V_{hit} + (1 - PPV) \times V_{CR} - C_{cc} > PPV \times V_{hit} - (1 - PPV) \times C_{FA} \\ \Rightarrow PPV < 1 - \frac{C_{cc}}{(V_{CR} + C_{FA})} \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} PPV \times V_{hit} + (1 - PPV) \times V_{CR} - C_{cc} > (1 - PPV) \times V_{CR} - PPV \times C_{Miss} \\ \Rightarrow PPV > \frac{C_{cc}}{(V_{hit} + C_{Miss})} \end{aligned} \tag{3.5}$$

Outside of the two bounds, the human operator should use the extreme response strategy.

$$P(y|alarm) = \begin{cases} 1 \quad when\ PPV > \max(\frac{V_{CR}+C_{FA}}{(V_{hit}+C_{Miss})+(V_{CR}+C_{FA})}, 1 - \frac{C_{cc}}{(V_{CR}+C_{FA})}) \\ 0 \qquad when\ PPV < \min(\frac{V_{CR}+C_{FA}}{(V_{hit}+C_{Miss})+(V_{CR}+C_{FA})}, \frac{C_{cc}}{(V_{hit}+C_{Miss})}) \end{cases}$$

Figure 3.1 plots the optimal compliance behavior a person should display assuming equal V_{hit}, V_{CR}, C_{FA} and V_{Miss}, and assume the C_{cc} is half of V_{hit}. In this case, $\frac{V_{CR}+C_{FA}}{(V_{hit}+C_{Miss})+(V_{CR}+C_{FA})} = 0.5$, $1 - \frac{C_{cc}}{(V_{CR}+C_{FA})} = 0.75$, and $\frac{C_{cc}}{(V_{hit}+C_{Miss})} = 0.25$

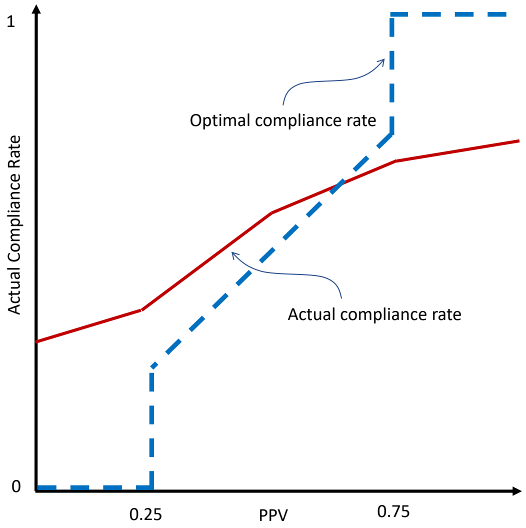


Figure 3.1: Red solid line indicates actual compliance rates and the blue dashed line indicates optimal compliance rates.

Figure 3.1 visually represents when automation's $PPV > 0.75$, human operator should always respond *yes*, i.e., optimal compliance rate = 1. When automation's $PPV < 0.25$, human operator should always blindly respond *no*, i.e., optimal compliance rate = 0. Between these thresholds, the operator should comply at a rate matching the automation's PPV and otherwise cross-check.

3.3 Methods

This study adhered to the ethical guidelines outlined by the American Psychological Association and received approval from the Institutional Review Board at the University of Michigan.

3.3.1 Participants

A total of 125 (50 males, 72 females, and 3 people who identified as non-binary) university students (mean age = 22.40 years old, SD = 3.34) with normal or corrected to normal vision participated in our experiment. Each participant was assigned randomly to one of the five experimental conditions and received a base rate of USD $20, plus a performance bonus ranging from $2.50 to $10.

3.3.2 Apparatus and Stimuli

The study was conducted utilizing an HP 24-inch monitor with a resolution of 1920 x 1200, coupled with a Logitech Extreme 3D Pro joystick. Participants engaged in a simulated surveillance mission, where they assumed the role of drone operators. Each participant undertook a series of 100 dual-task trials. Trials consisted of a compensatory tracking task and a threat detection task (Yang et al., 2017; Du et al., 2020a; Schuler and Yang, 2023). In the compensatory tracking task, participants were required to maintain a moving target within a specified region. Concurrently, the threat detection task involved identifying and flagging potential threats . Participants were restricted to viewing a single task display at a time (see Figure 3.2). Each trial was set to a duration of 10 seconds and commenced upon the tracking task display. Participants were provided the flexibility to transition between task displays anytime during a trial. If a participant chose to transition between task displays, a 0.5-second delay was introduced. A preparatory 3-second countdown preceded each trial. At the beginning of each trial, an automated aid recommended the presence or absence of threats. A scoring system was implemented based on prior work Du et al. (2020a), where participants could earn up to 15 points per trial. After each trial, participants received feedback about accuracy and performance. After the feedback for each trial, participants rated their trust in the automated aid with a slider bar.

3.3.2.1 Tracking Task

The tracking task was programmed based on the compensatory tracker task in the Psychology Experiment Building Language (PEBL) (Mueller and Piper, 2014). Participants used a

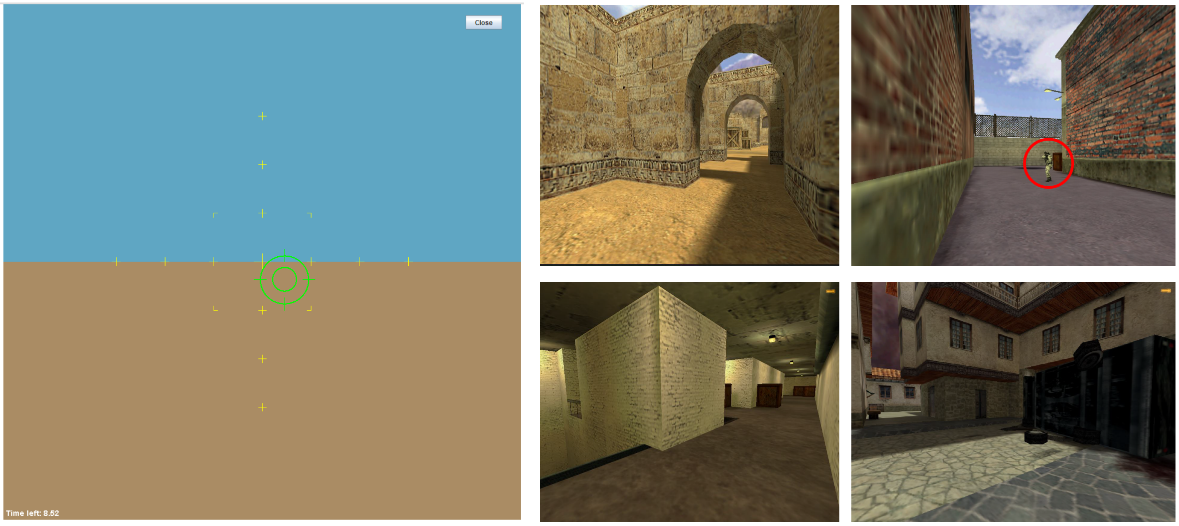


Figure 3.2: The tracking task display is on the left and threat detection images are on the right; a threat can be seen in the top right image (circled for illustrative purposes).

joystick to position a randomly drifting green circle over an immobile, central cross-hair. During each trial, up to 10 points could be acquired based on tracking task performance. Tracking error was quantified as the pixel distance between the cross-hair and circle; tracking data was collected at a sampling frequency of 20 Hz. Subsequently, the Root Mean Square Error (RMSE) for the tracing task was calculated as $\sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(Tracking\ Error)^2}$, where $n = 200$. The tracking score was calculated using a 10-bin histogram of the RMSE distribution determined from prior research with this test-bed Yang et al. (2017).

3.3.2.2 Threat Detection Task

Participants could *choose* to switch task displays and view the threat detection images; this allowed participants to cross-check the automation's recommendation with AVI. The images always correctly displayed the status of threats. Threats were presented as human figures and only one threat was shown at a time. No distraction stimuli were presented and participants were instructed to immediately report threats using the joystick. Threats were uniformly distributed across the four images. A maximum of 5 points could be earned for each threat detection task and a linear point penalty was applied based on each participant's decision-making response time. When participants correctly detected a threat, they earned a score of $5 - 5 \times \frac{detection\ time}{10}$; when participants correctly identified the absence of a threat, they earned 5 points. Participants received 0 points for an incorrect detection.

3.3.2.3 Automated Aid

Participants were supported by an imperfect automated aid for the threat detection task. The automation provided audio and visual alarms to recommend if threats were present or absent. However, participants were ultimately responsible for reporting threats, and would report by using the joystick's trigger. We bench-marked d' and c following the study of Wiczorek and Manzey (2014) (Table 3.1). Automation likelihood information was not disclosed to participants.

Table 3.1: Alarm characteristics of the 5 experimental conditions. There were 25 participants in each condition.

	Alarm System Characteristics								
Group ID	Base Rate	d'	c	PPV	NPV	Hit	Miss	FA	CR
G1	0.1	-1.09	-0.29	0.18	0.96	8	2	36	54
G2	0.2	-1.09	-0.29	0.33	0.92	16	4	32	48
G3	0.3	-1.09	-0.29	0.46	0.87	24	6	28	42
G4	0.4	-1.09	-0.29	0.57	0.82	32	8	24	36
G5	0.5	-1.09	-0.29	0.67	0.75	40	10	20	30

3.3.3 Design of Experiment

There were two independent variables: automation reliability level, a between-subjects factor shown in Table 3.1, and experimental quarter, a within-subjects factor measuring participants' interaction experience with the automated aid. Our study focused on the influence of repeated automation use; we divided the experimental session (i.e., 100 trials) into 4 quarters: Q1 = trials 1-25, Q2 = trials 26-50, Q3 = trials 51-75, and Q4 = trials 76-100.

3.3.4 Measures

The dependent variables included operators' dependence behaviors, decision-making strategies, trust, and dual-task performance measures.

3.3.4.1 Compliance, Reliance, and Cross-checking Behaviors

Upon receiving an automation recommendation, participants could choose one of the three responses: blindly follow the recommendation, blindly reject the recommendation, or cross-check using the four threat detection images. We measured operators' blind dependence and cross-checking behaviors for each trial. The operator's blind dependence was defined as the probability that the human blindly follows the automation's recommendation without manually cross-checking the alarm validity information. We separated the dependence behaviors based on the alarm state, an operator could comply with an alarm and rely on automation's non-alarm. We calculated blind dependence behaviors using the following equations:

$$\begin{aligned} &\textit{Blind Compliance} = P(\textit{report and not cross-checking} \mid \textit{automation alert}) \\ &\textit{Blind Reliance} = P(\textit{not report and not cross-checking} \mid \textit{automation slience}) \end{aligned} \tag{3.6}$$

Blind non-compliance and non-reliance behaviors refer to the situation when operators blindly reject automation's recommendation *without* cross-checking. The blind non-compliance and blind non-reliance were calculated with the following equations:

$$\begin{aligned} &\textit{Blind Non-Compliance} = P(\textit{not report and not cross-checking} \mid \textit{automation alert}) \\ &\textit{Blind Non-Reliance} = P(\textit{report and not cross-checking} \mid \textit{automation silence}) \end{aligned} \tag{3.7}$$

We also measured participants' cross-checking behaviors as:

$$\begin{aligned} &\textit{Alarm Cross-Checking} = P(\textit{cross-checking} \mid \textit{automation alert}) \\ &\textit{Non-Alarm Cross-Checking} = P(\textit{cross-checking} \mid \textit{automation silence}) \end{aligned} \tag{3.8}$$

3.3.4.2 Decision-Making Strategies

In addition to the dependence and cross-checking behaviors, we captured operators' decision-making strategies, specifically their extreme response strategies. Following prior research (Manzey et al., 2014; Yang, 2024), an extreme response strategy refers to when a participant uses the same strategy more than 90% of the time. The following definitions categorized the extreme automation strategy used:

- Extreme compliance strategy: Participants blindly comply with more than 90% of the alarms (i.e., *blind compliance* $\geq$ 90%)
- Extreme non-compliance strategy: Participants blindly reject more than 90% of the alarms (i.e., *blind non-compliance* $\geq$ 90%)
- Extreme alarm cross-checking strategy: Participants cross-check more than 90% of the alarms (i.e., *alarm cross-checking* $\geq$ 90%)
- Extreme reliance strategy: Participants blindly rely on more than 90% of the non-alarms (i.e., *blind reliance* $\geq$ 90%)
- Extreme non-reliance strategy: Participants blindly reject more than 90% of the non-alarms (i.e., *blind non-reliance* $\geq$ 90%)
- Extreme non-alarm cross-checking strategy: Participants cross-check more than 90% of the non-alarms (i.e., *non-alarm cross-checking* $\geq$ 90%)

3.3.4.3 Trust

Trust was measured immediately after each trial using a visual analogue scale from 1-100 (Wiczorek and Manzey, 2014; Bhat et al., 2023; Guo et al., 2023).

3.3.4.4 Performance

Participants could earn up to 15 points for each trial, with 10 points for the tracking task and 5 for the detection task. The cumulative dual-task score quantified total performance. Additionally, we calculated the human-automation team's response accuracy and separated the accuracy across alarm states.

3.3.5 Procedure

We recruited participants to the laboratory through email communication. After completing the informed consent form, each participant received a video orientation that explained the drone surveillance mission, tasks, and the scoring system. Each participant then completed 30 practice trials of only the tracking task, followed by 8 practice dual-task trials. During the practice session, a researcher monitored participants and answered questions raised by participants. Participants then started the experiment and took a mandatory 5-minute break halfway through the experiment.

3.3.6 Data Analyses

All statistical analyses were conducted using R (version 4.3.1). The analyses compared dependence behaviors, decision-making strategies, trust, and performance across the five experimental conditions as participants gained more interaction experience. This was quantified by separating the 100 trials into quarters: Q1 = trials 1-25, Q2 = trials 26-50, Q3 = trials 51-75, and Q4 = trials 76-100; experimental quarter was a within-subjects factors.

Two-way ANOVAs, implemented through the 'aov' package, were performed with automation reliability (i.e., PPV and NPV) and quarters as predictors. In instances of an interaction effect between the two predictors, a simple effect analysis was conducted. Post-hoc Bonferroni tests were employed when significant differences were found across quarters to capture temporal differences on the dependent variable. A one way ANOVA was conducted to compare total performance across reliability conditions and was followed up with a post-hoc Bonferroni test. Additionally, McNemar's test (Eliasziw and Donner, 1991) was conducted to compare proportions of extreme strategy counts between the beginning (Q1) and the end of the experiment (Q4). A continuity correction was applied to each group and the α level for all statistical tests was set at 0.05.

3.4 Results

First, we calculated the optimal strategy for our experiment and determine the optimal behaviors for operators in each reliability condition. Next we dissected the experimental data across four quarters to examine how operators, leveraging automated aids with different reliability levels, adjusted their behaviors and strategies. Finally, we illustrated how adjustments influenced dual-task performance.

3.4.1 The Optimal Strategy for Our Study

In section 3.2, we proposed the ideal strategy which incorporates alarm validity information and below we will calculate the decision making thresholds for the experiment. To calculate the average cost of cross-checking (C_{cc}), we first separated trials into two groups, based on if participants cross-checked or not. Next, we averaged the tracking task scores (per trial) for each group. Finally, we calculated the difference between the groups to calculate the cost of cross-checking (C_{cc}) per trial (see Equation 3.9). For the 12,500 trials, 100 trials for each of 125 participants, cross-checking caused an average point loss of 3.97 points per trial.

$$C_{cc} = \frac{\Sigma \textit{ tracking score} \mid \textit{no cross-check}}{\textit{total trials no cross-check}} - \frac{\Sigma \textit{ tracking score} \mid \textit{cross-check}}{\textit{total trials cross-check}} \tag{3.9}$$

Participants would gain 5 points for correctly identifying a signal (V_{hit}) and were not subjected to a penalty for a miss (C_{miss}).

For the ideal strategy, the human should cross-check if $EU(cc|alarm) > EU(y|alarm)$ and $EU(cc|alarm) > EU(n|alarm)$; thus, the operator should cross-check when:

$$\begin{aligned}
PPV &> \frac{C_{cc}}{(V_{hit} + C_{Miss})} \\
PPV &> \frac{3.97}{5} \\
PPV &> 0.794
\end{aligned} \tag{3.10}$$

or when:

$$\begin{aligned}
PPV &< 1 - \frac{C_{cc}}{(V_{CR} + C_{FA})} \\
PPV &< 1 - \frac{3.97}{5} \\
PPV &< 0.206
\end{aligned} \tag{3.11}$$

These calculations indicate that cross-checking was too costly because the lower bound was higher than the upper bound. Therefore, the theoretical optimal behavior for the experiment was to always blindly use or ignore the automation (i.e., the extreme responding strategy), depending on the automation's performance. To calculate the threshold between when extreme use and disuse are recommended, we used the formula below and inserted values from the experiment - value of Hit and CR were 5, and cost of incorrect automation was 0.

$$P(y|alarm) = \begin{cases} 1 \quad when\ PPV > \frac{5+0}{(5+0+5+0)} \\ 0 \quad when\ PPV < \frac{5+0}{(5+0+5+0)} \end{cases}$$

Therefore, the optimal behaviors would be extreme blind compliance when $PPV > 0.5$ (meaning for groups 4 and 5) and extreme blind non-compliance when $PPV < 0.5$ (groups 1, 2, and 3), see Table 3.1. Extending to non-alarm states, the optimal behavior would be to blindly rely on the automation, since all NPV's were above 0.5.

3.4.2 Compliance, Non-Compliance, and Alarm Cross-Checking

PPV had a significant effect on compliance, $F(1, 498) = 55.94$, $p < 0.001$. Quarter, however, exhibited no significant effect on compliance (p=0.78). An interaction effect emerged between PPV and quarter concerning compliance ($p < 0.001$), prompting separate ANOVAs for each variable. Operators in groups 3, 4, and 5 increased blind compliance as the experiment progressed, while groups 1 and 2 decreased. Except for group 3, all blind compliance rates were converging towards respective optimal values.

PPV significantly influenced non-compliance, $F(1, 498) = 79.34$, $p < 0.001$, while quarters also had a significant effect, $F(3, 496) = 6.03$, $p < 0.001$. A significant interaction effect between PPV and quarter was observed ($p < 0.001$), leading to separate ANOVAs for each variable. Post-hoc Bonferroni tests revealed significant differences in non-compliance between Q1-Q3 ($p = 0.01$) and Q1-Q4 ($p < 0.01$). As the experiment progressed, the two groups with the lowest PPV levels experienced an increase in non-compliance, while the other three groups maintained similar levels throughout the experiment.

In terms of cross-checking, PPV showed a marginally significant effect onto cross-checking rates, $F(1, 492) = 2.77$, $p = 0.097$, while quarter had a significant effect, $F(3, 492) = 3.71$, p =0.01.

3.4.3 Reliance, Non-Reliance and Non-Alarm Cross-Checking

NPV was found to have a significant effect on reliance, $F(1, 492) = 16.39$, $p < 0.001$. Quarter also had a significant effect onto reliance, $F(3, 492) = 4.40$, $p = 0.005$. A post-hoc Bonferroni test found significant differences in reliance between Q1-Q4 ($p < 0.01$). Throughout the experiment, operators across all groups increased their reliance behaviors.

NPV exerted a significant effect on blind non-reliance, $F(1, 492) = 12.45$, $p < 0.001$, while quarter was not found to be significant ($p = 0.82$).

NPV, $F(1,\ 492) = 10.06$, $p = 0.002$, and quarter, $F(3,\ 492) = 4.14$, $p = 0.006$ were significant predictors of non-alarm cross-checking rates.

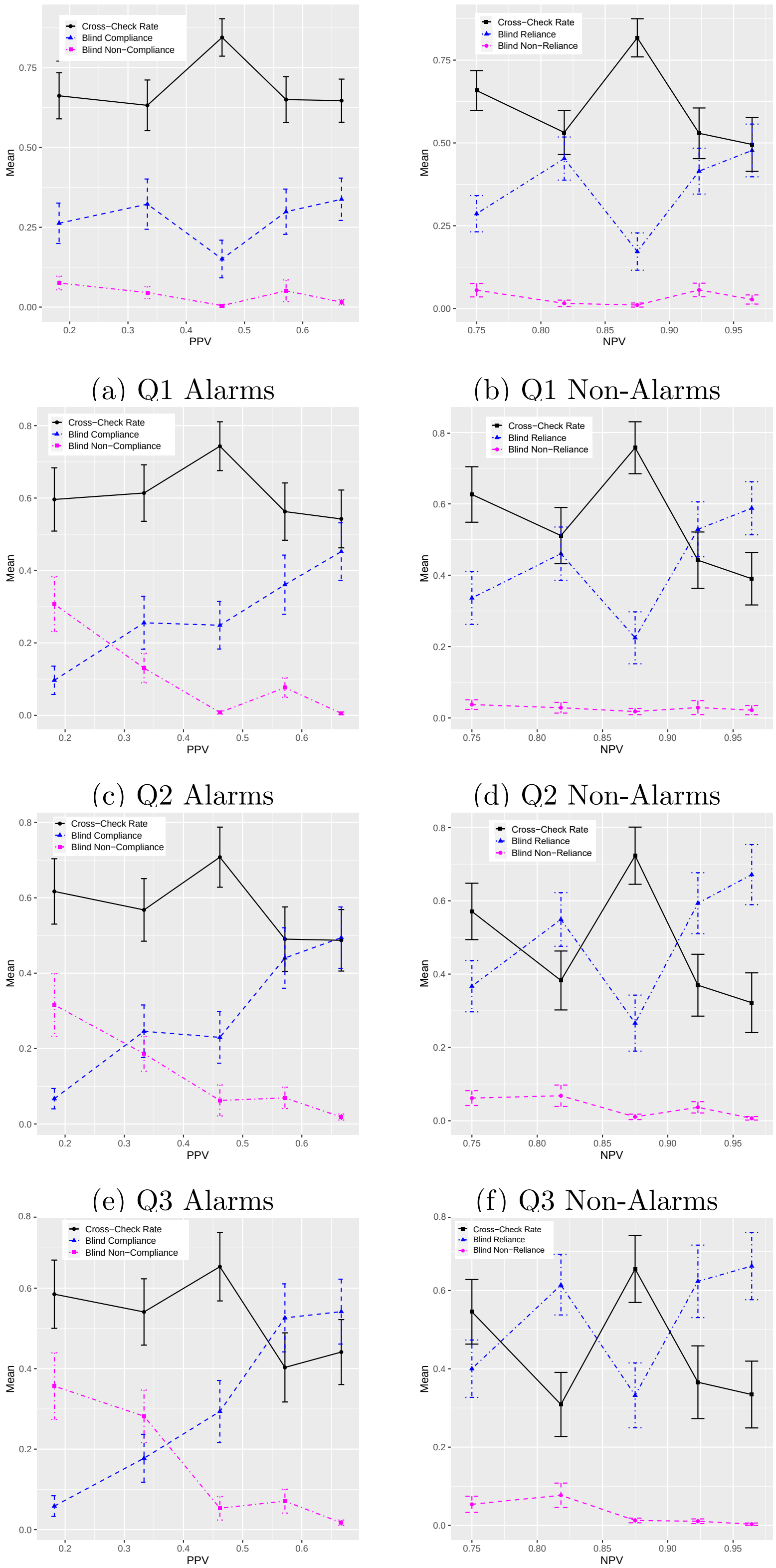
3.4.4 Trust

Trust in automation increased as the reliability increased, $F(1,\ 492) = 3.29$, $p = 0.01$. There was no significant effect of quarters on trust (p=0.93).

Descriptive data for our results are presented in Table 3.2. This is complemented by Figure 3.3, which illustrates graphs of dependence behaviors plotted against varying degrees of automation reliability. The graphs are organized into two distinct columns, the left column displays alarm response behaviors across Positive Predictive Values (PPVs), while the right column displays non-alarm response behaviors across Negative Predictive Values (NPVs). Each row corresponds to a distinct quarter.

Table 3.2: Mean (SE) values for dependence, cross-checking, and trust.

		Reliability	Trust	PPV	Blind	Blind	Alarm	NPV	Blind	Blind	Non-Alarm
					Compliance	Non-Compliance	Cross-Checking		Reliance	Non-Reliance	Cross-Checking
G1	Q1	0.62	0.48 (0.23)	0.18	0.26 (0.32)	0.08 (0.10)	0.66 (0.36)	0.96	0.47 (0.40)	0.03 (0.07)	0.50 (0.41)
	Q2	0.62	0.51 (0.27)	0.18	0.10 (0.20)	0.31 (0.38)	0.60 (0.44)	0.96	0.59 (0.37)	0.02 (0.07)	0.39 (0.37)
	Q3	0.62	0.52 (0.28)	0.18	0.07 (0.13)	0.32 (0.42)	0.62 (0.43)	0.96	0.67 (0.40)	0.01 (0.02)	0.32 (0.41)
	Q4	0.62	0.48 (0.29)	0.18	0.06 (0.13)	0.36 (0.41)	0.59 (0.42)	0.96	0.66 (0.43)	0.00 (0.02)	0.33 (0.43)
G2	Q1	0.64	0.55 (0.21)	0.33	0.32 (0.39)	0.05 (0.10)	0.63 (0.40)	0.92	0.42 (0.35)	0.06 (0.10)	0.53 (0.38)
	Q2	0.64	0.51 (0.28)	0.33	0.26 (0.36)	0.13 (0.20)	0.61 (0.39)	0.92	0.53 (0.39)	0.03 (0.10)	0.44 (0.40)
	Q3	0.64	0.48 (0.31)	0.33	0.26 (0.35)	0.19 (0.23)	0.57 (0.42)	0.92	0.59 (0.42)	0.04 (0.08)	0.37 (0.42)
	Q4	0.64	0.47 (0.33)	0.33	0.18 (0.30)	0.28 (0.32)	0.54 (0.42)	0.92	0.62 (0.4)	0.01 (0.03)	0.37 (0.47)
G3	Q1	0.66	0.60 (0.17)	0.46	0.15 (0.29)	0.00 (0.02)	0.85 (0.29)	0.87	0.17 (0.28)	0.01 (0.03)	0.81 (0.29)
	Q2	0.66	0.57 (0.23)	0.46	0.25 (0.33)	0.00 (0.02)	0.74 (0.34)	0.87	0.23 (0.36)	0.02 (0.04)	0.76 (0.36)
	Q3	0.66	0.53 (0.26)	0.46	0.23 (0.34)	0.06 (0.20)	0.71 (0.40)	0.87	0.27 (0.48)	0.01 (0.04)	0.72 (0.39)
	Q4	0.66	0.52 (0.28)	0.46	0.29 (0.39)	0.05 (0.15)	0.65 (0.43)	0.87	0.33 (0.41)	0.01 (0.03)	0.66 (0.43)
G4	Q1	0.68	0.57 (0.16)	0.57	0.30 (0.36)	0.05 (0.17)	0.65 (0.36)	0.82	0.45 (0.32)	0.02 (0.05)	0.53 (0.33)
	Q2	0.68	0.59 (0.20)	0.57	0.36 (0.41)	0.07 (0.13)	0.56 (0.40)	0.82	0.46 (0.38)	0.03 (0.08)	0.51 (0.39)
	Q3	0.68	0.57 (0.23)	0.57	0.44 (0.40)	0.07 (0.14)	0.49 (0.43)	0.82	0.55 (0.37)	0.07 (0.15)	0.38 (0.40)
	Q4	0.68	0.58 (0.26)	0.57	0.53 (0.42)	0.07 (0.15)	0.40 (0.43)	0.82	0.61 (0.38)	0.08 (0.15)	0.31 (0.41)
G5	Q1	0.70	0.55 (0.16)	0.67	0.34 (0.33)	0.02 (0.04)	0.65 (0.34)	0.75	0.29 (0.27)	0.06 (0.10)	0.66 (0.30)
	Q2	0.70	0.58 (0.19)	0.67	0.45 (0.40)	0.01 (0.02)	0.54 (0.40)	0.75	0.34 (0.37)	0.04 (0.07)	0.63 (0.39)
	Q3	0.70	0.60 (0.18)	0.67	0.49 (0.41)	0.02 (0.04)	0.49 (0.41)	0.75	0.38 (0.35)	0.06 (0.10)	0.57 (0.38)
	Q4	0.70	0.61 (0.19)	0.67	0.54 (0.40)	0.02 (0.03)	0.44 (0.40)	0.75	0.40 (0.37)	0.05 (0.10)	0.55 (0.41)



(a) Q1 Alarms

(b) Q1 Non-Alarms

(c) Q2 Alarms

(d) Q2 Non-Alarms

(e) Q3 Alarms

(f) Q3 Non-Alarms

(g) Q4 Alarms

(h) Q4 Non-Alarms

Figure 3.3: Left column displays alarm responses across PPV; right column displays non-alarm responses across NPV levels.

3.4.5 Strategy Counts

We tallied participants who employed extreme strategies throughout the experiment, as detailed in Table 3.3.

Table 3.3: Participant counts (out of 25) for the different strategies, separated by quarters.

		Alarms			Non-Alarms		
	Quarter	Extreme Blind Compliance	Extreme Cross-Checking	Extreme Blind Non-Compliance	Extreme Blind Reliance	Extreme Cross-Checking	Extreme Blind Non-Reliance
G1	Q1	0	11	0	6	5	0
	Q2	0	11	3	8	3	0
	Q3	0	12	7	15	4	0
	Q4	0	10	6	15	5	0
G2	Q1	3	11	0	1	7	0
	Q2	2	9	0	8	6	0
	Q3	2	7	0	9	5	0
	Q4	1	7	3	14	7	0
G3	Q1	1	18	0	1	17	0
	Q2	0	16	0	4	16	0
	Q3	2	15	0	4	14	0
	Q4	4	13	0	6	12	0
G4	Q1	3	10	0	3	5	0
	Q2	5	9	0	4	8	0
	Q3	6	9	0	5	5	0
	Q4	8	8	0	7	5	0
G5	Q1	2	9	0	1	7	0
	Q2	6	8	0	3	12	0
	Q3	6	7	0	3	9	0
	Q4	7	7	0	3	7	0

3.4.6 Extreme Alarm Response Strategies

An insufficient number of participants engaged in the extreme blind compliance or extreme blind non-compliance strategies to meet required assumptions for statistical testing. While the initial count of operators opting for extreme blind compliance was relatively consistent across groups (0-3 individuals), noteworthy shifts towards this strategy were observed in groups with higher PPVs. Groups 4 and 5 eventually saw just under 30% of participants employing the extreme blind compliance strategy.

A statistically significant difference in extreme cross-checking counts emerged between the experiment's commencement and conclusion ($\tilde{\chi}$ (1) = 5.63, p = 0.03). Group 3 made the largest shift, with 18 out of 25 participants who initially applied the extreme cross-checking strategy and 5 operators shifted away from this non-optimal strategy. The extreme blind non-compliance strategy was exclusively observed in groups 1 and 2, the number of participants adopting this strategy increased as the experiment unfolded.

3.4.7 Extreme Non-Alarm Response Strategies

A significant difference in the number of participants using an extreme blind reliance strategy was observed between the experiment's initiation and conclusion $\tilde{\chi}$ (1) =27.68, $p < 0.001$. The prevalence of individuals adopting an extreme reliance strategy increased throughout the experiment. Furthermore, groups with high NPV levels exhibited a higher frequency of operators resorting to extreme blind reliance than when NPV was lower.

In contrast, there was no statistically significant difference in extreme non-alarm cross-checking counts between the experiment's commencement and conclusion ($\tilde{\chi}$ (1) = 0.41, p = 0.52). Notably, group 3 stood out as the sole group with multiple participants discontinuing the extreme cross-checking strategy.

Participants did not use an extreme blind non-reliance strategy during the experiment.

3.4.8 Performance

Total performance scores had significant differences across both automation reliability levels and quarters (respectively, $F(1,\ 492) = 25.18$, $p < 0.001$ and $F(3,\ 492) = 23.81$, $p < 0.001$). Each group improved their total performance throughout the experiment, with the improvement spanning 40-50 points between the experiment's commencement and conclusion. A significant difference was found between groups ($F(1,\ 123) = 7.25$, $p = 0.008$), however, post-hoc testing found that groups 1 and 2 performed significantly better than groups 3, 4, and 5 ($p < 0.05$).

A significant difference in detection scores across reliability levels emerged ($F(1, 498) = 115.10$, $p < 0.001$). Conversely, quarter did not significantly influence detection scores ($p = 0.93$). Threat detection scores had a significant interaction effect between reliability and quarter ($p = 0.01$), prompting separate ANOVAs for each variable.

Detection task accuracy was separated by alarm state (i.e., alarms and non-alarms). Regarding detection accuracy during alarms, neither PPV, ($p = 0.89$), nor quarter, ($p = 0.65$), were found to be significant predictors of alarm response accuracy. When examining alarm accuracy changes over quarters, groups 1 and 2 improved during the experiment, while the other groups worsened. A significant interaction effect was observed between automation PPV and quarter ($p = 0.01$), leading to separate ANOVAs for each variable.

NPV significantly impacted detection task accuracy during non-alarms, ($F(1, 498) = 68.28$, $p < 0.001$), with higher response accuracy in conditions with better NPV. No significant effect of quarter on detection accuracy during non-alarms was identified ($p = 0.76$). However, group 2 improved while groups 4 and 5 worsened. Detection task accuracy during non-alarms had a significant interaction effect between automation NPV and quarter ($p = 0.0489$), leading to separate ANOVAs for each variable.

For tracking scores, a significant difference emerged across quarters ($F(3, 492) = 19.17$, $p < 0.01$). Tracking scores did not significantly differ across reliability groups ($p = 0.11$). All groups demonstrated tracking task improvement throughout the experiment.

Participants significantly increased the proportion of time spent on the tracking task across quarters ($F(3, 492) = 9.19$, $p < 0.001$). Conversely, no significant effect of automation reliability on proportion was identified ($p = 0.57$). Detailed performance data, organized by groups and quarters, is presented in Table 3.4.

Table 3.4: Mean and SD values for total performance scores, detection task score and accuracy, tracking task score, and the proportion of time spent on the tracking task display.

	Quarter	Total	Detection Task	Detection Task	Detection Task	Tracking Task	Tracking Task
		Score	Score	Accuracy (Alarms)	Accuracy (Non-Alarms)	Score	Proportion
G1	Q1	247.34 (56.63)	103.86 (19.30)	0.73 (0.30)	0.94 (0.09)	143.48 (61.27)	0.88
	Q2	281.95 (60.66)	108.99 (12.11)	0.81 (0.18)	0.95 (0.07)	172.96 (61.54)	0.91
	Q3	301.02 (56.90)	111.46 (6.74)	0.83 (0.13)	0.96 (0.09)	189.56 (57.40)	0.92
	Q4	297.24 (57.55)	111.84 (9.02)	0.86 (0.14)	0.96 (0.09)	185.40 (59.36)	0.92
G2	Q1	254.07 (43.30)	94.31 (21.07)	0.72 (0.24)	0.86 (0.13)	159.76 (50.76)	0.88
	Q2	271.22 (40.52)	95.50 (15.32)	0.68 (0.24)	0.90 (0.10)	175.72 (46.84)	0.91
	Q3	289.00 (38.29)	98.48 (15.41)	0.71 (0.25)	0.94 (0.05)	190.52 (46.96)	0.92
	Q4	296.79 (38.08)	102.03 (13.53)	0.76 (0.23)	0.93 (0.06)	194.76 (43.10)	0.92
G3	Q1	225.63 (48.49)	103.11 (12.18)	0.86 (0.15)	0.93 (0.10)	122.52 (53.27)	0.85
	Q2	248.89 (43.15)	99.01 (14.28)	0.80 (0.20)	0.93 (0.09)	149.88 (51.26)	0.87
	Q3	265.96 (47.32)	99.44 (12.05)	0.79 (0.20)	0.93 (0.10)	166.52 (51.70)	0.88
	Q4	272.14 (42.24)	96.26 (15.79)	0.76 (0.22)	0.92 (0.09)	175.88 (49.65)	0.90
G4	Q1	228.19 (49.83)	91.43 (14.09)	0.78 (0.15)	0.84 (0.15)	136.76 (57.95)	0.89
	Q2	246.07 (50.13)	92.99 (10.87)	0.76 (0.15)	0.89 (0.12)	153.08 (56.59)	0.90
	Q3	268.61 (46.32)	87.89 (16.37)	0.70 (0.25)	0.83 (0.14)	180.72 (57.33)	0.93
	Q4	270.28 (47.91)	85.28 (16.40)	0.70 (0.19)	0.76 (0.18)	185.00 (56.09)	0.94
G5	Q1	234.50 (39.58)	90.78 (14.75)	0.83 (0.17)	0.87 (0.15)	143.72 (40.81)	0.87
	Q2	255.26 (48.09)	91.78 (10.45)	0.84 (0.12)	0.82 (0.16)	163.48 (50.94)	0.90
	Q3	268.59 (43.33)	87.89 (15.35)	0.78 (0.15)	0.81 (0.18)	181.36 (48.65)	0.91
	Q4	275.53 (44.12)	87.33 (11.92)	0.78 (0.15)	0.84 (0.13)	188.20 (49.79)	0.92

3.5 DISCUSSION

We conducted a laboratory experiment to examine the influence of imperfect automation on operators' dependence behaviors, response strategies, and total human-automation performance. We anticipated that participants would change their behaviors to account for the automation's errors. To explore this prediction, we divided the experiment into four quarters and compared dependence behaviors, cross-checking rates, extreme strategy counts, trust, and dual-task performance across the reliability levels.

Furthermore, we introduced a theoretical optimal behavior incorporating the human's ability to choose when to validate an alarm. Experimental data was utilized to calculate the theoretical optimal behaviors specific to our dual-task scenario, with the identified optimal strategy aligning with the extreme response strategy.

3.5.1 Dependence Behaviors and Response Rates

As anticipated, automation performance significantly influenced corresponding operator dependence behaviors, aligning with prior research (Manzey et al., 2014; Du et al., 2020a; Meyer, 2001). Alarm performance exhibited a positive linear relationship with compliance behaviors, and this trend was mirrored in the reliance and non-alarm relationship, except for participants in group 3. Participants in this group utilized the alarm validity information at much higher rates when compared to other groups. This potentially stems from perceived ambiguity in the automation's performance and could lead to increased operator uncertainty and cross-checking. This result partially aligns with the results of Manzey et al. (2014), specifically in their third and fourth experiments. In those experiments, the authors systematically manipulated the base rate to control the PPV and NPV, while also increasing the effort required to cross-check (experiment 3) or workload (experiment 4). The trend in their alarm cross-checking results resembled an inverted 'u-shape', where cross-checking was highest when the PPV was 0.3. The tendency of cross-checking being a major response (i.e., cry-wolf phenomena) was also found in accordance with other previous work (Bliss and Dunn, 2000).

Significant changes in dependence behaviors occurred across our experiment, with most adjustments occurring in the first half, as operators were building their initial expectations of the automation. Both increased and decreased compliance were observed depending on alarm performance, resulting in compliance rates shifting towards optimal behaviors for each group. Reliance rates also significantly improved over time, with all groups increasing reliance towards our proposed optimal strategy. Notably, when automation alarm performance was poor (e.g., PPV of 0.18 and 0.33), operators exhibited a considerable amount of blind non-compliance behavior, often ignoring alarms rather than cross-checking with alarm validity information. This finding varied from the results reported by Manzey et al. (2014), where participants with AVI access decreased the extent to which they ignored alarms. The blind non-compliance behavior is essentially the ‘cry wolf’ effect, where a high proportion of false alarms leads to operator distrust and subsequent ignoring of true alarms.

Our findings indicate participants altered their response behaviors across alarm and non-alarm states, particularly in the choice to blindly ignore alarms and blindly follow non-alarms. This independence between compliance and reliance behaviors aligns with empirical findings in the literature (Yang et al., 2017; Meyer, 2004; Meyer and Bitan, 2002; Dixon and Wickens, 2006).

3.5.2 Trust in the Automation

Our results revealed that participants had higher trust in automation with better reliability. However, trust adjustments across experimental quarters were not significant, contrasting recent research on the dynamic nature of trust (De Visser et al., 2020; Yang et al., 2023). Possible explanations include our group-level aggregation of trust data, which may have masked individual-level variations. The trust increases or decreases by an individual could have been counter-balanced by adjustments made from other individuals, leading to a static appearance of group trust. Additionally, the use of a uni-variate trust measurement might oversimplify the nuanced nature of trust in the context of multiple alarm states; a uni-variate trust measurement attempted to capture trust in automation with multiple alarm states, this limitation was also noted in Du et al. (2020a). Future work could explore individual-level trust dynamics, considering personal factors known to influence trust in human-automation teams (Hoff and Bashir, 2015; Hancock et al., 2011, 2021; Schaefer et al., 2016).

3.5.3 Participant Strategies

Participants not only adjusted their response behaviors to the automation but also changed their strategies, and their strategy adjustments aligned with the automated aid’s performance. This aligns with findings from a dual-task experiment by Manzey et al. (2014) where alarm characteristics were manipulated similarly. More participants resorted to extreme reliance strategies than extreme compliance strategies in our data, reflecting the superior performance of the automation during non-alarm states.

Analysis of extreme strategy counts across quarters revealed ongoing adaptation throughout the experiment, with a more substantial shift in the first half. Individuals continued to adapt their strategies towards optimal in the last quarter. While this adaptation trend was observed, further research is needed to explore factors influencing these shifts and the limitations of the proposed optimal strategy. Additionally, the proposed optimal strategy could be extended to incorporate other scenario-specific factors, such as time restrictions, additional tasks, or the presence of multiple automated aids.

3.5.4 Performance

Our results indicated significant dual-task performance differences between reliability conditions, with group 1 ($M = 281.89$) and group 2 ($M = 277.52$) demonstrating better performance compared to the other groups (M respectively for groups 3, 4, 5 = 253.16, 253.29, 258.47). Our group with the highest performing automated aid did not exhibit the most proficient dual-task performance. Two groups with mid-range automation performance (66% and 68% reliability) recorded the lowest total human-automation performance. This contrasts previous research, where authors found that dual-task performance was better in conditions with more reliable automation (Walliser et al., 2016; Neyedli et al., 2011; Du et al., 2020a). This contrast may stem from our use of similar reliability levels across conditions (only changing by 2% between each condition), while the other studies had a much larger performance difference, ranging from a 10% - 30% between conditions. In addition, the discrepancy could be rooted in the way we manipulated our experimental base rate, which was also different than the mentioned studies, as our participants experienced a time penalty when reporting threats and no time penalty for correct non-threat decisions. Participants in conditions with a higher base rate would experience a greater cumulative time penalty, as the penalty is applied to trials with a signal (we cannot apply a time penalty for a correct non-action). Furthermore, this outcome could be linked to how uncertainty influenced participants' selection of strategies; groups with poorer total performance employed sub-optimal behaviors and strategies. More research is warranted to further explore the application of our proposed optimal strategy.

In our results, each group displayed a consistent overall improvement throughout the experiment, particularly in the tracking task, there emerged a significant difference in the threat detection performance among the groups. In our experiment, the overall performance improvement stems from focusing on the the tracking task, which in addition to dependence and cross-checking rates, is illustrated with the proportion measure. Groups 3, 4, and 5 regressed on the detection task score and accuracy throughout the experiment, resulting from heightened dependence on automation recommendations. In contrast, groups 1 and 2, despite also prioritizing detection, opted for blind non-compliance behaviors, effectively compensating for the sub-optimal alarm performance. Aligning with earlier experiments, participants consistently favored prioritizing tracking over threat detection (Du et al., 2020b; Wiczorek and Manzey, 2014), suggesting a strategic automation use to manage attention and improve non-automated task performance.

The group utilizing the most optimal extreme response strategies emerged as the top performing group. Nevertheless, further investigations are essential to delve deeper into these findings and to explore factors influencing the observed performance variations.

3.5.5 Limitations and Future Work

Despite efforts to balance the scoring trade-off for cross-checking (points gained and lost for using alarm validity information); the task difficulty and delay time penalty resulted in a higher cross-checking cost than anticipated. Future work could systematically manipulate cross-checking costs, calculate the optimal behavioral thresholds, and examine the relationship between operator strategy and total human-automation performance. We did not provide a priori automation information to participants, instead, participants were required to use trial-by-trial feedback to form opinions regarding automation performance. In many domains, experienced operators understand nuances of their task and the limitations of their automated aids. Our participants were all novices. Dividing the experiment into quarters was a trade-off decision (as opposed to halves, thirds, etc.), which required careful consideration of the alarm characteristics and of strategy operationalization. If aggregating trial data into larger set (such as the entire experiment), the results provide a coarse view of the human-automation interactions. If too few trials are selected for a set, the operationalization of strategies can artificially categorize behaviors as extreme. Future research is needed to explore alternatives, such as by using a smoothed average method. Finally, we examined a considerable amount of data at the group level; individual traits have been shown to largely influence trust (Schaefer et al., 2016; Hoff and Bashir, 2015), dependence (McBride et al., 2011), and operator strategy (Riley, 1996). Future research could examine similar research questions at the individual level.

3.5.6 Conclusions

A dual-task laboratory experiment was conducted to examine how operators adapt their dependence behaviors and strategies and they increase their automation use. The experiment was separated into four quarters and comparisons were made for dependence, response rates, extreme strategy counts, trust, and dual-task performance. We found that automation performance influenced dependence behaviors and response strategies. More specifically, operators used post trial feedback to adjust across alarm state; their behaviors and strategies were independently adapted to the automation's PPV and NPV. Additionally, we proposed an ideal optimal decision-making strategy that considers when operators have access to alarm validity information. The adjustments operators made in their automation use converged towards theoretical optimal levels, which is seen in their performance scores, as operators whose strategies began to converge on the ideal optimal strategy performed best.

CHAPTER 4

The Influence of Likelihood Information on Dependence Behaviors, Response Strategies, and Performance

Introduction

In order to address inappropriate automation use, researchers proposed disclosing automation performance information to operators as a design intervention. Researchers, such as McGuirl and Sarter (2006), Walliser et al. (2016), and Wang et al. (2009), found that likelihood information facilitated operators' dependence behaviors and improved total human-automation performance. Other investigations reported that operators used likelihood information sub-optimally (Bagheri and Jamieson, 2004; Fletcher et al., 2017; Wickens and Colcombe, 2007). Du et al. (2020a) proposed that the varying outcomes may stem from the specific information presentation to participants. Literature examining the role of likelihood information, specifically in frequency formats, indicated that disclosing natural frequencies can improve a decision-makers' sensitivity (Gigerenzer and Hoffrage, 1995; Hoffrage and Gigerenzer, 1998; McDowell and Jacobs, 2017).

The literature using natural frequency formats has focused on the single-task paradigm, we believe there is a reasonable expectation that frequency formats may also prove beneficial in dual-task paradigms. This chapter aims to investigate a design intervention focusing on the incorporation of likelihood information, specifically, to compare the effects of predictive values against a frequency format and a baseline condition, where no *a priori* information was provided.

4.0.1 Current Study

This chapter presents two human subject studies to investigate how operators are using likelihood information as they repeatedly interact with an imperfect automated aid. The first study consists of a data reanalysis of a previously published study (Du et al., 2020a), where they conducted a human-subjects experiment to address the observed inconsistencies within the literature. They presented participants with different types of likelihood information: Hit/CR rates, predictive values, and OSL. They examined how different likelihood information affected participant trust, dependence behaviors, and performance. We expand on their analyses by comparing operators' dependence and cross-checking behaviors, extreme response strategies, and performance between the first and second half of the experiment. We delved deeper into their data to explore the influence of likelihood information on operators and their adaptations to the imperfect automation.

The findings from the reanalysis served as the foundation for a laboratory experiment. We utilized the predictive values, which were identified as most informative in the reanalysis, and added a condition in where participants did not receive *a priori* information. Additionally, we explored the impact of an alternative form of predictive values - frequencies. Participants were assigned to one of three likelihood information conditions: baseline (no information provided), predictive values, and predictive values in a frequency format. The data was divided into four quarters and we examined participants' dependence and cross-checking rates, response strategies, and performance. The following sections detail the methods and outcomes of each experiment.

4.0.2 Contributions

The first study involved a reanalysis of an existing dataset, we compared operators' dependence and cross-checking behaviors, extreme response strategies, and performance between the first and second halves of the experiment. The influence of likelihood information on operators and their adaptations to the imperfect automation was explored, revealing that both likelihood information and reliability significantly affected how participants utilized the automation. Predictive values emerged as particularly useful to improve performance. The study highlighted that participants in all conditions adjusted their behaviors to the scenario and improved as the experiment progressed. However, a limitation of the relatively short experiment was noted, as participants may have still been adapting their behaviors at the end of the experiment.

To address this limitation, we conducted the second study, where we doubled the number of trials. Likelihood information was again manipulated, but with different conditions, including a baseline of no information, predictive values, and predictive values in a frequency format. By maintaining the same reliability levels, we were able to examine how the combination of likelihood information and experience affected operators' dependence and cross-checking behaviors, strategies, and performance.

Several themes emerged when considering both studies. Automation reliability exerted more influence on behaviors and strategies than likelihood information. When the automation's likelihood was perceived as clearly good or bad, participants were more likely to adjust their behaviors and strategies. Controlling for automation reliability revealed that participants with likelihood information made faster adjustments than those without information. Post-trial feedback played a crucial role, enabling participants without *a priori* likelihood information to perform similarly to those with information. The uncertainty around our selected PPV and NPV levels likely contributed to the challenge participants faced in determining how much they should follow, cross-check, or ignore the automated aid.

The insights gained from these findings offer guidance for engineers and system designers seeking to enhance operators' dependence behaviors. Such improvements can, in turn, contribute to enhanced human-automation performance.

4.1 Effects of presenting different likelihood information on operators' adaptation

Study 4.1 reanalyzed data from an existing dataset (Du et al., 2020a), where participants complete 2 blocks of 50 trials and randomly alternated the automation reliability levels (74% or 90%). The automation likelihood information provided to participants was a between-subjects variable and conditions were: Hit/CR rates, OSL, or the predictive values.

Differing from their analyses, we compared the dependence behaviors, response strategies, and performance between the first half of the experiment and the second half.

4.1.1 Participants

Data from the existing set was gathered from 61 (25 males and 36 females) university students (mean age = 22.28 years old, *SD* = 4.88).

4.1.2 Apparatus and Stimuli

The same experimental test-bed was used from Chapter 3, refer to section 3.3.2.

4.1.3 Automated Aid

Participants were assisted by an imperfect automated system that was designated for the threat detection task. This automation indicated the presence or absence of threats through both auditory and visual alarms. However, participants were ultimately responsible for identifying and reporting threats, and would use the trigger function on the joystick to report. The alarm characteristics were manipulated by researchers and can be seen below in Table 4.1.

Table 4.1: Automation performance characteristics for reanalysis.

	Reliability (%)	Positive Predictive Value (%)	Negative Predictive Value (%)	Hits	False Alarms	Misses	Correct Rejections
Low Reliability c = -0.25, d’ = 1.5	74	54	92	13	11	2	24
High Reliability c = -0.25, d’ = 3.0	90	78	97	14	4	1	31

4.1.4 Experimental Design

The data reanalysis adopted a 3 (likelihood information: overall success likelihood, predictive values, and Hit/CR rates) × 2 (automation reliability: low vs high) × 2 (experimental half: first half vs second half) mixed design. We divided the length of the experiment (50 trials) into halves (H): H1 = trials 1-25, H2 = trials 26-50. Likelihood information was treated as between-subject factor; experimental half and automation reliability were within-subjects factors.

4.1.5 Measures

Our dependent variables included the operator’s dependence behaviors, decision-making strategies, and dual-task performance measures, which can be reviewed in section 3.3.4.

4.1.6 Data Analyses

All statistical analyses were carried out in R (version 4.3.1). Factorial analysis of variance (ANOVA) tests were executed to examine relationships between dependent and independent variables. In instances of statistical significance, a post-hoc Bonferroni test was performed.

Simple effect analyses were conducted when a significant interaction effect was found between independent variables.

To examine relationships between the counts of extreme strategies and the independent variables, McNemar's test Eliasziw and Donner (1991) was employed, comparing the first half (H1) to the second half (H2) of the experiment. A continuity correction was applied to each group. The alpha level for all statistical tests was set at 0.05.

4.1.7 Study 4.1 Results

One participant's data were excluded from our reanalyses because their tracking task performance was more than three standard deviations below the group's mean. This participant's data was excluded from the original study.

4.1.7.1 Dependence and Cross-Checking Behaviors

The descriptive data for all dependence behaviors and cross-checking rates are presented in Table 4.2 and are complemented by Figure 4.1, which illustrates mean and error plots of our dependent variables, separated by reliability and likelihood conditions. Figure 4.1 is organized so the left column displays alarm responses and the right column displays non-alarm responses. Each row corresponds to an experimental half.

Table 4.2: Mean ± standard error (SE) values for dependence, non-dependence, and cross-checking rates.

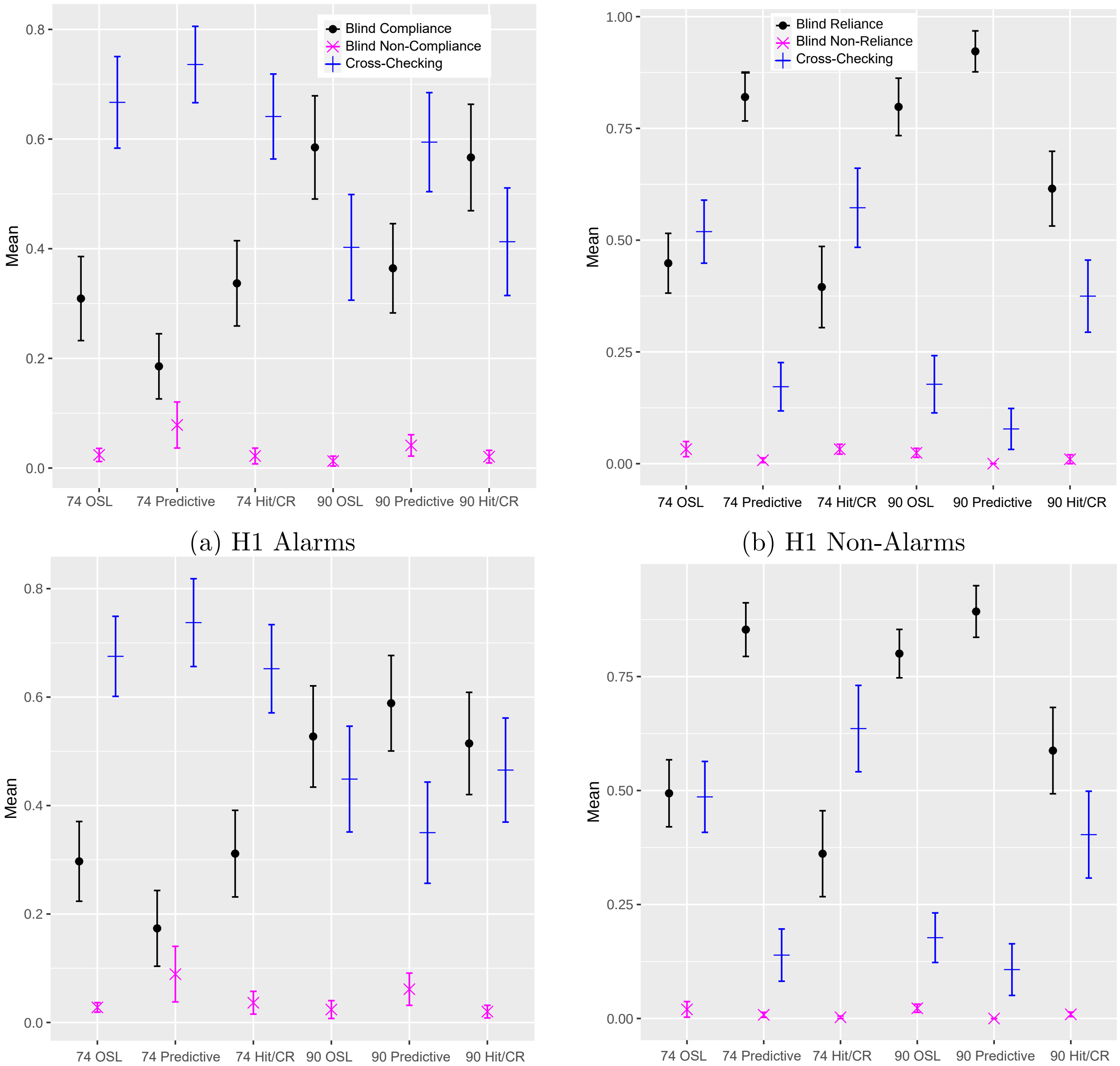
		Low Reliability, PPV = 0.54, NPV = 0.92			High Reliability, PPV = 0.78, NPV = 0.97		
		Overall Success Likelihood	Predictive Values	Hit/CR Rates	Overall Success Likelihood	Predictive Values	Hit/CR Rates
Blind Compliance	H1	0.31 ± 0.08	0.19 ± 0.06	0.34 ± 0.08	0.59 ± 0.09	0.36 ± 0.08	0.57 ± 0.10
	H2	0.30 ± 0.07	0.17 ± 0.07	0.31 ± 0.08	0.53 ± 0.09	0.59 ± 0.09	0.52 ± 0.09
Blind Non-Compliance	H1	0.02 ± 0.01	0.08 ± 0.04	0.02 ± 0.01	0.01 ± 0.01	0.04 ± 0.02	0.02 ± 0.01
	H2	0.03 ± 0.01	0.09 ± 0.05	0.04 ± 0.02	0.02 ± 0.02	0.06 ± 0.03	0.02 ± 0.01
Alarm Cross-Checking	H1	0.67 ± 0.08	0.74 ± 0.07	0.64 ± 0.08	0.40 ± 0.10	0.59 ± 0.09	0.41 ± 0.10
	H2	0.68 ± 0.07	0.74 ± 0.08	0.65 ± 0.08	0.45 ± 0.10	0.35 ± 0.09	0.47 ± 0.10
Blind Reliance	H1	0.45 ± 0.07	0.82 ± 0.05	0.40 ± 0.09	0.80 ± 0.06	0.92 ± 0.05	0.62 ± 0.08
	H2	0.49 ± 0.07	0.85 ± 0.06	0.36 ± 0.09	0.80 ± 0.05	0.89 ± 0.06	0.59 ± 0.09
Blind Non-Reliance	H1	0.03 ± 0.02	0.01 ± 0.01	0.03 ± 0.01	0.02 ± 0.01	0.00 ± 0.00	0.01 ± 0.01
	H2	0.02 ± 0.02	0.01 ± 0.01	0.00 ± 0.00	0.02 ± 0.01	0.00 ± 0.00	0.01 ± 0.01
Non-Alarm Cross-Checking	H1	0.52 ± 0.07	0.17 ± 0.05	0.57 ± 0.09	0.18 ± 0.06	0.08 ± 0.05	0.37 ± 0.08
	H2	0.49 ± 0.08	0.14 ± 0.06	0.64 ± 0.10	0.18 ± 0.05	0.11 ± 0.06	0.40 ± 0.10

4.1.7.2 Compliance, Blind Non-Compliance, and Alarm Cross-Checking

Increases in PPV led to increases in compliance rates, $F(1,\ 57) = 38.38$, $p < 0.001$. The main effects of likelihood information and experimental half were non-significant.

Likelihood information marginally affected blind non-compliance rates, $F(2,\ 57) = 2.45$, $p = 0.09$. Participants in the predictive values condition had increased blind non-compliance compared to participants in the OSL ($p = 0.08$) conditions. The effects of automation PPV and experimental half were non-significant.

Participant cross-checking rates increased as the automation's PPV decreased, $F(1,\ 57) = 32.76$, $p < 0.01$. The effects of likelihood information and experimental half were non-significant.



(a) H1 Alarms

(b) H1 Non-Alarms

(c) H2 Alarms

(d) H2 Non-Alarms

Figure 4.1: Left column displays alarm responses across conditions; right column displays non-alarm responses across conditions.

4.1.7.3 Reliance, Non-Reliance and Non-Alarm Cross-Checking

Reliance rates were significantly affected by automation NPV, $F(1, 57) = 35.34$, $p < 0.001$, and likelihood information, $F(2, 57) = 9.59$, $p < 0.001$. Reliance rates increased as NPV increased. A significant two-way interaction effect was found between likelihood information and NPV, $F(1,\ 56) = 4.77$, $p = 0.01$. Reliance in the predictive values condition was greater than the OSL condition ($p < 0.01$) and the OSL condition was greater than the Hit/CR condition ($p < 0.01$). The effects of experimental half were non-significant. Blind non-reliance rates were not significantly affected by the independent variables.

Non-alarm cross-checking rates were significantly affected by NPV, $F(1,\ 57) = 32.51$, $p < 0.001$, and likelihood information, $F(2,\ 57) = 8.88$, $p < 0.001$. Decreases in automation NPV led to increases in cross-checking rates (p=0.01). Non-alarm cross-checking in the predictive condition occurred less than the OSL condition ($p < 0.01$) and the OSL condition had significantly less than the Hit/CR condition ($p = 0.01$). A significant two-way interaction effect was found between likelihood information and NPV, $F(2,\ 57) = 4.78$, $p < 0.001$. The effects of experimental half were non-significant.

Extreme Response Strategies

We tallied participants who employed extreme strategies throughout the experiment, as detailed in Table 4.3.

Table 4.3: Illustrates participant counts (out of 20) for the different strategies, separated by experimental half.

		Low Reliability, PPV = 0.54, NPV = 0.92			High Reliability, PPV = 0.78, NPV = 0.97		
		Overall Success Likelihood	Predictive Values	Hit/CR Rates	Overall Success Likelihood	Predictive Values	Hit/CR Rates
Extreme Blind Compliance	H1	1	0	2	8	2	7
	H2	2	2	2	5	6	4
Extreme Blind Non-Compliance	H1	0	0	0	0	0	0
	H2	0	1	0	0	0	0
Extreme Alarm Cross-Checking	H1	10	8	7	6	7	5
	H2	9	11	9	7	5	7
Extreme Blind Reliance	H1	2	11	5	12	17	7
	H2	2	14	5	10	16	8
Extreme Blind Non-Reliance	H1	0	0	0	0	0	0
	H2	0	0	0	0	0	0
Extreme Non-Alarm Cross-Checking	H1	2	0	6	1	0	3
	H2	3	1	11	0	1	6

Automation's PPV appeared to affect how many participants used the extreme blind compliance strategy, as the high reliability conditions had more than three times the amount of participants than lower reliability. However, an insufficient number of participants resorted to this strategy in the low reliability condition to meet assumptions for statistical testing. The data illustrates that participants were changing strategies as the experiment progressed, but the trends in extreme compliance are unclear across experimental half or likelihood conditions.

An insufficient number of participants engaged in the extreme blind non-compliance or non-reliance strategies to meet required assumptions for statistical testing.

The number of participants using extreme cross-checking was significantly different across reliability conditions ($\tilde{\chi}$ (1) = 7.61, $p = 0.01$). More participants were applying the extreme cross-checking strategy in lower reliability condition than the high reliability condition.

The differences in counts of participants using the extreme blind reliance across reliability conditions were marginally significant ($\tilde{\chi}$ (1) = 3.34, $p = 0.07$). The number of participants using the extreme blind reliance strategy was higher for the predictive values likelihood condition were when compared to other likelihood conditions ($\tilde{\chi}$ (1) = 14.40, $p < 0.001$).

There were not enough participants using the extreme cross-checking during non-alarms for statistical testing. The data illustrate that the Hit/CR likelihood condition had more operators extreme cross-checking non-alarms that other conditions and that operators in this condition were adapting to cross-check more as the experiment progressed. The participants in this group also altered their strategy based on automation reliability, higher reliability resulted in less non-alarm cross-checking.

Performance

Participant's total performance score was significantly different across automation reliability ($F(1,\ 57) = 57.22$, $p < 0.001$), likelihood information, ($F(1,\ 57) = 3.89$, $p = 0.03$), and experimental half ($F(1,\ 57) = 9.64$, $p < 0.001$). Participants significantly improved as the experiment progressed ($p = 0.002$).

Regarding tracking task performance scores, significant main effects were found for automation reliability ($F(1,\ 57) = 7.27$, $p < 0.001$), likelihood information, ($F(2,\ 57) = 3.26$,

p= 0.046), and experimental half, ($F(2, 57) = 4.33$, $p = 0.042$) onto the tracking scores. Participants increased the proportion of time spent on the tracking task as automation reliability increased ($F(1, 57) = 62.23$, $p < 0.001$).

Automation reliability ($F(1, 57) = 4.80$, $p < 0.001$), and experimental half ($F(1, 57) = 7.94$, $p < 0.001$), had significant main effects onto the detection task score. Post hoc testing indicated that participants scored more detection task points in the high reliability condition compared to the low reliability condition ($p < 0.001$); additionally, participants improved their detection task scores (p= 0.01) across halves.

Regarding detection accuracy during alarms, reliability conditions were found to be marginally significant, ($F(1, 57) = 3.16$, $p = 0.08$). During non-alarms, automation reliability, ($F(1, 57) = 41.33$, $p < 0.001$) and experimental half, ($F(1, 57) = 4.45$, $p = 0.04$), were found to significantly affect detection accuracy. Participants improved their detection task accuracy during non-alarms across halves and were more accurate in the high reliability condition compared to the low reliability condition.

Detailed performance data, organized by experimental conditions and halves, is presented in Table 4.4.

Table 4.4: Mean and SD values for total performance scores, detection task score and accuracy, tracking task score, and the proportion of time spent on the tracking task display.

		Low Reliability, PPV = 0.54, NPV = 0.92			High Reliability, PPV = 0.78, NPV = 0.97		
		Overall Success Likelihood	Predictive Values	Hit/CR Rates	Overall Success Likelihood	Predictive Values	Hit/CR Rates
Total Score	H1	254.51 ± 8.10	271.05 ± 8.18	240.44 ± 8.34	294.67 ± 10.33	302.58 ± 8.18	284.19 ± 8.18
	H2	264.85 ± 7.78	277.55 ± 7.45	256.23 ± 6.81	297.47 ± 9.96	308.63 ± 8.41	284.60 ± 8.63
Tracking Task Score	H1	159.05 ± 8.54	172.50 ± 9.11	144.5 ± 9.66	188.38 ± 10.30	194.75 ± 7.86	177.90 ± 7.86
	H2	164.04 ± 8.43	174.85 ± 8.66	153.95 ± 7.61	189.71 ± 10.27	200.65 ± 8.47	176.00 ± 8.82
Tracking Task Proportion	H1	0.91 ± 0.01	0.91 ± 0.01	0.89 ± 0.01	0.96 ± 0.01	0.95 ± 0.01	0.94 ± 0.01
	H2	0.91 ± 0.01	0.91 ± 0.01	0.89 ± 0.01	0.95 ± 0.01	0.97 ± 0.01	0.93 ± 0.02
Detection Task Score	H1	95.47 ± 3.05	98.55 ± 2.32	95.93 ± 2.71	106.28 ± 1.51	107.83 ± 1.39	106.29 ± 1.88
	H2	100.81 ± 2.38	102.70 ± 2.06	102.28 ± 2.05	107.75 ± 1.44	107.98 ± 1.31	108.60 ± 1.30
Detection Task Accuracy (Alarms)	H1	0.78 ± 0.04	0.76 ± 0.04	0.75 ± 0.05	0.82 ± 0.03	0.85 ± 0.03	0.80 ± 0.02
	H2	0.77 ± 0.04	0.85 ± 0.04	0.77 ± 0.04	0.81 ± 0.03	0.79 ± 0.04	0.82 ± 0.03
Detection Task Accuracy (Non-Alarms)	H1	0.91 ± 0.02	0.91 ± 0.01	0.92 ± 0.01	0.94 ± 0.01	0.97 ± 0.01	0.95 ± 0.01
	H2	0.92 ± 0.02	0.94 ± 0.01	0.95 ± 0.01	0.96 ± 0.01	0.97 ± 0.01	0.98 ± 0.01

4.1.8 Discussion

We conducted a reanalysis of an existing dataset to explore how operators adapt their dependence behaviors, extreme response strategies, and performance to an automated aid. In each section of this discussion, we first note the results from the initial analyses and then transition to discuss the novel aspects of our reanalysis.

4.1.8.1 Dependence and Cross-Checking Behaviors

Consistent with the original analysis, our examination revealed a correlation between operator compliance rates and automation reliability. Upon dissecting the experiment into halves, it became evident that compliance rates remained steady in the low reliability conditions. Notably, within the high reliability conditions, interesting compliance adjustments emerged: participants in the OSL and Hit/CR Rate conditions each reduced compliance by 6% during the experiment. This reduction stood in stark contrast to the predictive values group, where participants initially demonstrated lower compliance compared to other groups but underwent a substantial increase, nearly doubling their compliance rates from 36% to 59%. Our reanalysis concerning reliance over halves found that participants did not significantly modify their reliance behaviors throughout the experiment.

Operator cross-checking rates exhibited an inverse relationship with blind behaviors, particularly noteworthy in the high reliability conditions where the reduction in alarm cross-checking precisely corresponded to the increase in compliance rates. This measure was more informative in studies where participants adjust their blind non-dependence behaviors, as in Chapter 3.

In contrast to the original analysis, our examination encompassed blind non-compliance and blind non-reliance behaviors. Our analyses indicated that automation reliability did not significantly impact blind non-compliance behaviors. Instead, likelihood information had a modest impact on blind non-compliance behaviors, particularly evident among participants in the predictive values condition. In the low-reliability, predictive values condition, participants exhibited blind non-compliance during nearly 10% of alarms. Conversely, blind non-reliance rates were low, which is appropriate as the NPV values were either 92% in the low reliability condition or 97% in the high reliability condition.

4.1.8.2 Response Strategies

Our reanalysis found that the majority of participants resorted to an extreme strategy and their strategies were appropriate to the automation's alarm and non-alarm performance. The original analysis did not examine the individual response strategies.

Automation reliability played a pivotal role in shaping the adoption of extreme use strategies among participants, leading to heightened levels of extreme automation use in the high reliability group. Moreover, groups with lower reliability exhibited a higher frequency of extreme cross-checking compared to their high reliability counterparts.

Likelihood conditions also exerted a notable impact on participants' response strategies. Participants assigned to the predictive values condition exhibited a higher prevalence of the extreme reliance compared to other groups, this was observed in both reliability conditions. Nearly every participant in the high reliability predictive values condition adopted the extreme reliance strategy.

Interestingly, by end of the experiment, a greater number of individuals applied the extreme reliance strategy in the low reliability, predictive values condition (14 participants) than in the high reliability conditions of OSL (10 participants) and Hit/CR (8 participants). Which might suggest that the likelihood information presented to operators might wield more influence on reliance behaviors than the automation's reliability. However, more research should explore the nuances of information presented to participants with widely varying automation PPV and NPV.

Upon dividing the experiment in half, few strategy changes were observed among operators, except in the Hit/CR conditions. In this condition, the number of participants employing an extreme non-alarm cross-checking strategy doubled across both reliability levels.

Additionally, operators in this group increased the use of extreme cross-checking strategies in both alarm and non-alarm states.

4.1.8.3 Performance

Our performance analyses yielded much of the same findings as the initial analyses. Including that participants in the predictive values condition outperformed counterparts in other likelihood conditions, and that human-automation team performance increased as automation reliability increased.

When we separated experimental halves, we illustrated that performance improved significantly. All groups improved at the tracking task, with improvements ranging from 1-9 points. Operators in the low reliability conditions improved at the detection task score, while the high reliability detection remained stable.

Our reanalysis found variations in detection task accuracy between alarm and non-alarm conditions. This is likely a consequence of automation performance across alarm states, given that non-alarm performance was better than alarm performance. Detection task accuracy during non-alarms remained relatively stable across halves. However, the predictive values condition stood out as having the most pronounced adjustments in detection task accuracy. These adjustments varied across reliability condition. In the high reliability group, there was a reduction in accuracy, while the low reliability group exhibited an accuracy increase.

The observed performance improvements can be attributed to operators' adjustments in dependence and cross-checking behaviors. Both alarm and non-alarm responses aligned appropriately with the corresponding automation performance condition.

4.1.8.4 Limitations

A limitation in assessing performance changes arises from the relatively small number of trials in the experiment, consisting of only 50 trials for each reliability condition with 20 participants in each condition (note that reliability was a within-subject variable). In Chapter 3, we observed that behaviors and adjustments tended to plateau around or after 50 trials with this specific experimental setup. Moreover, consistent with the original study by Du et al. (2020a), our analysis encountered experimental limitations, including that participants were required to estimate base rates and that and the criterion c was liberal, leading to a higher rate of false alarms than misses.

4.1.8.5 Conclusions

We reanalyzed data from a previous experiment to explore how likelihood information affects operators' dependence behaviors, cross-checking rates, and strategies. We divided the experimental session into halves to explore how operators used the alarm validity information and adjusted their behaviors. Beyond confirming the findings of the original study, our results revealed novel trends regarding how likelihood information influenced operators' dependence behaviors, strategies, and performance. These additional insights contribute to a more nuanced understanding of the impact of likelihood information on operator decision-making. Operators receiving likelihood information in the predictive values format established better dependence behaviors then their counterparts in alternative information conditions; which was best illustrated in the low-reliability condition, where 14 out of 20 individuals eventually adopted an extreme blind reliance strategy (the NPV was 0.92). Within the same group, 11 individuals employed an extreme cross-checking strategy during alarms (the PPV was 0.54). In contrast, participants in other information conditions did not demonstrate such a difference in their behaviors and strategies across alarms and non-alarms. Additionally, operators in the predictive value condition appeared to be more likely to make behavioral adjustments. In the high reliability condition, operators nearly doubled their compliance rates over the course of the experiment. Conversely, in the low automation condition, operators were more inclined to blindly ignore alarms.

We leveraged the insights gained to inform the design of a laboratory experiment, discussed in detail below. This iterative approach allowed us to refine methods and address potential challenges, resulting in a more comprehensive investigation.

4.2 Effects of presenting or withholding likelihood information on operators' adaptation

Study 4.2 consists of a laboratory experiment designed to further explore the influence of likelihood information onto operator dependence behaviors, response strategies, and human-automation team performance. We used the same dual-task scenario as completed in Study 4.1. Our experiment is different in that the number of trials was doubled (i.e., 100 trials) and we presented participants with different likelihood information. Specifically, we removed the likelihood conditions that resulted in inferior HAT performance and replace them with another form of likelihood information and a baseline condition. The additional form of likelihood information we introduced is a frequency format of predictive values. Participants in the baseline condition received no *a priori* information about automation performance. In Study 4.1, automation reliability was a significant factor for dependence and cross-checking behaviors, strategies selected, and performance. All participants in this experiment used the same automated aid so we could isolate the effects of the likelihood information presented.

4.2.1 Participants

A total of 78 (36 males, 40 females, and 1 person who identified as non-binary, and 1 non-response) university students (mean age = 22.97 years old, SD = 3.93) with normal or corrected to normal vision participated in our experiment. Participants were randomly assigned to an experimental condition and were paid USD $20, along with a performance bonus - which ranged from $2.50 to $10. Participant data from the no likelihood information condition was collected as part of Chapter 3, the data was group 5 from Chapter 3.1.

4.2.2 Design of Experiment

This study adopted a 3 (likelihood information: no information, predictive values, and predictive values in a frequency format) × 4 (experimental quarters) mixed design. Likelihood information was treated as between-subject factor and experimental quarter was a within-subjects factor.

We doubled the length of the reanalysis study to 100 trials separated by quarters: Q1 = trials 1-25, Q2 = trials 26-50, Q3 = trials 51-75, and Q4 = trials 76-100. Of the 100 trials, the experiment consisted of 40 hits, 30 correct rejections, 10 misses, and 20 false alarms (Group 5 in Table 3.1). Previous research was bench-marked to select the d' and c for our study (Wiczorek and Manzey, 2014).

4.2.3 Procedure

Email communication was used to recruit participants to the laboratory. After completing the informed consent form, each participant received a video orientation about the drone surveillance mission, each task, and the scoring system. Each participant completed 30 practice tracking task trials, followed by 8 dual-task practice trials. A researcher accompanied participants during practice session to answer participant questions. To mitigate participant fatigue, a mandatory 5-minute break was given halfway through the session.

4.2.4 Data and Analyses

The same measures were collected as in section 3.3.4. The experimental session consisted of 100 trials and were separated into quarters: Q1 = trials 1-25, Q2 = trials 26-50, Q3 = trials 51-75, and Q4 = trials 76-100. Factorial analysis of variance (ANOVA) tests were executed to examine relationships between our dependent and independent variables. In instances of statistical significance, a post-hoc Bonferroni test was performed. Simple effect analyses were conducted when a significant interaction effect was found between independent variables.

A chi-square test of independence was conducted to determine strategy count differences between likelihood information conditions. To examine changes in the counts of extreme strategies, McNemar's test Eliasziw and Donner (1991) was employed to compare the first quarter (Q1) to the last quarter (Q4) of the experiment. A continuity correction was applied to each group.

4.2.5 Results

We present the experiment's results starting with dependence and cross-checking behaviors, followed by operator extreme strategy counts, and finally, performance. We dissected the data across four quarters to examine the influence of likelihood information onto how operators adapt as they increase automation usage.

4.2.5.1 Compliance, Blind Non-Compliance, and Alarm Cross-Checking

Compliance behaviors increased as the experiment progressed, $F(3, 75) = 27.12,\ p < 0.001$. Post hoc testing indicated that there were significant differences between Q1-Q2 ($p < 0.001$) and Q2-Q3 ($p = 0.004$), but not between Q3-Q4. Likelihood information did not significantly affect operators' compliance rates.

Blind non-compliance was significantly different across likelihood conditions $F(2,\ 75) = 3.11$, $p = 0.05$. Post hoc testing found significant differences between frequency and baseline conditions ($p = 0.001$), and marginal significance for differences between predictive values and baseline conditions ($p = 0.09$). The effects of experimental quarters were not significant.

Alarm cross-checking reduced as participants progressed through the experiment, $F(3,\ 75) = 24.48$, $p < 0.001$. The effects of likelihood information were non-significant on alarm cross-checking ($p = 0.58$).

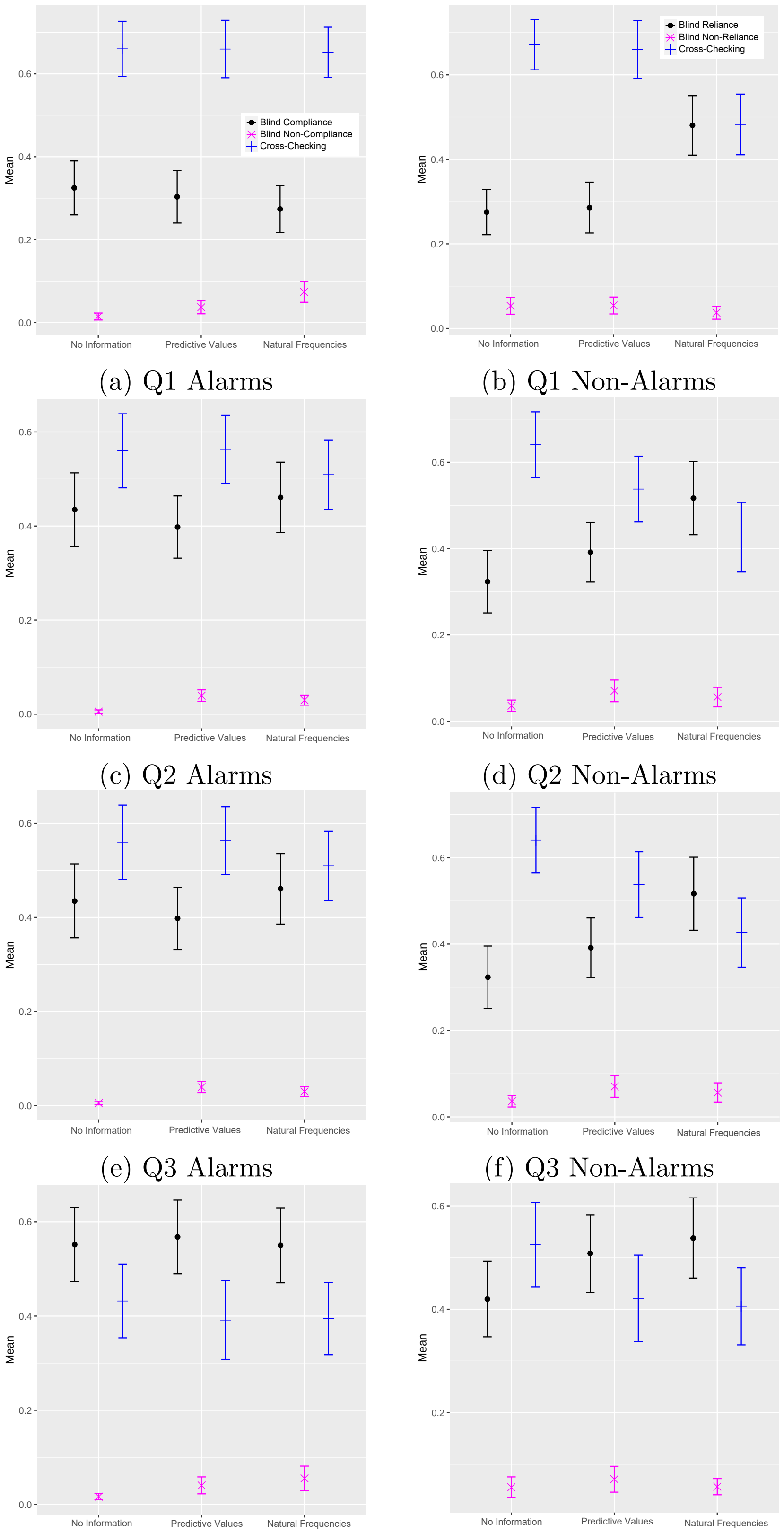
4.2.5.2 Reliance, Non-reliance and Non-alarm Cross-checking

Reliance rates increased as the experiment progressed $F(3,\ 75) = 8.69$, $p < 0.01$. The differences in reliance across likelihood conditions were not found to be significant ($p = 0.13$). Neither likelihood information nor experimental quarter significantly influenced blind non-reliance. Non-alarm cross-checking rates decreased as the experiment progressed, $F(2, 75) = 9.53$, $p < 0.001$. The differences in non-alarm cross-checking between likelihood conditions were not found to be significant ($p = 0.17$).

The dependence and cross-checking descriptive results are presented in Table 4.5 and are complemented by Figure 4.2, which illustrates graphs of dependence behaviors plotted for the three likelihood conditions. The graphs are organized into two distinct columns; the left column displays alarm responses (i.e., compliance, blind, non-compliance, and alarm cross-checking) across conditions, while the right column displays non-alarm responses across conditions. Each row corresponds to a distinct quarter.

Table 4.5: Mean ± standard error (SE) values for dependence, non-dependence, and cross-checking rates.

		PPV = 0.67, NPV = 0.75		
		No Likelihood Information	Predictive Values	Frequency Format
Blind Compliance	Q1	0.33 ± 0.07	0.30 ± 0.06	0.27 ± 0.06
	Q2	0.43 ± 0.08	0.40 ± 0.07	0.46 ± 0.07
	Q3	0.48 ± 0.08	0.57 ± 0.08	0.56 ± 0.07
	Q4	0.55 ± 0.08	0.57 ± 0.08	0.55 ± 0.08
Blind Non-Compliance	Q1	0.02 ± 0.01	0.04 ± 0.02	0.07 ± 0.03
	Q2	0.01 ± 0.00	0.04 ± 0.01	0.03 ± 0.01
	Q3	0.02 ± 0.01	0.03 ± 0.01	0.06 ± 0.02
	Q4	0.02 ± 0.01	0.04 ± 0.02	0.06 ± 0.03
Alarm Cross-Checking	Q1	0.66 ± 0.07	0.66 ± 0.07	0.65 ± 0.06
	Q2	0.56 ± 0.08	0.56 ± 0.07	0.51 ± 0.07
	Q3	0.51 ± 0.08	0.40 ± 0.09	0.38 ± 0.07
	Q4	0.43 ± 0.08	0.40 ± 0.08	0.40 ± 0.08
Blind Reliance	Q1	0.28 ± 0.05	0.29 ± 0.06	0.48 ± 0.07
	Q2	0.32 ± 0.07	0.39 ± 0.07	0.52 ± 0.09
	Q3	0.35 ± 0.07	0.50 ± 0.08	0.57 ± 0.08
	Q4	0.42 ± 0.07	0.51 ± 0.07	0.54 ± 0.08
Blind Non-Reliance	Q1	0.05 ± 0.02	0.05 ± 0.02	0.04 ± 0.02
	Q2	0.04 ± 0.01	0.07 ± 0.03	0.06 ± 0.02
	Q3	0.06 ± 0.02	0.07 ± 0.02	0.06 ± 0.02
	Q4	0.06 ± 0.02	0.07 ± 0.03	0.06 ± 0.02
Non-Alarm Cross-Checking	Q1	0.67 ± 0.07	0.66 ± 0.07	0.48 ± 0.07
	Q2	0.64 ± 0.08	0.54 ± 0.08	0.43 ± 0.08
	Q3	0.59 ± 0.08	0.43 ± 0.09	0.37 ± 0.08
	Q4	0.53 ± 0.08	0.42 ± 0.08	0.41 ± 0.08



(a) Q1 Alarms

(b) Q1 Non-Alarms

(c) Q2 Alarms

(d) Q2 Non-Alarms

(e) Q3 Alarms

(f) Q3 Non-Alarms

(g) Q4 Alarms

(h) Q4 Non-Alarms

Figure 4.2: Left column displays alarm responses; right column displays non-alarm responses.

4.2.5.3 Extreme Response Strategies

We tallied participants who employed extreme response strategies throughout the experiment, as detailed in Table 4.6.

Table 4.6: Illustrates participant counts (out of 26) for the different strategies, separated by experimental quarter (Q).

		PPV = 0.67, NPV = 0.75		
		No Likelihood Information	Predictive Values	Frequency Format
Extreme Blind Compliance	Q1	2	1	1
	Q2	6	3	4
	Q3	6	8	9
	Q4	7	8	7
Extreme Blind Non-Compliance	Q1	0	0	0
	Q2	0	0	0
	Q3	0	0	0
	Q4	0	0	0
Extreme Alarm Cross-Checking	Q1	10	11	8
	Q2	9	8	6
	Q3	8	7	3
	Q4	7	7	4
Extreme Blind Reliance	Q1	1	0	4
	Q2	3	4	9
	Q3	3	5	9
	Q4	4	5	10
Extreme Blind Non-Reliance	Q1	0	0	0
	Q2	0	0	0
	Q3	0	0	0
	Q4	0	0	0
Extreme Non-Alarm Cross-Checking	Q1	8	11	6
	Q2	13	8	6
	Q3	10	8	6
	Q4	7	4	7

Likelihood information did not significantly impact the utilization of extreme blind compliance strategy ($p = 0.87$). During the experiment, a significant amount of participants changed their response strategy to the extreme blind compliance ($\tilde{\chi}(1) = 16.06,\ p < 0.001$).

No participants engaged in the extreme blind non-compliance or non-reliance strategies.

Likelihood information marginally impacted the utilization of extreme cross-checking strategy, ($\tilde{\chi}\ (2) = 4.97,\ p = 0.08$). Additionally, the number of participants using extreme alarm cross-checking decreased as the experiment progressed, ($\tilde{\chi}$ (1) = 5.88, p = 0.02).

Likelihood information significantly affected the number of participants utilizing the extreme blind reliance strategy ($\tilde{\chi}(2) = 16.61,\ p < 0.001$). Post hoc testing indicated that more participants used this strategy when presented the frequency format when compared to predictive values ($\tilde{\chi}$ (1) = 8.07, p = 0.005), or baseline ($\tilde{\chi}$ (1) = 11.73, p = 0.0006). Participants significantly increased their extreme reliance strategies as the experiment progressed, ($\tilde{\chi}$ (1) = 10.56, p = 0.001).

Likelihood information influenced the number of participants using the extreme cross-checking strategy during non-alarms ($\tilde{\chi}(2) = 6.33,\ p = 0.04$). Post hoc testing indicated that the frequency condition was significantly lower when compared to baseline ($\tilde{\chi}$ (1) = 5.27, p = 0.02). The number of participants using extreme non-alarm cross-checking did not significantly change (p = 0.12).

4.2.5.4 Performance

Detailed performance data, organized by condition and quarters, is presented in Table 4.7. Likelihood information did not significantly affect total performance. Total performance significantly improved during the experiment, (F(3, 75) = 79.84, p < 0.001).

Likelihood information did not affect tracking task performance. Participants significantly improved at the tracking task during the experiment (F(3, 75) = 93.13, p< 0.001). Participants differed the proportion of time on the tracking task during the experimental quarters, (F(3, 75) = 37.95, p < 0.001). Post hoc testing indicated that participants increased tracking task proportion from Q1-Q2 (p < 0.001),and from Q2-Q3 (p < 0.001).

Neither likelihood condition (p = 0.78), nor experimental quarter (p = 0.42), significantly affected detection task score. Regarding detection task accuracy during alarms, there were significant differences between experimental quarters (F(3, 75) = 7.03, p< 0.001). Post hoc testing indicated significant differences between Q1-Q3 (p = 0.005), Q2-Q3 (p = 0.051), and marginal differences between Q2-Q4 (p = 0.08). Detection task accuracy during non-alarms did not significantly vary across likelihood conditions ($p = 0.62$) or between experimental quarters ($p = 0.42$).

Table 4.7: Mean ± standard error (SE) values for total performance scores, tracking task scores and proportion, and detection task scores and accuracy.

		PPV = 0.67, NPV = 0.75		
		No Likelihood Information	Predictive Values	Frequency Format
Total Score	Q1	234.54 ± 38.78	231.04 ± 48.85	241.97 ± 55.41
	Q2	256.11 ± 47.31	251.04 ± 49.00	257.36 ± 52.76
	Q3	268.46 ± 42.46	277.26 ± 43.49	277.17 ± 47.23
	Q4	276.46 ± 43.49	280.07 ± 41.20	283.67 ± 49.13
Tracking Task Score	Q1	144.27 ± 40.08	141.54 ± 53.25	152.73 ± 57.22
	Q2	164.27 ± 50.07	161.73 ± 53.69	170.46 ± 57.46
	Q3	181.50 ± 47.67	187.69 ± 55.00	192.15 ± 54.28
	Q4	188.85 ± 48.89	192.46 ± 49.73	195.11 ± 50.69
Tracking Task Proportion	Q1	0.87 ± 0.07	0.88 ± 0.07	0.90 ± 0.05
	Q2	0.90 ± 0.07	0.91 ± 0.07	0.92 ± 0.06
	Q3	0.91 ± 0.07	0.93 ± 0.08	0.94 ± 0.06
	Q4	0.92 ± 0.07	0.93 ± 0.08	0.94 ± 0.06
Detection Task Score	Q1	90.27 ± 14.68	89.50 ± 13.55	89.24 ± 10.11
	Q2	91.83 ± 10.25	89.31 ± 15.13	86.90 ± 9.40
	Q3	86.95 ± 15.11	89.57 ± 14.90	85.01 ± 16.90
	Q4	87.62 ± 11.77	87.61 ± 15.49	88.56 ± 13.33
Detection Task Accuracy (Alarms)	Q1	0.84 ± 0.17	0.84 ± 0.14	0.80 ± 0.14
	Q2	0.84 ± 0.12	0.81 ± 0.17	0.76 ± 0.14
	Q3	0.76 ± 0.16	0.78 ± 0.15	0.73 ± 0.16
	Q4	0.77 ± 0.15	0.74 ± 0.17	0.76 ± 0.16
Detection Task Accuracy (Non-Alarms)	Q1	0.87 ± 0.15	0.81 ± 0.17	0.83 ± 0.16
	Q2	0.82 ± 0.15	0.81 ± 0.16	0.82 ± 0.13
	Q3	0.82 ± 0.17	0.82 ± 0.17	0.77 ± 0.19
	Q4	0.83 ± 0.13	0.81 ± 0.18	0.80 ± 0.14

4.2.6 Discussion

In the current study, we predicted that participants would use likelihood information to adapt their behaviors to the automated aid, resulting in improved performance. To test this prediction, we conducted a laboratory experiment comparing a baseline condition, where no *a priori* information was provided to participants, against conditions where participants received likelihood information regarding automation alarm and non-alarm performance. We examined operators' dependence behaviors, extreme strategy counts, and dual-task performance.

4.2.6.1 Dependence and Cross-Checking Behaviors

Our findings indicate that likelihood information did not result in significant differences in dependence rates across conditions. However, it did appear to influence the rate at which participants adjusted their behaviors. Participants without likelihood information exhibited a gradual increase in dependence rates and relied on trial-by-trial feedback for understanding. In contrast, participants provided with information adapted more swiftly and stabilized earlier. Bettman et al. (1990) suggested that decision-makers base their strategy selection on a combination of cognitive efforts and system accuracy. Providing performance information to our participants likely reduced their cognitive load, as they were not required to estimate automation's accuracy from a state of informational emptiness, potentially accelerating their strategy selection process. Further research is essential to validate this finding.

When dissecting dependence into compliance and reliance, we observed that likelihood information did not impact participants' compliance behaviors but did influence their reliance on automation. This could be attributed to the uncertainty associated with a PPV of 0.67, which did not necessarily reduce operator uncertainty, while an NPV of 0.75 could be perceived as more reliable. Initial differences in reliance behaviors were evident, with participants in the frequency format condition relying more on automation. However, as the experiment progressed, participants using predictive values gradually aligned their reliance rates with participants in the frequency format, while the uninformed group's behavior appeared to still be changing.

Participants with likelihood information exhibited similar compliance and reliance rates, which is likely due to the comparable PPV and NPV. This contrasted participants without information, who were required to estimate automation performance through repeated interactions. Additionally, they tended to follow automation recommendations more in alarms than non-alarms, despite better automation performance during non-alarms.

Surprisingly, participants with likelihood information were more willing to blindly ignore alarms, they did this at similar rates during non-alarms. In contrast, participants without likelihood information only seemed to blindly ignore non-alarms. However, the rates for ignoring the automation was low across all conditions and this behavior did not lead to better performance.

4.2.6.2 Extreme Response Strategies

More than half of the participants eventually adopted extreme strategies, where they consistently blindly depended on or cross-checked against the automated aid. Notably, participants distinguished these strategies between alarm and non-alarm states.

In alarm response strategies, likelihood information did not influence how participants blindly followed automation alarms. Similar numbers of participants exhibited extreme blind compliance and extreme cross-checking across alarm states. As participants prolonged their automation use, the number of participants employing an extreme compliance strategy increased. This contrasted non-alarm response strategies, which were significantly impacted by likelihood information. More participants had an extreme amount of blind reliance in the frequency condition compared to other conditions. They also refrained from consistently validating non-alarm recommendations. As the experiment progressed, more participants exhibited the extreme blind reliance strategy.

Examining cross-checking strategies allowed us to explore how participants utilized alarm validity information, revealing differences across information conditions. The frequency condition exhibited considerably less extreme cross-checking than other conditions, indicating a more balanced approach influenced by the presented information.

4.2.6.3 Performance

The likelihood information presented did not significantly affect total performance. Participants improved total performance by prioritizing the tracking task over the detection task, which the data illustrated through the proportion of time on the tracking task.

Considering performance by task, we found that likelihood information did not affect participants detection task scores. There was a slight reduction in the detection task performance, which resulted from increased compliance on an imperfect automated aid. It appears that participants accepted the reduction in alarm detection task accuracy to increase their tracking task scores. The non-alarm detection accuracy was similar throughout the experiment and across likelihood conditions.

4.2.7 Limitations

Several limitations were inherent in our approach. We deliberately chose a fine-grained view of one automation reliability level, which was close to the thresholds identified in Chapter 2 of around 70%. This decision allowed for a nuanced exploration. However, it also created a scenario where participants, even with knowledge of automation performance, might remain uncertain about whether to blindly follow or validate the automation's recommendation. Although opting for a clearly strong or weak automation performance was an alternative, it was avoided as it would yield less interesting information — participants would likely follow highly performing automation and ignore poorly performing automation.

Another limitation was the requirement for participants to estimate the base rate, introducing an additional cognitive load. Furthermore, the liberal criterion c resulted in a higher number of false alarms than misses.

Despite doubling the experimental length in the reanalysis, our findings are constrained by the fact that participants were novices transitioning from no experience to a limited amount. Research by others has highlighted that experts are less likely to follow imperfect automation, especially when they have the option to manually complete tasks (Sanchez et al., 2014).

While the majority of participants adopted extreme response strategies, there exist other strategies that could be operationalized for examination. For instance, the probabilistic matching strategy could operationalized when compliance is within ±10% of the automation's PPV. Further work is needed to explore and define appropriate thresholds, as the selected range can significantly impact findings and might be perceived as either too liberal or conservative.

4.3 Chapter Conclusions

This chapter presented two human subject studies aimed at investigating how operators utilize likelihood information while interacting with an imperfect automated aid.

Study 4.1 involved a reanalysis of a previous study, extending its findings by comparing operators' dependence and cross-checking behaviors, extreme response strategies, and performance between the first and second halves of the experiment. The influence of likelihood information on operators and their adaptations to the imperfect automation was explored, revealing that both likelihood information and reliability significantly affected how participants utilized the automation. Predictive values emerged as particularly useful in improving performance. The study highlighted that participants in all conditions adjusted their behaviors to the scenario and improved as the experiment progressed. However, a limitation of the relatively short experiment was noted, as participants may have still been adapting their behaviors at that point.

To address this limitation, we conducted the second study, where we doubled the number of trials. Likelihood information was again manipulated, but with different conditions, including a baseline of no information, predictive values, and a frequency format of predictive values. By maintaining the same reliability levels, we were able to examine how the combination of likelihood information and experience affected operators' dependence and cross-checking behaviors, strategies, and performance.

Several themes emerged when considering both studies. Automation reliability exerted more influence on behaviors and strategies than likelihood information. When the automation's likelihood was perceived as clearly good or bad, participants were more likely to adjust their behaviors and strategies.

Controlling for automation reliability revealed that participants with likelihood information made faster adjustments than those without information. Post-trial feedback played a crucial role, enabling participants without *a priori* likelihood information to perform similarly to those with information. The uncertainty around our selected PPV and NPV levels likely contributed to the challenge participants faced in determining how much they should follow or ignore the automated aid.

Surprisingly, both studies uncovered instances of blind non-compliance, where participants ignored the automation and AVI. Participants with likelihood information were more inclined to attempt blind disuse, even when the worst PPV was still over 50%. Exploring the impact of likelihood information on operators across diverse automation performance conditions presents an opportunity for future research. A deeper understanding could provide valuable insights on when to provide likelihood information or consider alternative solutions, offering practical guidance for system design and implementation.

Finally, a consistent preference for prioritizing tracking over threat detection was observed across both studies, echoing findings in Chapter 3 and other experiments (Wiczorek and Manzey, 2014). This strategic use of automation suggests that participants accept a loss in threat detection accuracy to allocate more time to the tracking task.

The insights gained from these findings offer guidance for engineers and system designers seeking to enhance operators' dependence behaviors. Such improvements can, in turn, contribute to enhanced total human-automation performance.

CHAPTER 5

Conclusions

5.1 Summary

To address the problem of inappropriate use of imperfect automation and contribute to the knowledge gaps surrounding the role of alarm validity information, the aims of this dissertation were to:

1. Systematically examine how operators' dependence behaviors (i.e., compliance and reliance) and cross-checking behaviors are affected by automation performance, separated across alarm states.

2. Evaluate how operators adapt their dependence behaviors, cross-checking rates, and response strategies to varying degrees of imperfect automation.

3. Investigate a design intervention focusing on the incorporation of likelihood information, specifically, to compare the effects of predictive values against a frequency format and a baseline condition, where no *a priori* information was provided.

To address aim 1, we conducted a meta-analysis regarding human operator use of imperfect diagnostic automation in dual task scenarios. We systematically extracted dependence and cross-checking behavior data to examine them across a range of alarm and non-alarm performances. We found that the human operators not only varied their compliance and reliance behaviors to the automation, but also varied how often they used additional information to verify the automation's recommendation. Human operators' blind compliance ($\beta_1 = .74$) and reliance ($\beta_1 = .89$) behaviors increased as automation's PPV and NPV increased. Alternatively, when automation performed worse, the operators were more likely to verify automation's recommendation. Operators' cross-checking behaviors were marginally ($p = 0.08$) more sensitive to non-alarm errors ($\beta_1 = -.90$) than alarm errors ($\beta_1 = -.52$). Our results, in light of a separate meta-analysis by Wickens and Dixon (2007), suggest that operators tend to blindly follow automation when the benefits outweigh the costs and cross-check with AVI when the costs outweigh the benefits.

To address aim 2, we utilized a dual-task laboratory experiment to evaluate how operators adapt their dependence behaviors, cross-checking rates, and response strategies as they increase their automation use. We found that automation performance influenced dependence behaviors and response strategies. More specifically, operators used trial-by-trial feedback to adjust to alarms and non-alarm; their behaviors and strategies were independently adapted to the automation's PPV and NPV. We introduced a novel optimal decision-making strategy that considers operators' access to alarm validity information. In the experiment, adjustments in behaviors converged towards the theoretical optimal behavior. However, more research is needed to empirically validate the proposed optimal strategy. Future researchers could systematically manipulate cross-checking costs, calculate the optimal strategy, and examine the relationships between the strategies that operators selected and performance.

To address aim 3, we presented two human subject studies to compare how operators utilize different likelihood information. The first study was a reanalysis of previous research, where we compared operators' dependence and cross-checking behaviors, extreme response strategies, and performance between the first and second halves of the experiment. Likelihood information and reliability significantly affected how participants interacted with the automation. Predictive values emerged as particularly useful in facilitating total performance. The study highlighted that participants in all conditions adjusted their behaviors to the scenario and improved as the experiment progressed. The second study compared the same dependent measures across different likelihood information conditions, including a baseline of no information, predictive values, and frequency formats. We observed that likelihood information did not impact participants' compliance behaviors but did influence their reliance on automation. More than half of the participants eventually adopted extreme strategies, where they consistently blindly followed or validated the automated aid's recommendation. The likelihood information presented did not significantly affect performance. However, we found that participants strategically changed their behaviors to improve total performance; they accepted a loss in threat detection accuracy to allocate more time to the tracking task. Participants with likelihood information made slightly faster behavioral adjustments than those without information.

5.2 Intellectual Merit and Broad Impact

The study significantly contributes to a further comprehension of operator dependence behaviors, use of alarm validity information, and individual strategies in response to imperfect automation.

First, the findings enrich our understanding of how operators depend on, validate, or ignore automated systems during dual-task performance. Insights into operators' decision-making behaviors under uncertainty addresses a fundamental question: how well must automation perform for operators to follow recommendations, and at what point do operators choose to manually validate recommendations? The implications of this work extend to various domains, including military, maritime, industrial settings, and aviation, and provides valuable knowledge for enhancing human-automation interactions.

Second, the introduction of a theoretical optimal standard is a useful advancement. This theoretical optimal can serve as a benchmark, enabling operators to calibrate their dependence behaviors. Knowledge of dependence rates becomes actionable when coupled with a calibration standard. Improvements to decision-making can enhance automation's benefits, which include improved safety, cost reduction, and decreased operator workload.

Finally, the insights into how operators adapt their behaviors can facilitate an accelerated learning process for operators and support more effective solution implementation. This understanding is pivotal for researchers and can provide deeper insights on the impact of solutions, training interventions, and other human-automation enhancements. A nuanced understanding of operator behaviors can inform system designers on whether to incorporate likelihood information or to consider alternative solutions, offering practical guidance for the design and implementation of automated systems. In essence, this research not only broadens our knowledge base but also has implications for the improvement of human-automation interactions and the design of intelligent systems.

5.3 General Limitations and Future Work

Through a series of studies, this dissertation contributes to understanding of operator dependence behaviors, use of alarm validity information, response strategies, and performance. As with all research, there are current limitations that can be treated as future research avenues.

We conducted a thorough exploration on how operators are using and adapting to an imperfect automated aid. Using the same dual-task scenario allowed us to isolate the effects of reliability, error types, and forms of likelihood information. The scenario was useful for basic research, as participants did not require prior experience and the insights gained from fundamental studies can inform the development of solutions to real-world challenges. However, other scenarios and scoring systems should be evaluated to inform on the nuances of human behaviors and behavioral adaptations. This would allow for empirical validation of our proposed optimal solution, which considered the costs and benefits of when an operator can manually validate an automated recommendation. Furthermore, our objectives were pursued within an experimental scenario that consisted of 10-second trials. This design was useful for our aims, however, the design may limit the generalizability of our findings to brief, discrete interactions with automation. Real-world automation operators continuously use automation over extended periods of time, which may not be fully captured within the confines of our experimental setup. Another limitation stemming from our experimental configuration was the use of a binary alarm system. While employing the binary alarm enabled us to assess alarm performance across various alarm states utilizing SDT, prior research indicated potential benefits associated with supplying operators with a likelihood alarm when compared to binary alarms (Sorkin et al., 1988; Wiczorek et al., 2014; Clark et al., 2009).

Second, we analyzed the dependence and cross-checking behaviors of multiple operators across multiple experiments. This was followed by a more focused examination of operators' behaviors with varying automation errors during repeated interactions. We separated the data into quarters to illustrate how operators are adapting to the automation and scenario. Dividing the experiment into quarters was a trade-off decision (as opposed to halves, thirds, etc.), which required careful consideration of the alarm characteristics and of strategy operationalization. If aggregating trial data into a larger set (such as the entire experiment), the results provide a coarse view of the human-automation interaction. If too few trials are selected for a set, the operationalization of strategies can artificially categorize behaviors as extreme. Each research project establishes the requisite level of detail for its specific objectives. For those investigating dynamic aspects of measures, alternative approaches, like a smoothed average method, may be considered for analyses. Other research has examined behaviors from the beginning of the experimental session to trial i and compared results for i at different points in time.

By categorizing an individual's strategy, we were able to illustrate individuals' 'all or nothing' behaviors. We chose to focus on the extreme strategy because the majority of participants used it and the strategy aligned with our proposed optimal strategy. However, there are additional individual strategies that can be operationalized and explored, and are potentially more relevant to other researchers. An example would be the probabilistic matching strategy (Bliss et al., 1995), which is where the operator roughly matches their compliance and reliance behaviors to the automation's PPV and NPV, respectively. Further work is needed to define appropriate thresholds, as the selected range could be perceived as either too liberal or conservative. Additionally, there could be an overlap between probabilistic matching and extreme strategy counts in high ($> 90\%$) or low ($< 10\%$) performance conditions.

Finally, we investigated individual strategies to improve understanding on how individuals utilize automation. We manipulated aspects of the human-automation interaction and focused on how operators adapted to the automation. Technological progress has enabled the calibration of autonomous systems to an individual, a concept that has been referred to as "individuation" (Hancock et al., 2009). In order to appropriately calibrate automation to the human, it is crucial to understand individual characteristics and traits. Researchers have created fairly concise instruments for participants to self report aspects aspects of themselves, including an individual's culture (Yoo et al., 2011), personality (Donnellan et al., 2006), propensity to trust (Merritt, 2011; Merritt et al., 2013), and attentional control (Derryberry and Reed, 2002). Additionally, previous studies indicated that individual factors significantly impact trust (Schaefer et al., 2016; Hoff and Bashir, 2015), dependence (McBride et al., 2011), and operator strategy (Riley, 1996). Future research could extensively investigate the influence of these characteristics and traits within human-automation interactions and explore their potential impact on human dependence behaviors and adaptations to the automation.

BIBLIOGRAPHY

Note: References marked with asterisks were used for meta-analyses.

Allendoerfer, K. R., Pai, S., and Friedman-Berg, F. J. (2008). The complexity of signal detection in air traffic control alert situations. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52, pages 54–58. SAGE Publications Sage CA: Los Angeles, CA.

Bagheri, N. and Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 1, pages 212–217. IEEE.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., and Frith, C. D. (2010). Optimally interacting minds. *Science*, 329(5995):1081–1085.

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

*Barg-Walkow, L. H. (2013). *Understanding the role of expectations on human responses to an automated system*. PhD thesis, Georgia Institute of Technology.

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

Bartlett, M. L. and McCarley, J. S. (2017). Benchmarking aided decision making in a signal detection task. *Human factors*, 59(6):881–900.

Bartlett, M. L. and McCarley, J. S. (2021). Ironic efficiency in automation-aided signal detection. *Ergonomics*, 64(1):103–112.

Bettman, J. R., Johnson, E. J., and Payne, J. W. (1990). A componential analysis of cognitive effort in choice. *Organizational behavior and human decision processes*, 45(1):111–139.

Bhat, S., Lyons, J. B., Shi, C., and Yang, X. J. (2023). Evaluating the impact of personalized value alignment in human-robot interaction: Insights into trust and team performance outcomes. *arXiv preprint arXiv:2311.16051*.

Bliss, J. P. (1993). The cry-wolf phenomenon and its effect on operator responses. *Unpublished doctoral dissertation, University of Central Florida, Orlando.*

*Bliss, J. P. (1997). Alarm reaction patterns by pilots as a function of reaction modality. *The International Journal of Aviation Psychology*, 7(1):1–14.

Bliss, J. P. (2003). An investigation of extreme alarm response patterns in laboratory experiments. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 47, pages 1683–1687. SAGE Publications Sage CA: Los Angeles, CA.

*Bliss, J. P. and Chancey, E. (2010). The effects of alarm system reliability and reaction training strategy on alarm responses. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54, pages 2248–2252. SAGE Publications Sage CA: Los Angeles, CA.

*Bliss, J. P. and Dunn, M. C. (2000). Behavioural implications of alarm mistrust as a function of task workload. *Ergonomics*, 43(9):1283–1300.

Bliss, J. P., Gilson, R. D., and Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, 38(11):2300–2312.

Breznitz, S. (1984). *Cry wolf: The psychology of false alarms.* Psychology Press.

Bustamante, E. A. (2005). A signal detection analysis of the effects of workload, task-critical and likelihood information on human alarm response. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 49, pages 1513–1517. SAGE Publications Sage CA: Los Angeles, CA.

*Bustamante, E. A. and Bliss, J. P. (2005). Effects of workload and likelihood information on human response to alarm signals. In *2005 International Symposium on Aviation Psychology*, page 97.

Chancey, E. T., Bliss, J. P., Proaps, A. B., and Madhavan, P. (2015). The role of trust as a mediator between system characteristics and response behaviors. *Human factors*, 57(6):947–958.

*Chancey, E. T., Bliss, J. P., Yamani, Y., and Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human factors*, 59(3):333–345.

*Chancey, E. T., Proaps, A., and Bliss, J. P. (2013). The role of trust as a mediator between signaling system reliability and response behaviors. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 57, pages 285–289. SAGE Publications Sage CA: Los Angeles, CA.

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

Clark, R. M., Peyton, G. G., and Bustamante, E. A. (2009). Differential effects of likelihood alarm technology and false-alarm vs. miss prone automation on decision making. In *Proceedings of the Human Factors and Ergonomics Society annual meeting*, volume 53, pages 349–353. Sage Publications Sage CA: Los Angeles, CA.

Comstock Jr, J. R. and Arnegard, R. J. (1992). The multi-attribute task battery for human operator workload and strategic behavior research. Technical report.

De Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., and Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International journal of social robotics*, 12(2):459–478.

Degani, A. (2001). *Taming HAL: Designing Interfaces Beyond 2001*. Palgrave Macmillan, New York, NY.

Denkiewicz, M., Migda, P., Plewczynski, D., et al. (2013). Information-sharing in three interacting minds solving a simple perceptual task. In *Proceedings of the annual meeting of the cognitive science Society*, volume 35.

Derryberry, D. and Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of abnormal psychology*, 111(2):225.

*Dixon, S. R. (2006). *Imperfect diagnostic automation: How adjusting bias and saliency affects operator trust*. PhD thesis, University of Illinois at Urbana-Champaign.

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

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

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

Donnellan, M. B., Oswald, F. L., Baird, B. M., and Lucas, R. E. (2006). The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192.

Dorfman, D. D. (1969). Probability matching in signal detection. *Psychonomic Science*, 17(2):103–103.

*Du, N., Huang, K. Y., and Yang, X. J. (2020a). Not all information is equal: effects of disclosing different types of likelihood information on trust, compliance and reliance, and task performance in human-automation teaming. *Human factors*, 62(6):987–1001.

Du, N., Zhou, F., Pulver, E., Tilbury, D. M., Robert, L. P., Pradhan, A. K., and Yang, X. J. (2020b). Examining the effects of emotional valence and arousal on takeover performance in conditionally automated driving. *arXiv preprint arXiv:2001.04509*.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *International journal of human-computer studies*, 58(6):697–718.

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

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., and Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology*, 13(3):147–164.

Elghoneimy, E. and Gruver, W. A. (2012). Agent-based decision support and simulation for wood products manufacturing. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1656–1668.

Eliasziw, M. and Donner, A. (1991). Application of the mcnemar test to non-independent matched pair data. *Statistics in medicine*, 10(12):1981–1991.

Fletcher, K. I., Bartlett, M. L., Cockshell, S. J., and McCarley, J. S. (2017). Visualizing probability of detection to aid sonar operator performance. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 61, pages 302–306. SAGE Publications Sage CA: Los Angeles, CA.

*Gérard, N. and Manzey, D. (2010). Are false alarms not as bad as supposed after all? a study investigating operators' responses to imperfect alarms. *Human factors. A system view of human, technology and organisation*, pages 55–69.

Getty, D. J., Swets, J. A., Pickett, R. M., and Gonthier, D. (1995). System operator response to warnings of danger: A laboratory investigation of the effects of the predictive value of a warning on human response time. *Journal of experimental psychology: applied*, 1(1):19.

Gigerenzer, G. (2008). Moral intuition= fast and frugal heuristics? In *Moral psychology*, pages 1–26. MIT Press.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: Frequency formats. *Psychological review*, 102(4):684.

Gigerenzer, G. and Todd, P. M. (1999). *Simple heuristics that make us smart.* Oxford University Press, USA.

Gigerenzer, G. E., Hertwig, R. E., and Pachur, T. E. (2011). *Heuristics: The foundations of adaptive behavior.* Oxford University Press.

Guo, Y., Shi, C., and Yang, X. J. (2021). Reverse psychology in trust-aware human-robot interaction. *IEEE Robotics and Automation Letters*, 6(3):4851–4858.

Guo, Y. and Yang, X. J. (2021). Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, 13(8):1899–1909.

Guo, Y., Yang, X. J., and Shi, C. (2023). Enabling team of teams: A trust inference and propagation (tip) model in multi-human multi-robot teams. *arXiv preprint arXiv:2305.12614*.

Hancock, P., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, 53(5):517–527.

Hancock, P., Kessler, T. T., Kaplan, A. D., Brill, J. C., and Szalma, J. L. (2021). Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors*, 63(7):1196–1229.

Hancock, P. A., Hancock, G. M., and Warm, J. (2009). Individuation: the n= 1 revolution. *Theoretical Issues in Ergonomics Science*, 10(5):481–488.

Hocraffer, A. and Nam, C. S. (2017). A meta-analysis of human-system interfaces in unmanned aerial vehicle (uav) swarm management. *Applied ergonomics*, 58:66–80.

Hoff, K. A. and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434.

Hoffrage, U. and Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic medicine*, 73(5):538–40.

Hoffrage, U., Gigerenzer, G., Krauss, S., and Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, 84(3):343–352.

Hofmann, D. A., Griffin, M. A., and Gavin, M. B. (2000). The application of hierarchical linear modeling to organizational research. In Klein, K. J. and Kozlowski, S. W. J., editors, *Multilevel theory, research, and methods in organizations*, page 467–511. Jossey-Bass, San Francisco.

Hutchinson, J., Strickland, L., Farrell, S., and Loft, S. (2022). Human behavioral response to fluctuating automation reliability. *Applied ergonomics*, 105:103835.

Hutchinson, J., Strickland, L., Farrell, S., and Loft, S. (2023). The perception of automation reliability and acceptance of automated advice. *Human Factors*, 65(8):1596–1612.

Kaplan, A. D., Kessler, T. T., Brill, J. C., and Hancock, P. (2021). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, page 00187208211013988.

Koehler, D. J. and James, G. (2014). Probability matching, fast and slow. In *Psychology of learning and motivation*, volume 61, pages 103–131. Elsevier.

Laris, M. (2020). Fatal Tesla crash tied to technology and driver failures, NTSB says. *The Washington Post*.

Lee, J. D. (2008). Review of a pivotal human factors article:“humans and automation: use, misuse, disuse, abuse”. *Human Factors*, 50(3):404–410.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80.

*Lin, J., Matthews, G., Wohleber, R. W., Funke, G. J., Calhoun, G. L., Ruff, H. A., Szalma, J., and Chiu, P. (2020). Overload and automation-dependence in a multi-uas simulation: Task demand and individual difference factors. *Journal of Experimental Psychology: Applied*, 26(2):218.

Loftus, T. J., Filiberto, A. C., Balch, J., Ayzengart, A. L., Tighe, P. J., Rashidi, P., Bihorac, A., and Upchurch Jr, G. R. (2020). Intelligent, autonomous machines in surgery. *journal of surgical research*, 253:92–99.

Macmillan, N. A. and Creelman, C. D. (2005). *Detection theory: a user's guide*. Lawrence Erlbaum, Mahwah, NJ.

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

*Manzey, D., Gérard, N., and Wiczorek, R. (2014). Decision-making and response strategies in interaction with alarms: the impact of alarm reliability, availability of alarm validity information and workload. *Ergonomics*, 57(12):1833–1855.

Manzey, D., Reichenbach, J., and Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, 6(1):57–87.

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

*Mayer, A. K., Sanchez, J., Fisk, A. D., and Rogers, W. A. (2006). Don't let me down: The role of operator expectations in human-automation interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 2345–2349. SAGE Publications Sage CA: Los Angeles, CA.

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

McBride, S. E., Rogers, W. A., and Fisk, A. D. (2014). Understanding human management of automation errors. *Theoretical Issues in Ergonomics Science*, 15(6):545–577.

McDowell, M. and Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on bayesian reasoning. *Psychological bulletin*, 143(12):1273.

*McGuirl, J. M. and Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors*, 48(4):656–665.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, 53(4):356–370.

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

Merritt, S. M. and Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human factors*, 50(2):194–210.

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

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

Meyer, J. and Bitan, Y. (2002). Why better operators receive worse warnings. *Human Factors*, 44(3):343–353.

Meyer, J., Wiczorek, R., and Günzler, T. (2014). Measures of reliance and compliance in aided visual scanning. *Human Factors*, 56(5):840–849.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group*, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Annals of internal medicine*, 151(4):264–269.

Mueller, S. T. and Piper, B. J. (2014). The psychology experiment building language (pebl) and pebl test battery. *Journal of neuroscience methods*, 222:250–259.

Neyedli, H. F., Hollands, J. G., and Jamieson, G. A. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human factors*, 53(4):338–355.

NTSB (1995). Grounding of the Panamanian passenger ship Royal Majesty on Rose and Crown shoal near Nantucket, Massachusetts, June 10, 1995. Technical Report NTSB/MAR-97-01, National Transportation Safety Board, Washington DC.

NTSB (2020). Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator, Mountain View, California, March 23, 2018. Technical Report NTSB/HAR-20/01, National Transportation Safety Board, Washington DC.

*Okamura, K. and Yamada, S. (2020). Adaptive trust calibration for human-ai collaboration. *Plos one*, 15(2):e0229132.

Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2):230–253.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE transactions on systems, man, and cybernetics*, (3):257–266.

Rice, S. (2009). Examining single-and multiple-process theories of trust in automation. *The Journal of general psychology*, 136(3):303–322.

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

Riley, V. (1996). Operator reliance on automation: Theory and data. In *Automation and human performance*, pages 19–35. CRC Press.

Robinson, D. E. and Sorkin, R. D. (1985). A contingent criterion model of computer assisted detection. *Trends in ergonomics/human factors*, 2:75–82.

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

Sanchez, J. (2009). Conceptual model of human-automation interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 53, pages 1403–1407. SAGE Publications Sage CA: Los Angeles, CA.

Sanchez, J., Fisk, A. D., and Rogers, W. A. (2006). What determines appropriate trust of and reliance on an automated collaborative system? effects of error type and domain knowledge. In *2006 9th International Conference on Control, Automation, Robotics and Vision*, pages 1–6. IEEE.

*Sanchez, J., Rogers, W. A., Fisk, A. D., and Rovira, E. (2014). Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical issues in ergonomics science*, 15(2):134–160.

Sarter, N. B. (2006). Multimodal information presentation: Design guidance and research challenges. *International journal of industrial ergonomics*, 36(5):439–445.

Sarter, N. B. and Woods, D. D. (1997). Team play with a powerful and independent agent: Operational experiences and automation surprises on the airbus a-320. *Human factors*, 39(4):553–569.

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

Schuler, P. T. and Yang, X. J. (2023). Adapting to imperfect automation: The impact of experience on dependence behavior and response strategies. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, page 21695067231192408. SAGE Publications Sage CA: Los Angeles, CA.

Skitka, L. J., Mosier, K. L., and Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006.

Sorkin, R. D., Hays, C. J., and West, R. (2001). Signal-detection analysis of group decision making. *Psychological review*, 108(1):183.

Sorkin, R. D., Kantowitz, B. H., and Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors*, 30(4):445–459.

Sorkin, R. D. and Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human-computer interaction*, 1(1):49–75.

*Stanton, N. S., Ragsdale, S. A., and Bustamante, E. A. (2009). The effects of system technology and probability type on trust, compliance, and reliance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 53, pages 1368–1372. SAGE Publications Sage CA: Los Angeles, CA.

Tanner, W. P. and Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, 61(6):401–409.

Trevena, L. J., Bonner, C., Okan, Y., Peters, E., Gaissmaier, W., Han, P. K., Ozanne, E., Timmermans, D., and Zikmund-Fisher, B. J. (2021). Current challenges when using numbers in patient decision aids: advanced concepts. *Medical Decision Making*, 41(7):834–847.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Walliser, J. C., de Visser, E. J., and Shaw, T. H. (2016). Application of a system-wide trust strategy when supervising multiple autonomous agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 133–137. SAGE Publications Sage CA: Los Angeles, CA.

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

Wang, Y. and Yang, X. J. (2022). Humans working with un-reliable automation: Reverse psychology versus disuse Model. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 66(1):707–710. _eprint: https://doi.org/10.1177/1071181322661452.

Wickens, C. and Colcombe, A. (2007). Dual-task performance consequences of imperfect alerting associated with a cockpit display of traffic information. *Human factors*, 49(5):839–850.

Wickens, C. D. (2002a). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2):159–177.

Wickens, C. D. (2002b). Situation awareness and workload in aviation. *Current directions in psychological science*, 11(4):128–133.

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

Wickens, C. D., Gempler, K., and Morphew, M. E. (2000). Workload and reliability of predictor displays in aircraft traffic avoidance. *Transportation Human Factors*, 2(2):99–126.

Wickens, C. D., Helton, W. S., Hollands, J. G., and Banbury, S. (2021). *Engineering psychology and human performance*. Routledge.

Wickens, T. D. (2001). *Elementary signal detection theory*. Oxford university press.

*Wiczorek, R. and Manzey, D. (2014). Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior, and performance. *Human factors*, 56(7):1209–1221.

Wiczorek, R., Manzey, D., and Zirk, A. (2014). Benefits of decision-support by likelihood versus binary alarm systems: Does the number of stages make a difference? In *Proceedings of the human factors and ergonomics society annual meeting*, volume 58, pages 380–384. SAGE Publications Sage CA: Los Angeles, CA.

*Wohleber, R. W., Matthews, G., Lin, J., Szalma, J. L., Calhoun, G. L., Funke, G. J., Chiu, C.-Y. P., and Ruff, H. A. (2019). Vigilance and automation dependence in operation of multiple unmanned aerial systems (uas): A simulation study. *Human factors*, 61(3):488–505.

Yang, X. J. (2024). Humans working with imperfect diagnostic automation: A rational operator's behavior. Working paper.

Yang, X. J., Guo, Y., and Schemanske, C. (2022). From trust to trust dynamics: Combining empirical and computational approaches to model and predict trust dynamics in human-autonomy interaction. In *Human-Automation Interaction: Transportation*, pages 253–265. Springer.

Yang, X. J., Schemanske, C., and Searle, C. (2023). Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors*, 65(5):862–878.

Yang, X. J., Unhelkar, V. V., Li, K., and Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pages 408–416.

Yeh, M. (2000). *Attention and trust biases in the design of augmented reality displays.* University of Illinois at Urbana-Champaign.

Yoo, B., Donthu, N., and Lenartowicz, T. (2011). Measuring hofstede's five dimensions of cultural values at the individual level: Development and validation of cvscale. *Journal of international consumer marketing*, 23(3-4):193–210.