Trust and Performance in AI-Assisted Decision Making: The Role of Multi-class Error Types, Automation Stages, and Error Consequences

by

Jin Yong Kim

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
2026

Doctoral Committee:

Associate Professor X. Jessie Yang, Chair
Professor Lionel Robert
Professor Leia Stirling
Assistant Professor Oshin Tyagi

Jin Yong Kim
kjinyong@umich.edu
ORCID iD: 0009-0002-1626-7204

© Jin Yong Kim 2026

DEDICATION

To the community that sustained me: this work is a testament to the collective belief, patience, and love of those who walked beside me every step of the way.

ACKNOWLEDGEMENTS

I'm grateful to everyone who has been part of this journey. Reaching this milestone was never an individual effort. It was made possible by a community that consistently encouraged, challenged, and supported me. I feel incredibly fortunate to have been surrounded by such a dedicated family, friends, mentors, and peers.

First, I want to thank my advisor, Dr. Jessie Yang. I joined your lab as a research assistant during my undergraduate years, and I'm so thankful you saw something in me and encouraged me to pursue a PhD. That decision has shaped my life in ways I could not have imagined. From you, I learned not only how to conduct meaningful, rigorous research, but also what it means to mentor with purpose and care. Thank you for believing in me and for being such an influential presence in my life.

I'm also grateful to Dr. Leia Stirling. Since my senior year as your research assistant, you have been an important mentor to me in both research and leadership. I especially valued our work together on departmental K-12 outreach initiatives. Those experiences were meaningful and helped shape how I approach service and community engagement.

I would like to thank Drs. Lionel Robert and Oshin Tyagi for serving on my dissertation committee. I appreciate the time and effort you invested in my growth, the insightful feedback that strengthened this dissertation, and your guidance as I look toward the next stage of my career.

I am thankful for the mentors who offered feedback on my research goals and inspired me to lead and mentor others. At the University of Michigan, I thank Drs. Albert Berahas, Na Du, Marina Epelman, Corey Lester, Manhua Wang, and Amelia Warden. Within the HFES community, I thank Drs. Myrtede Alfred, HeeSun Choi, Heejin Jeong, Jeongho Kim, Minji Kim, Junho Park, Mike Rayo, and Changwon Son. I also want to honor the memory of the late Dr. Yili Liu, whose impact continues to shape and inspire our community.

I'm grateful to the incredible staff at IOE and C4E: Akosua, Alison, Chris, Cornelius, Denise, Mint, Rod, Saroya, Sheryl, Tyra, and Valerie. So much of your work happens behind the scenes, from coordinating research logistics to supporting outreach and departmental events. Thank you for being such a dedicated team.

This research would not have been possible without the students and healthcare professionals who generously gave their time to participate in these studies. I had the privilege

of working with undergraduate and graduate students through research (Alice, Catherine, David, Olivia, and Szu-Tung). Your curiosity and collaboration strengthened this work. I also appreciate the support of the Rackham Graduate School and the National Library of Medicine at the National Institutes of Health, whose funding made my doctoral training possible.

I feel lucky to have been part of such motivated and supportive student communities at Michigan, including the ICRL, Stirling Group, HFES, GAMP Team, SLB, and our various outreach teams. The friendships and shared experiences we built mean a lot to me. Thank you to Caleb, Cheoljoon, Doowon, Drew, Hannah, Hyesun, Jaeshin, Jacqueline, Jen, Maggie, Patrik, Robert, Sangyoon, Shreyas, and Szu-Tung. You brought so much laughter, encouragement, and perspective to this journey, especially during the most challenging moments.

To my host families: thank you for the steady love you've given me since I moved to the U.S. Your support created the stability that allowed me to build my life here with confidence.

To my parents, any success I've had begins with you. Your love shaped who I am, and I'm grateful for that. To my family-in-law, thank you for your kindness and unwavering support. Hearing about your own experiences building a life in the U.S. gave us perspective and reassurance as we navigated our own challenges.

Finally, to my wife, Grace: thank you for being there through every step of this process. Having you by my side made this journey possible.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES viii

LIST OF TABLES xi

LIST OF APPENDICES xii

LIST OF ACRONYMS xiii

ABSTRACT xiv

CHAPTER

1 Introduction **1**
- 1.1 Motivation 1
- 1.2 Background 5
  - 1.2.1 SDT, Trust, Automation Compliance and Reliance 5
  - 1.2.2 Levels, Stages, and Degrees of Automation 7
  - 1.2.3 Properties of Trust Dynamics 10
  - 1.2.4 Uncertainty Presentation 12
- 1.3 Research Objectives 14
- 1.4 Dissertation structure 15

2 The Effects of Presenting AI Uncertainty Information on Pharmacists' Trust in Automated Pill Recognition Technology: an Exploratory Study **17**
- 2.1 Introduction 17
- 2.2 Methods 19
  - 2.2.1 Ethical Considerations 19
  - 2.2.2 Recruitment and Participants 19
  - 2.2.3 AI Model 19
  - 2.2.4 Experimental Testbed 20
  - 2.2.5 Experimental Design 24
  - 2.2.6 Measures 24
  - 2.2.7 Experimental procedure 25

2.2.8 Statistical Analysis . . . 25
2.3 Results . . . 26
2.3.1 Trust Propensity . . . 26
2.3.2 End-Trust Towards AI Aid . . . 26
2.3.3 Average Trust Towards AI Aid . . . 26
2.3.4 Trust Adjustment . . . 27
2.4 Discussion . . . 29
2.4.1 Principal Findings . . . 29
2.4.2 Comparison to Prior Work . . . 29
2.4.3 Limitations and Future Directions . . . 33
2.5 Conclusion . . . 34

3 Beyond Binary Decisions: Evaluating the Effects of AI Error Type on Trust and Performance in AI-Assisted Tasks 35
3.1 Introduction . . . 35
3.2 Methods . . . 38
3.2.1 Ethical Considerations . . . 38
3.2.2 Recruitment and Participants . . . 38
3.2.3 Experimental apparatus and stimuli . . . 38
3.2.3.1 Stimuli development . . . 38
3.2.3.2 Experimental Task . . . 40
3.2.4 Experimental Design . . . 43
3.2.5 Measures . . . 44
3.2.6 Experimental procedure . . . 44
3.2.7 Statistical Analysis . . . 45
3.3 Results . . . 45
3.3.1 Comparing Patterns WC, WI, and WI_{ref} . . . 46
3.3.2 Evaluating Continuity, Negativity Bias, and Stabilization . . . 48
3.4 Discussion . . . 50
3.5 Conclusion . . . 52

4 Evaluating the Effects of Consequences of Error on Trust in AI-Aided Task: An Eye Tracking Study 54
4.1 Introduction . . . 54
4.2 Methods . . . 55
4.2.1 Ethical Considerations . . . 55
4.2.2 Recruitment and Participants . . . 55
4.2.3 Experimental Testbed and Stimuli . . . 55
4.2.3.1 Mental Rotation Task . . . 55
4.2.3.2 Eye tracking . . . 57
4.2.4 Experimental Design . . . 58
4.2.5 Measures . . . 58
4.2.6 Experimental procedure . . . 59
4.2.7 Statistical Analysis . . . 59
4.3 Results . . . 59

4.3.1 Occurrence of Error Patterns Across Consequence Levels 59
4.3.2 Within-Pattern Effects of Consequence 60
4.4 Discussion 61
4.5 Conclusion 64

5 Evaluating the Effects of Stages of Automation (SoA) on Performance and Trust in AI-Assisted Tasks 65
5.1 Introduction 65
5.1.1 Proposition 66
5.1.2 Hypotheses 68
5.2 Methods 69
5.2.1 Ethical Considerations 69
5.2.2 Participants 69
5.2.3 Experimental Task 69
5.2.4 Experimental Design 72
5.2.5 Measures 72
5.2.6 Experimental procedure 74
5.2.7 Statistical Analysis 74
5.3 Results 75
5.3.1 Comparing between SoA when AI is correct and incorrect 75
5.3.2 Comparing between SoA when Unintended Safeguards occur 76
5.3.3 Comparing Between SoA (Stage-2-AI vs. Stage-3-AI) 77
5.4 Discussion 78
5.4.1 Limitations 81
5.5 Conclusion 81

6 Conclusion and Future Directions 83
6.1 Summary, Intellectual Merit, and Practical Impact 83
6.1.1 Connecting to Existing Literature 83
6.1.2 Intellectual Merit 86
6.1.3 Practical Impact 86
6.2 Future Work 87
6.2.1 Considerations for Pharmacy Technology Implementation 87
6.2.2 Extensions of the Proposed Theoretical Frameworks 88

APPENDICES 90

BIBLIOGRAPHY 95

LIST OF FIGURES

FIGURE

1.1 Difference in the performance patterns when choices are binary or non-binary. In the binary choice (a), the Wrong-Incorrect pattern always matches the reference. However, in the non-binary choice (b), the Wrong-Incorrect pattern could be categorized into two, based on whether incorrect predictions match the reference (prescribed pill). 4
1.2 Signal Detection Theory (SDT) 5
1.3 The degrees of automation (DoA) framework combines the dimensions of levels of automation (LoA) and stages of automation (SoA), represented as the diagonal-dashed line. (Wickens et al., 2010) 8

2.1 Interface for the AI aid without and with the uncertainty plot. Checkboxes indicate AI's recommendation. When all four checkboxes are green, the AI advises to accept; otherwise, to reject. The histogram shows the distribution of the 50 predicted probabilities associated with the predicted NDC. 22
2.2 Four potential AI performance patterns according to Signal Detection Theory. 23
2.3 Mean end (left) and average (right) trust by AI aid help type (with or without the uncertainty plot). The error bars represent a 95% confidence interval. . . 27
2.4 Trust adjustment for different AI performance patterns. Red shades represent negative trust adjustment, and blue shades represent positive trust adjustment. 28
2.5 Autocorrelation of trust as a function of time separation. The blue solid line represents AI aid with the uncertainty plot, and the red dashed line represents the AI aid without the uncertainty plot. The error bars represent +/- 2 standard errors. 29
2.6 Uncertainty plot with narrower (left) and wider (right) interquartile range (IQR). IQR is a measure of statistical dispersion, or how spread out the data points are. 31
2.7 Mean interquartile range (IQR) by outcome pattern. The error bars represent a 95% confidence interval. 32

3.1 Illustration of stimuli development. After creating the (a) reference images, one (b) mirrored image, fourteen 45-degree (c) horizontal axis rotated figures, and fourteen 45-degree (d) vertical axis rotated figures were created for each reference image. 39

3.2 For the beta test, participants were asked to determine whether the image pairs are the same items in different rotations (by clicking "Yes, they are the same item") or different items (by clicking "No, they are different") within 10 seconds. 40
3.3 Flowchart of the experiment 41
3.4 Participants were asked to make their initial answer choice within 15 seconds. 41
3.5 Participants were presented with the AI system's recognition and chose between sticking with or rejecting their initial choice within 10 seconds. If no selection was made, they were skipped to the next trial. 42
3.6 Participants were presented with the performance feedback page, showing their performance on the initial and final answer choices and the validity of the AI's prediction. 43
3.7 Trust adjustment (left) and Performance (right) by patterns of performance. The error bars represent +/- 2 standard errors. 46
3.8 Reaction Time by patterns of performance. The error bars represent +/- 2 standard errors. 48
3.9 Autocorrelation of trust as a function of time separation. The error bars represent +/- 2 standard errors. 49
3.10 Trust adjustment by patterns of performance. The error bars represent +/- 2 standard errors. 50

4.1 Participants were presented with the reference shape and five answer selections, and were instructed to make an initial choice. The low-consequence trials appeared on blue backgrounds. 56
4.2 Pattern *RI* (Right-Incorrect) in low- and high-consequence appeared on blue and red backgrounds, respectively. 56
4.3 Performance feedback of high-consequence pattern *RI*. In high-consequence scenarios, initial selection, final selection, and performance feedback appeared on red backgrounds. 57
4.4 Eye tracking areas of interests (AOIs). 57

5.1 AI-assisted medication dispensing procedure is displayed with the probabilities. The red text indicates incorrect classifications and advice. Note that the last row of AI incorrectly predicting Y as Z (pattern *WI*, Stage-2-AI) results in a correct advice to "refill" (pattern $WC*$, Stage-3-AI). 67
5.2 Flowchart of the SoA experiment. 70
5.3 Participants selected an initial answer within 15 seconds. Their initial selection was highlighted in yellow. 71
5.4 AI advice was presented in Stage-2-AI or Stage-3-AI format. The Stage-2-AI displayed its prediction, oriented to closely match the angle of the reference image. The Stage-3-AI provided explicit advice. 71
5.5 Feedback was presented following the final verification selection. Feedback included the accuracy of participants' initial selection (right or wrong), the accuracy of AI prediction (correct or incorrect), and the accuracy of participants' final selection (correctly committed or incorrectly committed). 72

5.6 Comparing system usability score (SUS) rating, post-condition trust rating, and post-condition perceived reliability between the two SoA 78

B.1 After each trial i, participants reported their $trust(i)$ in the AI. 93
B.2 Participants rated their confidence using a visual analog scale. 94
B.3 Participants rated their trust and perceived reliability using a visual analog scale. 94

LIST OF TABLES

TABLE

2.1 Percentage of medication characteristics (colors on the left, shapes on the right) featured in the training dataset. 20
2.2 Percentage of medication characteristics (colors on the left, shapes on the right) featured in trials for reference images and filled images. 21

3.1 Description of the five cases and two outcomes. Bold text indicates potential incorrect outcomes resulting from automation errors. 36
3.2 Mean and sd of confidence rating, trust adjustment, final performance, final reaction time, number of patterns, and number of participants grouped by patterns and outcomes of performance. (*Note: Performance is based on the final answer choice. Therefore, for each pattern, there is only one output.*) . 47

4.1 Results table summarizing dependent measures by error patterns. 60
4.2 Results table summarizing eye tracking measures by error patterns. 62

5.1 Description of the five performance patterns based on human initial selection, AI classification, and AI advice. At Stage 2, pattern *WI* constitutes an incorrect classification that inadvertently prompts a safety check. At Stage 3, although the underlying AI classification remains incorrect, the AI recommendation to re-check is correct from a decision-support perspective, explicitly highlighting the unintended safeguard. 66
5.2 Outcome type was defined based on the correctness of the Stage-2 classification and the resulting Stage-3 recommendation, and was categorized as simultaneously correct, simultaneously incorrect, and unintended safeguard. . 73
5.3 Counts of ΔPerformance categories and means (with standard deviations) of reaction time and trust adjustment, grouped by SoA and AI Correctness. . . 75
5.4 Counts of ΔPerformance categories and means (with standard deviations) of reaction time and trust adjustment, grouped by Final Response Accuracy and SoA, for unintended safeguard. 76
5.5 Counts of ΔPerformance categories and means (with standard deviations) of reaction time and trust adjustment, grouped by SoA. The table presents the marginal average. 77

LIST OF APPENDICES

APPENDIX

A Questionnaires **90**

B Testbed Supporting Figures **93**

LIST OF ACRONYMS

AI Artificial Intelligence

AOI Areas Of Interest

BNN Bayesian Neural Network

CID Combat Identification

CNN Convolutional Neural Network

CR Correct Rejection

DoA Degree of Automation

FA False Alarm

IQR Interquartile Range

LoA Level of Automation

MRT Mental Rotation Task

NDC National Drug Code

SDT Signal Detection Theory

SoA Stages of Automation

ABSTRACT

Artificial intelligence (AI) systems are increasingly integrated into decision-making processes across a wide range of domains, including safety-critical settings such as healthcare, transportation, and defense. While these systems show significant potential for enhancing performance and safety, their effectiveness depends critically on how humans interact with them. This dissertation investigates the challenges and complexities of AI-assisted decision-making to enhance overall system safety by addressing the following questions: (1) How does AI transparency influence human trust? (2) How do different AI error patterns affect human trust and performance? (3) How do the consequences of errors influence human trust and reliance? and (4) How do varying stages of automation affect human-AI interaction?

First, this work presents one of the earliest empirical studies of pharmacists' trust in automated pill verification technology, introduced to mitigate the risk of incorrect medication dispensing. Grounded in human factors and cognitive systems engineering principles, and informed by user feedback, the study demonstrates that transparent presentation of AI uncertainty enhances pharmacists' trust, highlighting the importance of user-centered design in clinical AI systems.

Second, this dissertation moves beyond traditional binary performance categorization using Signal Detection Theory (SDT), which classifies outcomes as correct or incorrect, by introducing a multi-class decision framework that captures the complexity of modern AI systems (i.e., an image classification AI system identifying one medication among thousands of alternatives). This framework reveals distinct AI error patterns with varying effects on trust and performance, including cases of false reassurance that degrade performance and paradoxical errors that inadvertently improve safety by prompting verification.

Third, the work examines how varying error consequences influence trust and reliance, showing that users adapt their behavior in high-consequence scenarios by reducing reliance on AI and increasing vigilance, regardless of system accuracy. These findings highlight the importance of incorporating consequence awareness into AI system design.

Finally, this dissertation investigates the role of stages of automation (SoA), defined as the allocation of automation support across information-processing stages (e.g., information acquisition, analysis, decision selection, and action implementation) (Parasuraman

et al., 2000). The results demonstrate that higher stages of automation can yield superior interaction-level performance when evaluated through the proposed multi-class framework, particularly when unintended safeguards are present, cases in which classification errors at Stage-2 result in correct Stage-3 recommendations that are not captured by the binary SDT framework.

Together, this dissertation advances the theoretical understanding of trust in automation/autonomy/AI systems by expanding existing human-AI interaction models as a dynamic, multi-class process influenced by error type, consequence severity, and automation stages. Practically, these findings provide design-relevant insights for developing transparent, human-centered AI systems to support safety-critical decision-making.

CHAPTER 1

Introduction

1.1 Motivation

Automation/autonomy/AI systems have the potential to enhance safety and operational efficiency across a wide spectrum of human activities. While routine tasks are increasingly supported by AI assistants, advances in large language models (LLMs) have enabled AI systems to generate, summarize, and reason over complex datasets, offloading human cognitive effort. Beyond everyday applications, AI systems are increasingly deployed into safety-critical domains, such as healthcare (Cheikh et al., 2022; Asan et al., 2020; Gombolay et al., 2018; Wang et al., 2015), transportation (e.g., aviation, space, and autonomous driving (Azevedo-Sa et al., 2020; Lee and Kolodge, 2020; Ayoub et al., 2023; Larson and Stirling, 2025)), workforce management (Robert et al., 2020), and defense (Bhat et al., 2024a). In healthcare, for instance, AI systems that interpret complex data to generate actionable suggestions are recognized by practitioners as valuable tools for enhancing clinical decision-making (Asan et al., 2020; Lester et al., 2021; Tsai et al., 2025).

The integration of these systems represents a fundamental shift toward AI-assisted decision-making paradigms aimed at mitigating limitations in human judgment and addressing uncertainty inherent in real-world environments to support safer task performance. However, despite AI's potential to strengthen these fields, the effectiveness of human interactions with automation/autonomy/AI systems, and their safe integration into practice, depends on the establishment of trust in these systems (Du et al., 2020; Hoff and Bashir, 2015; Yang et al., 2023a; Schaefer et al., 2016). In this dissertation, trust is defined according to Lee and See (2004) as "the attitude that an agent will help achieve an individual's goals in situations characterized by uncertainty and vulnerability," highlighting human reliance and situational uncertainty.

Existing trust in automation/autonomy/AI systems research typically involves a human subject performing tasks, with certain tasks being automated. For example, an automated

combat identification (CID) aid can scan the environment, identify a friend or a foe, and make recommendations to soldiers (Wang et al., 2009; Du et al., 2020; Neyedli et al., 2011). To model the performance of the automation, most studies used the signal detection theory (SDT) (Tanner and Swets, 1954; Hautus et al., 2021). SDT evaluates how AI systems distinguish signals from noise, categorizing the state of the world as either "signal present" or "signal absent." Based on SDT categorization, AI performance results in four outcomes: hit (error flagged correctly), miss (error not flagged), false alarm (FA: no error, but flagged incorrectly), and correct rejection (CR: no error, no flag) (Tanner and Swets, 1954; Hautus et al., 2021). Building on this framework, prior research has examined how error types (misses and FAs) distinctly induce distinct behavioral responses, such as reliance and compliance, referred to as operator dependence (Chancey et al., 2017; Meyer, 2001). Correct identifications (hit and CR) increase trust, while incorrect ones (FA and miss) decrease trust (Lee and Moray, 1992; Swets, 1992). Research also indicates that FAs typically reduce trust less than misses, prompting designers to design more liberal systems (i.e., more willing to flag an error) with a higher rate of false alarms to ensure potential issues are flagged (Davenport and Bustamante, 2010; Stanton et al., 2009).

The binary categorization using SDT has proven effective for diagnostic decision-making (Bhat et al., 2022; Yang et al., 2017; Chancey et al., 2017). However, this framework oversimplifies the complex dynamics and variances observed in real-world AI-assisted decision-making contexts, which are often "multi-class" in nature, involving multiple possible classification outcomes associated with an AI prediction, and consequently resulting in non-binary error structures.

Consider the following hypothetical scenario:

> Sarah, a highly skilled pharmacist, is responsible for filling the medication bottles for each prescription order. Recently, the pharmacy Sarah works for has introduced an AI computer vision system that can scan the filled bottle and identify the specific medication that has been filled. The AI system is introduced as another layer of verification before medication dispensing. Today, Sarah receives a prescription order for one patient, Noah, who needs to take medication $\boldsymbol{X}$.

Upon receiving the order, one of the following five cases could happen:

> Case A: Sarah correctly fills the bottle with medication $\boldsymbol{X}$. The AI system scans the bottle and correctly identifies that the filled medication is $\boldsymbol{X}$ and signals a green light.

Case B: Sarah correctly fills the bottle with medication ***X***. The AI system scans the bottle but incorrectly identifies that the filled medication is ***Y*** (or any medication that is **not** X). As medication ***Y*** is not the ordered medication, the AI system signals a red flag and indicates the need to double-check.

Case C: Sarah has a lapse when filling the bottle and incorrectly fills it with medication ***Z***. Fortunately, the AI system correctly identifies the filled medication as ***Z***. As medication ***Z*** is not the ordered medication, the AI system signals a red flag.

Case D: Sarah has a lapse when filling the bottle and incorrectly fills it with medication ***Z***. The AI system scans the filled medication and identifies it as medication ***Y*** (or any medication that is **not** ***X*** and **not** ***Z***). As medication ***Y*** is not the ordered medication, the AI system signals a red flag.

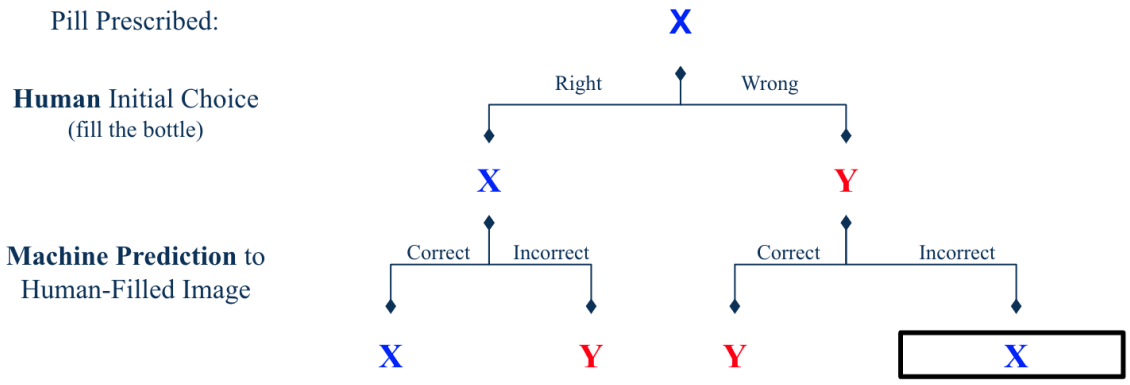
Case E: Sarah has a lapse when filling the bottle and incorrectly fills it with medication ***Z***. The AI system scans the filled medication and incorrectly identifies it as medication ***X*** and signals a green light.

This hypothetical scenario, which reflects recent developments in medical dispensing (Lester et al., 2021; Zheng et al., 2023; Chen et al., 2024), presents unique characteristics that are largely overlooked in existing research on trust in and dependence on automation/autonomy and AI systems. Let's examine the five cases presented in the hypothetical scenario, focusing particularly on Case D and Case E shown in Figure 1.1b, which are especially thought-provoking. In Case D, although the AI erroneously identifies medication Y as Z, this error can be considered beneficial from a utilitarian perspective. The incorrect identification differs from the prescribed medication X, which rightly triggers a safety alert. Conversely, Case E also involves an error by the AI, but this mistake could potentially lead to catastrophic results, as it mistakenly confirms the incorrectly filled medication as correct, misleading the human operator and providing false reassurance.

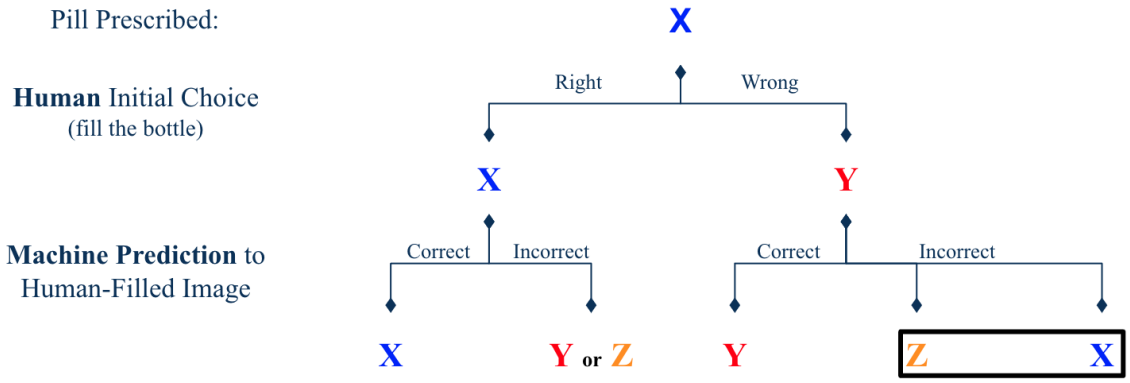
As illustrated in Figure 1.1a, when the traditional SDT framework is applied under the assumption of binary diagnostic decision-making, Cases D and E are combined into a single category of "incorrect" classification. Examining trust in and dependence on automation/autonomy/AI systems beyond the binary classification is therefore important, as the SDT framework does not capture the inherent multi-class classification scenarios that are increasingly prevalent in advanced AI systems.

The following literature review provides important background information for the topics addressed in this dissertation. Section 1.2.1 introduces Signal Detection Theory (SDT) and its application in assessing the effectiveness of human-automation interaction, with particular

emphasis on trust, compliance, and reliance. Section 1.2.2 reviews human factors principles related to automation implementation, including frameworks of levels, stages, and degrees of automation. Section 1.2.3 examines prior work on trust dynamics, highlighting the need to investigate trust as a dynamic, moment-to-moment construct. Finally, Section 1.2.4 reviews approaches to presenting uncertainty aimed at enhancing system transparency.



(a) Patterns from **binary choices**. Wrong-Incorrect always matches the reference (prescribed pill).



(b) Patterns from **non-binary choices**.

Figure 1.1: Difference in the performance patterns when choices are binary or non-binary. In the binary choice (a), the Wrong-Incorrect pattern always matches the reference. However, in the non-binary choice (b), the Wrong-Incorrect pattern could be categorized into two, based on whether incorrect predictions match the reference (prescribed pill).

1.2 Background

1.2.1 SDT, Trust, Automation Compliance and Reliance

SDT models the relationship between signals and noise, as well as the automation's ability to detect signals among noise (Tanner and Swets, 1954; Sorkin and Woods, 1985). The state of the world is characterized by either "signal present" or "signal absent," which may or may not be identified correctly by automation. The combination of the state of the world and the automation's detection results in four possible states: *hit, miss, false alarm (FA)*, and *correct rejection (CR)* (see Figure 1.2).

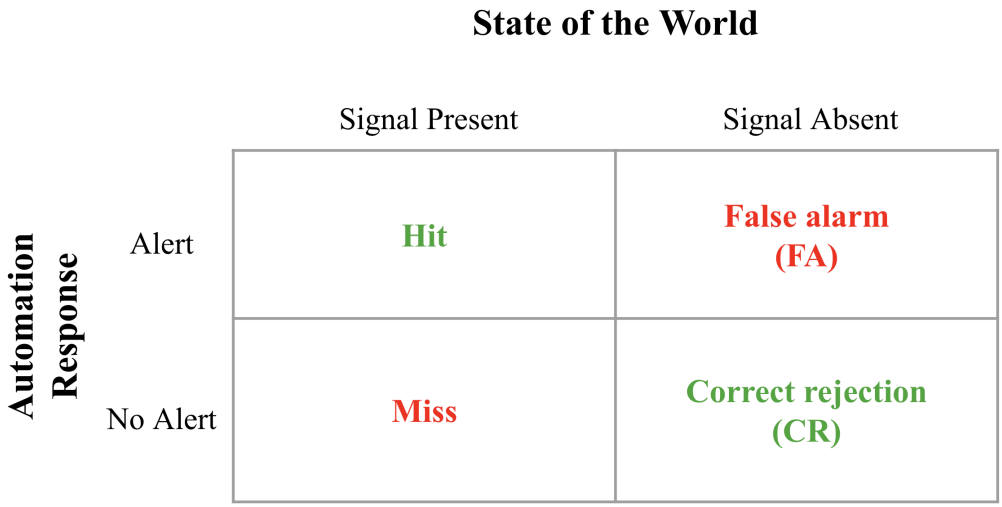


Figure 1.2: Signal Detection Theory (SDT)

SDT is widely employed to examine trust in and dependence on automation/autonomy and AI systems. Typically, SDT is used to model the reliability of automation with which the human operators interact in both single-task (Wang et al., 2009; Neyedli et al., 2011; Lacson et al., 2005; Madhavan et al., 2006; Wiegmann et al., 2001; Bhat et al., 2022, 2024b) and multi-task paradigms (Yang et al., 2017; Chung and Yang, 2024; Du et al., 2020; Wiczorek and Manzey, 2014; McBride et al., 2011). The general findings indicate that as automation reliability increases, trust and dependence on automation increase accordingly.

When error types (*FAs* and *misses*) are examined more closely, findings suggest they lead to distinct behavioral patterns (Meyer, 2001, 2004; Chancey et al., 2017). Dependence on automation can be categorized into compliance and reliance. Compliance is characterized by operators responding as though the error exists when the system indicates an error. *FAs* often result in commission errors when the operator follows the automation's false indication. Reliance, by contrast, occurs when operators believe the automated system's

indication of safety. When operators fail to respond to events due to automation *misses*, omission errors arise (Meyer, 2001, 2004; Skitka et al., 1999). Research suggests that *FAs* primarily influence compliance, though some studies indicate they may also affect reliance, whereas *misses* predominantly impact reliance (Dixon and Wickens, 2006; Dixon et al., 2007; Chancey et al., 2017; Meyer, 2001; Chancey et al., 2015).

For example, Chancey et al. (2017) investigated automation reliability and error types (*FA*-prone vs. *miss*-prone) in a simulated flight task. Participants performed a primary task of maintaining level flight and a secondary resource management task supported by automated aids with high (90%) and low (60%) reliability. Trust was measured at an aggregate level halfway through each session. The results showed that reliability significantly affected compliance in the *FA*-prone condition, with participants more likely to comply with the highly reliable systems. In contrast, in the *miss*-prone condition, reliability primarily affected reliance, with greater reliance observed on highly reliable systems. Additionally, *FA*s had a stronger impact on trust than *misses*; trust mediated the relationship between *FAs* and compliance but not between *misses* and reliance (Chancey et al., 2017).

Other research indicates that *misses* may cause operators to trust automation less. For instance, in a study simulating an unmanned aerial vehicle mission with an automation-assisted secondary weapon-deployment task, Davenport and Bustamante (2010) investigated participants' trust, reliance, and compliance following *FA*-prone and *miss*-prone aid. Trust was assessed through post-session questionnaires, revealing a greater trust decrement for systems with more *misses* compared to *FAs*. Compliance was greater for the *miss*-prone aid, while reliance was greater for the *FA*-prone aid (Davenport and Bustamante, 2010). Conversely, other studies indicate that operators may exhibit lower trust in systems with a high frequency of *FAs*, which may be attributed to the heightened prominence of *FAs* that reduces the credibility as operators need to actively verify the presence of errors, making the error more salient (Johnson et al., 2004; Breznitz, 2013).

Some studies implicitly assessed compliance and reliance using performance and reaction time metrics. Dixon et al. (2007) examined error types in tracking and system monitoring tasks. Participants were grouped by automation aid type: no help, perfectly reliable aid, *FA*-prone aid (60% reliability), and *miss*-prone aid (60% reliability). Reliance was evaluated through performance deficits, while compliance was measured by comparing response times to those observed in the perfectly reliable aid condition. Results indicated that the *FA*-prone system reduced both compliance and reliance, while the *miss*-prone system only reduced reliance. In the *FA*-prone condition, failure detection rates and response times in the monitoring task were worse than in the perfectly reliable automation condition, reflecting reduced compliance. Both error types affected reliance as evidenced by poorer tracking task

performance after silent automation trials. Notably, the *FA*-prone condition resulted in lower performance and longer detection times compared to the *miss*-prone condition (Dixon et al., 2007). In Sanchez et al. (2014)'s multi-task simulation experiment, participants performed a collision avoidance driving task supported by 94.5% reliable automation and a secondary tracking task. Reliance was assessed based on whether participants chose to double-check after receiving support from the automated aid. Results revealed a trend of increasing reliance in the *FA* condition as participants became more familiarized with automation, whereas reliance decreased in the *miss* condition (Sanchez et al., 2014).

In high-stakes scenarios where the consequences of missing critical events could be catastrophic, automation thresholds are often calibrated to favor *FAs* to minimize the chance of overlooking abnormalities (Sorkin and Woods, 1985; Sanchez et al., 2014). However, excessive *FAs* may affect trust more than *misses*, leading to disuse (neglect or under-dependence), deviating from the original purpose (Chancey et al., 2017; Parasuraman and Riley, 1997; Dixon and Wickens, 2006). These studies treated the error type (*FA*-prone and *miss*-prone) as a between-subject variable. Future research could benefit from evaluating within-subject differences to provide a more comprehensive understanding of human behavioral patterns in response to automation errors.

1.2.2 Levels, Stages, and Degrees of Automation

Researchers have developed frameworks to describe automation implementation, notably levels, stages, and degrees of automation (Parasuraman et al., 2000; Wickens et al., 2021, 2004; Sheridan et al., 1978).

In a seminal study, Sheridan et al. (1978) explored how control should be shared between humans and automation in underwater teleoperation systems, such as deep-sea exploration and maintenance, in which manual operation raises a significant risk to human life. The authors introduced the levels of automation (LoA) framework, which categorizes the degree of control shared between humans and automation to reflect increasing levels of automation authority and involvement. The 10-level spectrum ranged from level 1, representing full manual control without any automation assistance, to level 10, representing full automation control without human input.

Expanding on LoA, Parasuraman et al. (2000) proposed stages of automation (SoA), which categorize automation by functional role in the human information processing. These stages are: (1) information acquisition/filtering that assists selective attention, (2) information analysis that supports human working memory by interpreting and processing the data, (3) decision and action selection assisting decision making, and (4) action implementation

supporting human execution response. The performance gap between stage 2 and stage 3 automation has also been conceptualized as the distinction between status (informs the operator of system status) and command (provides action recommendations to the operators) automation (Wickens et al., 2021).

To integrate LoA and SoA, the degrees of automation (DoA) framework combined both dimensions into a two-dimensional space, plotting LoA vertically and SoA horizontally (Wickens et al., 2010). In this representation, each stage of automation can operate at different levels of automation, meaning that a given stage may involve lower or higher levels of automated support depending on how much authority is allocated to the system (Parasuraman et al., 2000; Hart and Valasek, 2009). DoA was therefore defined as an ordinal scale (diagonal-dashed line, Figure 1.3) representing the overall amount of automated support across stages and levels, which provides a more detailed measure of how automation supports human operators (Wickens et al., 2010). In this framework, increasing DoA has been suggested to impair situational awareness (SA), contributing to the operators being "out-of-the-loop," which could result in unexpected and undetected failures, phenomena described as the lumberjack analogy and black swan events (Wickens et al., 2010, 2004; Sebok and Wickens, 2017).

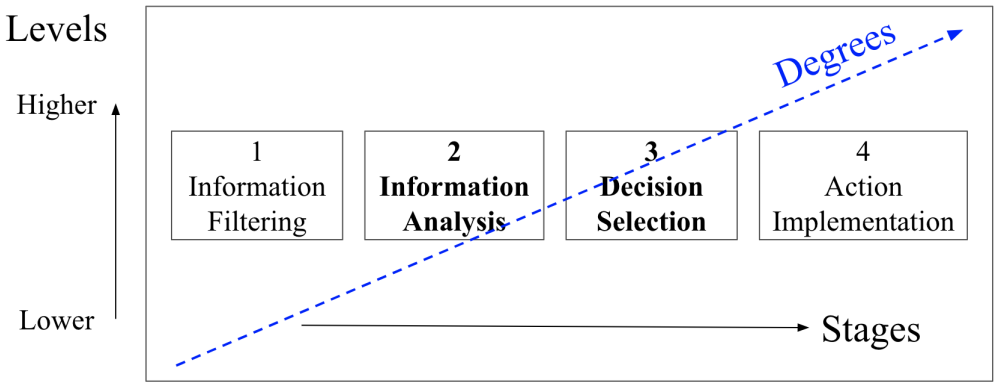


Figure 1.3: The degrees of automation (DoA) framework combines the dimensions of levels of automation (LoA) and stages of automation (SoA), represented as the diagonal-dashed line. (Wickens et al., 2010)

Empirical studies have applied these frameworks to examine performance outcomes across varying stages and degrees (Li et al., 2014; Picco et al., 2025; Gegoff et al., 2025; Wickens et al., 2007; Crocoll and Coury, 1990; Sarter and Schroeder, 2001; Rovira et al., 2002, 2007; Sethumadhavan, 2009; Tatasciore et al., 2020; Wang et al., 2025). We highlight representative examples of studies comparing information analysis (IA) and decision selection (DS) automation, the two stages selected for investigation in Chapter 5, as this transition repre-

sents the critical boundary where the operator's cognitive role shifts from active information processing to supervisory monitoring (Onnasch and Hösterey, 2019; Onnasch et al., 2014). The studies reviewed below are organized to illustrate common experimental paradigms based on two factors: trial reliability (whether performance was reported separately for correct and incorrect trials, or presented as an overall comparison) and task structure (single- or multi-task, defined as an automated task embedded within additional manual or monitoring tasks).

In single-task experiments under reliable conditions, DS automation often led to better performance and faster reaction time (Picco et al., 2025; Gegoff et al., 2025; Wickens et al., 2007). In these studies, DS automation transformed or prioritized IA automation information into actionable guidance. For example, in a driving simulator study, Picco et al. (2025) examined feedback systems that varied in SoA. Under IA conditions, feedback provided situational assessments (i.e., numerical values of the speed and distance to the lead vehicle), whereas under DS conditions, the same information was synthesized into explicit maneuver advice (i.e., "slow down"). Participants in the DS condition exhibited better speed-control performance with fewer speeding violations. Similarly, Wickens et al. (2007) reported overall benefits of DS automation in a flight recovery simulation. The IA display indicated aircraft attitude information, whereas the DS display presented an icon indicating the appropriate recovery direction. Pilots using DS displays initiated corrective actions more quickly and committed fewer flight control errors than those using IA displays. Gegoff et al. (2025) conducted an uninhabited vehicle (UV) management experiment in which participants selected the optimal UV for mission execution. The DS-equivalent condition provided recommendations without additional information, whereas the IA-equivalent condition provided additional informational cues without explicit recommendations. The DS condition led to faster decision times than the IA condition, although accuracy did not differ between stages.

In single-task experiments that distinguished between correct and incorrect trials, DS automation generally led to worse performance during incorrect trials (Crocoll and Coury, 1990; Sarter and Schroeder, 2001; Rovira et al., 2007). For instance, in an aircraft identification task, Crocoll and Coury (1990) compared IA automation that identified aircraft status (friend, hostile, or unidentified) with DS automation that recommended whether to "fire." Identification performance was comparable across SoA in correct trials, but DS automation produced significantly worse outcomes in incorrect trials. Similarly, in a flight simulation experiment by Sarter and Schroeder (2001), pilots used no aid, IA aid (top-down view indicating ice location), or DS aid (maneuver recommendations) during icing encounters. In correct trials, both aids improved icing escaping performance compared to no aid. How-

ever, in incorrect trials, DS aid resulted in greater performance decline than IA (Sarter and Schroeder, 2001). Rovira et al. (2007) observed similar effects in a sensor-to-shooter targeting simulation. IA automation presented a complete list of possible engagement combinations, and DS automation prioritized the options, presenting the best choice first. Compared to IA automation, DS automation improved performance and response time in correct trials, but led to significantly worse performance in incorrect trials (Rovira et al., 2007).

In multi-task experiments, findings were mixed (Rovira et al., 2002; Sethumadhavan, 2009; Tatasciore et al., 2020; Wang et al., 2025). Rovira et al. (2002) conducted a Multi-Attribute Task Battery (MATB) simulation experiment involving a manual tracking task, a fuel management task, and an automation-assisted malfunction monitoring task. IA automation identified faulty gauges, whereas DS automation recommended corrective actions. Malfunction detection performance was similar across stages for correct trials; however, DS automation worsened performance than IA automation during incorrect trials. worse performance. Across all trials, DS automation resulted in quicker detection times (Rovira et al., 2002). Contrasting results were found in Wang et al. (2025)'s MATB experiment, using similar automation system designs as Rovira et al. (2002). DS automation improved response speed under correct condition but no performance difference emerged between DS and IA automation under failure conditions. Sethumadhavan (2009) examined an air traffic conflict detection task in which participants used automation to avoid collisions, while monitoring a weather display. The IA system highlighted aircraft pairs projected to be in conflict, whereas DS automation recommended conflict-resolving actions. Participants using IA automation detected automation failures more quickly, mediated by higher situational awareness (SA). In contrast, Tatasciore et al. (2020) found no difference in SA between SoA under incorrect trials. In their simulated submarine track management experiment, participants performed two tasks assisted by automation (contact classification & closest point of approach estimation) and a manual dive task. IA automation predicted track history to support classification, whereas DS automation added salient alerts (i.e., flashing history). Under reliable trials, DS aid enhanced performances. However, under unreliable trials, performance differences between IA and DS automation disappeared (Tatasciore et al., 2020).

1.2.3 Properties of Trust Dynamics

Over the past 30 years, many researchers have investigated people's trust in automated/autonomous and AI systems. Historically, most studies have approached trust from a *snapshot* view, typically using end-of-experiment questionnaires to evaluate trust levels. Research within this view has identified many factors that can influence people's trust in

these systems (Hoff and Bashir, 2015; Kaplan et al., 2023). More recently, acknowledging that a person's trust can change dynamically while interacting with automated/autonomous technologies, there is a shift of research focus from snapshot trust to trust dynamics – how humans' trust in autonomy forms and evolves due to moment-to-moment interaction with automated/autonomous technologies (de Visser et al., 2020; Yang et al., 2023a; Wischnewski et al., 2023; Yang et al., 2023b).

One body of research on trust dynamics explores how trust can be violated and restored after trust violations. Trust violations happen after negative experiences (*false alarms* and *misses*), which substantially decreases trust to a level below the original trust level (Esterwood and Robert Jr, 2023; Esterwood and Robert, 2022; Kim et al., 2004; Lewicki and Brinsfield, 2017; Sebo et al., 2019; Kim and Song, 2021; Baker et al., 2018). Some trust violation research suggests that *misses* may lead to lower trust in automation. For example, Azevedo-Sa et al. (2020) investigated how failures in semi-autonomous driving systems influence drivers' trust. Participants engaged in a non-driving task in a semi-autonomous vehicle, which occasionally provided *FAs* or *missed* obstacles. Trust dynamics were evaluated after each interaction, revealing that *misses* had a more harmful effect on trust than *FAs* (Azevedo-Sa et al., 2020). In contrast, other studies suggest no differences between error types. For instance, Guzman-Bonilla and Patton (2024) investigated participants' responses to automation errors (*FAs* or *misses*) during a block pair matching task. Trust was measured after each block, and their findings indicated that error type did not affect the participants' trust (Guzman-Bonilla and Patton, 2024). After trust violations occur, efforts can be made to restore trust, and these efforts often involve various actions/strategies such as apologies (Mahmood et al., 2022; Natarajan and Gombolay, 2020; Kohn et al., 2019), denials (Kohn et al., 2019), explanations (M. Faas et al., 2021; Ashktorab et al., 2019; Natarajan and Gombolay, 2020), and promises (Albayram et al., 2020).

The second body of research is focused on developing real-time trust prediction models (Xu and Dudek, 2015; Guo and Yang, 2021; Bhat et al., 2022; Chen et al., 2018). Examples of trust prediction models include the online probabilistic trust inference model (OPTIMo) by (Xu and Dudek, 2015), the Beta random variable model by (Guo and Yang, 2021), and the Bayesian model combining Gaussian processes and recurrent neural networks by (Soh et al., 2020). For a detailed review, please refer to (Kok and Soh, 2020).

The third body of research, which is examined in this study, uncovers general properties that govern how a person's trust in automated/autonomous and AI systems changes over time. Recently, Yang et al. (2023a) summarized three key properties of trust dynamics: *continuity* (Trust at the present moment i is significantly associated with trust at the previous moment $i-1$.), *negativity bias* (Negative experiences due to autonomy failures have a greater

influence on trust than positive experiences due to autonomy successes), and *stabilization* (A person's trust will stabilize over repeated interactions with the same autonomy).

The first two properties have been reported consistently in prior research (Lee and Moray, 1992; Yang et al., 2017; Manzey et al., 2012; Yang et al., 2023b). The last property, stabilization, was first empirically found in Yang et al. (2023a). Building on these properties, a predictive computational model was developed, which claims that trust at any point follows a Beta distribution (Guo and Yang, 2021). This model updates trust dynamically based on observed automation performance and surpasses previous models in prediction accuracy while trust generalizability and explanatory power, highlighting the importance of the three properties of trust dynamics.

The studies in this dissertation also examine how trust evolves during moment-to-moment interactions with AI systems, capturing how trust continuously adjusts as humans observe AI performance. Through this approach, the thesis provides further empirical support for the properties of continuity, negativity bias, and stabilization in the context of human-AI collaboration.

1.2.4 Uncertainty Presentation

People frequently encounter decisions under uncertainty, such as whether to bring an umbrella when a weather forecast indicates a 30% chance of rain or whether to refuel immediately when a vehicle's remaining range displays only 10 miles. Prior research has shown that presenting uncertainty estimates to users can improve weather-related decision-making and driving performance (Joslyn and LeClerc, 2013; Jung et al., 2015).

Human operators interacting with automated decision aids face similar challenges, and several human factors approaches have been developed to address them. One approach involves presenting likelihood alarms based on probability (e.g., presenting no alert, "Warning" alert, or "Danger" alert) rather than binary alarms (no alert or alert). This design has been shown to reduce workload and enhance situational awareness by reducing uncertainty and mental strain associated with traditional alarm systems (Fallon et al., 2005). Such likelihood alarms mitigate the negative effects of unreliable automation by helping operators more effectively disregard false alarms. Similarly, Hoff and Bashir (2015) found that transparent automation, systems that provide accurate feedback about reliability or operational characteristics, can better facilitate a more appropriate understanding of the aid's performance. For example, Seong and Bisantz (2008) demonstrated that presenting performance-related information explaining why automation failures occurs led to higher trust ratings in low-reliability automation (users ignored aid when appropriate) and lower trust ratings in high-reliability

automation (users recognized the imperfect system).

As more automation systems utilize artificial intelligence (AI) models, the complexity of AI performance has increased substantially. This complexity introduces a dilemma that even though presenting all possible uncertainty-related information can improve users' understanding of the system performance and support more accurate decision-making, excessive information may increase cognitive load and lead to confusion (Miller, 2019; Fernandes et al., 2018). Accordingly, uncertainty representations have been suggested to be designed to support rapid and accurate decision-making while minimizing ambiguity (Kay et al., 2016). Moreover, future studies have been encouraged to incentivize user decisions by incorporating rewards or consequences that more closely resemble real-world scenarios beyond controlled experimental settings, which has been lacking in prior research on uncertainty visualization (Fernandes et al., 2018).

Building on these design considerations for uncertainty communication, researchers have developed various methods to enhance people's (proper) trust in automation/autonomy (Chatzimparmpas et al., 2020b; Rudin and Ustun, 2018), including the use of various graphical representations (Dasgupta et al., 2016; Yang et al., 2020; Sacha et al., 2016; Fisher et al., 2012; Mercado et al., 2016).

For example, a military perimeter defense experiment conducted by Mercado et al. (2016) aimed to investigate the role of intelligent agent transparency on operator trust. Participants were tasked with selecting optimal routes for unmanned vehicles, assisted by an AI agent. The AI agent operated at three levels of transparency: 1) basic details only, 2) basic details supplemented with reasoning and rationale, and 3) comprehensive information, including basic details, reasoning, rationale, and uncertainty indication, in a text description. They observed a positive correlation between transparency levels and participant trust. They concluded that providing operators with the agent's reasoning process and uncertainty metrics fostered a deeper understanding of the system's capabilities, thereby enhancing trust and increasing usability (Yang et al., 2023a). This finding emphasizes the importance of transparent AI systems in supporting effective human-machine collaboration.

Another study investigated the impact of visual explanations on human trust in machine-learning systems (Yang et al., 2020). Participants performed leaf classifying tasks with or without visual explanations. The leaf examples were presented in two formats: images and feature charts. Results revealed that providing visual explanations enhanced trust and confidence in participants' decision-making. Interestingly, the feature charts were designed with intentional omissions of detailed explanations of features to prevent information overload. However, this simplification led participants to struggle to interpret the charts, and expert users expressed a need for more comprehensive feature descriptions to inform their decisions

(Yang et al., 2020). This insight reveals the importance of integrating visual explanations with a thoughtfully managed information load for appropriate human trust.

1.3 Research Objectives

The literature review summarizes prior work that informs the four sequential aims of this research. Building on this foundation, this dissertation investigates human trust and performance in AI-assisted decision-making with the goal of enhancing overall system safety. The objectives of the dissertation research are as follows:

- **Aim 1.** First, this dissertation evaluates pharmacists' trust in automated pill verification technology designed to support medication dispensing. Dispensing errors are a significant contributor to adverse drug events, leading to substantial healthcare costs and patient harm. In response, AI pill verification systems with high model accuracy have been developed to aid pharmacists during the dispensing process. The implementation of these technologies holds potential for safer AI-assisted medication dispensing and to reduce error and patient harm. However, despite the advances in model accuracy, realizing the full potential of these systems depends on pharmacists establishing appropriate trust in them, as trust in automation is one of the most crucial factors determining the use of automation. Prior research has shown that graphical representations that enhance AI transparency can effectively promote proper trust in automation/autonomy. Accordingly, this work examines how presenting AI uncertainty information affects pharmacists' trust in the system.

- **Aim 2.** Second, this dissertation investigates how different error patterns of AI-aided decision-making in non-binary decision scenarios influence human operators' trust in AI systems and their task performance. In the first aim, Signal Detection Theory (SDT) was used to model autonomy performance in the study of pharmacists' trust in automated pill verification technologies, reflecting the predominant use of SDT in prior research on trust in automation/autonomy. SDT characterizes the decision environments into binary terms and has proven effective for modeling diagnostic decision-making. For example, in the context of pill dispensing, false alarms may lead to minor workflow disruptions, whereas misses could lead to dispensing errors, indicating that a more conservative AI system prioritizing safety by minimizing misses may better support safety.

  However, this binary classification reveals a theoretical limitation. SDT oversimplifies the interaction observed in real-world AI-assisted decision-making contexts, which of-

ten involve numerous possible prediction outcomes. To address this gap, this research abstracts the pill-dispensing scenario into a generalizable experimental testbed that enables systematic investigation of multi-class classification scenarios, which is broadly applicable to AI-assisted decision-making contexts beyond healthcare and allows examination of how distinct error types shape human trust and performance.

- **Aim 3.** Third, this dissertation analyzes how varying consequences of error influence human trust in AI systems. Although AI technologies are widely recognized as valuable tools for their potential to enhance decision-making, the cost of error associated with AI errors can differ substantially in safety-critical domains. Some errors may result in little or no harm, whereas others may carry severe consequences, including serious injury or loss of life. Existing literature reports mixed findings regarding the impact of error consequences on human trust, underscoring the need for further investigation. Building on the second research aim, this work incorporates the multi-class error pattern framework to extend beyond a general understanding of trust in AI, enabling a more nuanced examination of how trust and performance behaviors vary across different types of AI errors and levels of consequence.

- **Aim 4.** Fourth, this dissertation examines how varying stages of automation (SoA) influence human-AI system performance in AI-assisted tasks. Prior research on SoA suggests that although higher stages of automation can improve performance under reliable conditions, system failures at higher stages may lead to more severe negative consequences. By extending the binary SDT accuracy classification (correct or incorrect) into the multi-class error framework, this work identifies distinct and previously underexplored error patterns in which wrong AI predictions can lead to beneficial outcomes, termed "unintended safeguards." Notably, these "unintended safeguards" are found to explicitly differ between Stage 3 and Stage 2 automation. In addition to evidence from a human subject experiment, this work introduces a mathematical proposition to further support the findings, demonstrating that AI systems operating at higher stages (Stage-3) will achieve greater interaction-level accuracy than systems operating at lower stages (Stage-2).

1.4 Dissertation structure

This dissertation is organized into six chapters. Chapter 1 introduces the research motivation, reviews related work, and outlines the aims of this dissertation. Chapter 2 presents a domain-specific study investigating pharmacists' trust in AI technology. Chapter 3 addresses

a key research gap by developing a generalizable experimental testbed that formalizes the beyond binary decision framework. Chapter 4 extends this framework by examining how varying consequences of error influence trust in AI-aided tasks. Chapter 5 explores stages of automation and demonstrates the benefits of higher stages of automation identified through the beyond binary decision framework. Chapter 6 integrates findings across all chapters, discusses the broader impact and intellectual merit of the work, and outlines directions for future research.

CHAPTER 2

The Effects of Presenting AI Uncertainty Information on Pharmacists' Trust in Automated Pill Recognition Technology: an Exploratory Study

2.1 Introduction

Pharmacists play a critical role in ensuring that patients receive medications exactly as prescribed by healthcare providers. A main component of this responsibility is the verification process, in which pharmacists confirm that the medication dispensed in filled bottles corresponds accurately to the prescription. Dispensing errors, cases in which patients receive an incorrect drug or dosage, represent a major source of preventable adverse drug events, accounting for approximately 700,000 emergency room visits and 100,000 hospital admissions each year (Reiner et al., 2020). These errors arise from multiple factors, including technological limitations, the absence of standardized procedures, and the substantial cognitive demands placed on pharmacy personnel who frequently manage numerous concurrent tasks (Reiner et al., 2020; Chui and Mott, 2012; Chui et al., 2014; Balchaites et al., 2022). Improving patient outcomes, lowering avoidable healthcare expenditures, and reducing pharmacists' workload depend on the development and deployment of dependable tools to mitigate dispensing errors (Larios Delgado et al., 2019).

From the 1990s, barcode scanning systems were promoted as a strategy for reducing medication errors (Bates, 2000). Their implementation in pharmacies and broader healthcare settings has been associated with notable declines in medication error rates (Poon et al., 2006, 2010; Paoletti et al., 2007; Young, 2002). However, subsequent research has shown that barcode systems remain vulnerable to workarounds and human error (Debono et al.,

This chapter has been published as a journal article: Kim, J. Y., Marshall, V. D., Rowell, B., Chen, Q., Zheng, Y., Lee, J. D., Kontar, R. A., Lester, C., and Yang, X. J. (2025b). The effects of presenting AI uncertainty information on pharmacists' trust in automated pill recognition technology: Exploratory mixed subjects study. JMIR Human Factors, 12:e60273.

2013; Koppel et al., 2008; Oldland et al., 2015). Furthermore, such systems do not sufficiently resolve the workload pressures faced by pharmacists (Buchanan et al., 1991; Gidman, 2011; Lea et al., 2012; Peterson et al., 1999).

To address these limitations, image-based pill counting and verification/recognition systems have been introduced (Caban et al., 2012; Hamilton, 2004; Popovich Jr and Jordan, 2011; Wootton et al., 2003). Technologies such as Eyecon and VIVID apply vision-based approaches to count pills placed on a tray. Advances in automated pill identification have also leveraged feature engineering techniques. For instance, Yu et al. (2014) and Yu et al. (2015) developed imprint-based automatic pill recognition methods that achieved accuracies of 86.01% and 90.46%, respectively. In a related approach, Caban et al. (2012) employed a modified shape distribution method to extract shape, color, and imprint features for drug identification. When evaluated on 568 of the most frequently prescribed drugs in the United States, their method attained an accuracy of 91.13%.

More recently, deep learning has further advanced automated pill recognition capabilities (Larios Delgado et al., 2019; Lester et al., 2021). For example, Larios Delgado et al. (2019) introduced a two-stage deep learning framework consisting of a deep Convolutional Neural Network for pill blob detection, followed by a classifier that identified the five most probable pills, achieving 94% accuracy. Similarly, Wong et al. (2017) proposed a Deep Convolutional Network model that achieved a mean accuracy of 95.35%. Additionally, Lester et al. (2021) trained a ResNet-18 model to predict labeled features of medication products using images of pills inside filled prescription bottles. On a test set of 65,274 images representing 345 distinct oral drug products, the macro-average precision reached 98.5%.

Despite these substantial improvements in predictive performance, the successful integration of such systems into practice depends on users' ability to develop appropriately calibrated trust in these technologies (Lee and See, 2004).

This chapter aimed to examine pharmacists' trust in an automated pill verification system intended to support the critical task of medication dispensing. In particular, we investigated how the presentation of AI uncertainty information influences pharmacists' trust in the system.

> **H1**: Presenting AI uncertainty information of predicted probability in a visualization format will increase AI transparency, leading to enhanced pharmacists' trust in pill verification technology.

In addition to the primary aim, this chapter also explored how pharmacists' trust-related behaviors varied across different AI performance patterns (within the framework of signal detection theory (SDT)):

H2: Misses would result in a more significant decline in trust compared to false alarms.

H3: Furthermore, given the differing consequences associated with the four SDT patterns, we speculate that presenting AI uncertainty information might have varying effects depending on the specific type of patterns.

2.2 Methods

2.2.1 Ethical Considerations

This study was exempt from IRB oversight by the University of Michigan Institutional Review Board (HUM#00213493). Prior to participation, participants provided electronic informed consent, and all data were collected in an anonymous manner. Participants were compensated with $150 upon completing the study.

2.2.2 Recruitment and Participants

Recruitment emails were sent to pharmacists via the Minnesota Pharmacy Practice-Based Research Network and the University of Michigan College of Pharmacy Pharmacist Preceptor Network. Eligibility criteria required participants to 1) hold a U.S. pharmacist license, 2) be at least 18 years of age, and 3) have access to a laptop or desktop computer with a webcam. Pharmacists were excluded if they (1) required assistive technologies for computer use, (2) wore eyeglasses with more than one power, (3) had uncorrected cataracts, intraocular implants, glaucoma, or permanently dilated pupils, and/or (4) exhibited eye movement or alignment abnormalities (e.g., amblyopia, strabismus, or nystagmus). A total of N = 30 licensed pharmacists in the United States (mean age = 39.40 years, SD = 11.23) completed the study.

2.2.3 AI Model

The AI model employed in this study was a Bayesian neural network (BNN) designed to predict the National Drug Code (NDC), a unique identifier assigned by the United States Food and Drug Administration (FDA) to catalog drug products (Chen et al., 2024; Zheng et al., 2023). The BNN used a random dropout technique (Gal and Ghahramani, 2016) within a ResNet-34 Convolutional Neural Network (CNN) architecture (He et al., 2016) to estimate the class-level probability. Specifically, the model generated 50 predicted probabilities for each input image, with each output represented as a probability vector over all NDC

classes. The final predicted NDC was determined by selecting the class with the highest mean probability across the 50 predictions.

The model was trained on a dataset of 432,974 images obtained from a mail-order pharmacy. This pharmacy employs a robotic system that counts pills, fills and labels prescription bottles, captures images of the bottle contents, and seals the bottles. The dataset consisted one year of robot-acquired images of oral medications (e.g., tablets and capsules) filled in prescription bottles. Each image was labeled with an NDC and associated data, including color, shape, size, manufacturer, tablet scoring, and imprint. The number of images per NDC ranged from 3 to 12,105 (median = 540). The medications were classified into twelve color classes and seven shape categories, as summarized in Table 2.1.

Table 2.1: Percentage of medication characteristics (colors on the left, shapes on the right) featured in the training dataset.

Colors	Percentage (%)	Shapes	Percentage (%)
White	42.1	Round	49.6
Yellow	12.3	Oval	33.4
Pink	9.1	Capsule	16.2
Orange	7.1	Hexagon-6-sided	0.4
Multi-color	5.9	Triangle	0.3
Green	5.2	Trapezoid	0.1
Red	5.2	Pentagon-5-sided	0
Blue	4.8		
Brown	3.8		
Purple	3.1		
Turquoise	0.7		
Gray	0.7		

2.2.4 Experimental Testbed

Participants completed a pill verification task with the assistance of an imperfect AI system that recommended whether to accept or reject a filled medication. Participants were instructed to determine whether the filled medication matched the reference image. If the filled medication did not match the reference image, the correct response was to select "reject." If the two images matched, the correct response was to select "accept."

The user interface was developed based on pharmacists' feedback obtained through a focus group study (Zheng et al., 2023). The interface presented a filled medication image, a reference image, relevant prescription information, and the AI recommendation. Two versions

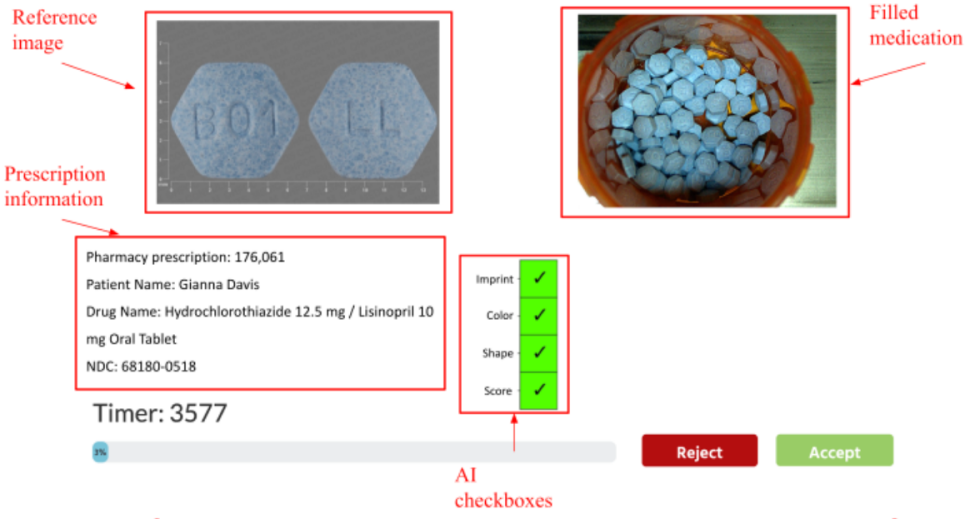
of AI aids were implemented, both powered by the same AI model. One version included an uncertainty visualization indicating the degree of certainty/uncertainty associated with the AI system recommendation, while the other aid did not have this feature. In both conditions, the AI system provided its recommendation using four visual checkboxes. When all four checkboxes were green, the AI recommended "accept" (see Figure 2.1a). Otherwise, the recommendation was "reject." In the uncertainty-augmented condition, an additional histogram was displayed (Figure 2.1b). The histogram illustrated the distribution of fifty predicted probabilities for the selected NDC, derived from fifty iterations of prediction outputs. The visual representation was intended to convey the level of certainty or uncertainty of AI's NDC prediction.

Participants completed a block of 100 pill verification trials with the assistance of an AI aid. The stimuli for these trials, including the reference NDC and corresponding filled medication images, were systematically selected from a dataset of 432,974 images. The selection process ensured broad representation across key visual characteristics, such as color and shape (see Table 2.2), while excluding blurry images to preserve visual clarity. To reduce potential learning or familiarity effects, each reference NDC was presented no more than twice during the experiment.

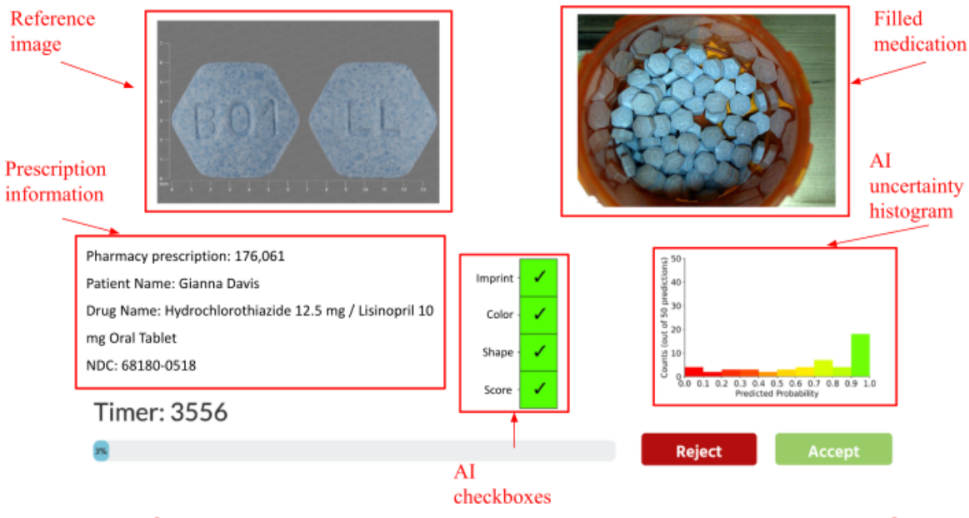
Table 2.2: Percentage of medication characteristics (colors on the left, shapes on the right) featured in trials for reference images and filled images.

Colors	Percentage Reference (%)	Percentage Filled (%)	Shapes	Percentage Reference (%)	Percentage Filled (%)
White	42	43	Round	60	59
Yellow	10	10	Oval	18	19
Pink	10	11	Capsule	19	19
Orange	7	6	Hexagon-6-sided	2	2
Multi-color	4	5	Triangle	0	0
Green	4	4	Trapezoid	1	1
Red	3	3	Pentagon-5-sided	0	0
Blue	7	8			
Brown	9	8			
Purple	2	1			
Turquoise	1	1			
Gray	1	0			

Furthermore, the AI aid was imperfect across the 100 trials and occasionally generated



(a) Interface Without Uncertainty Plot



(b) Interface With Uncertainty Plot

Figure 2.1: Interface for the AI aid without and with the uncertainty plot. Checkboxes indicate AI's recommendation. When all four checkboxes are green, the AI advises to accept; otherwise, to reject. The histogram shows the distribution of the 50 predicted probabilities associated with the predicted NDC.

incorrect recommendations. Using Signal Detection Theory (SDT), we characterized the relationship between the AI's recommendation and the true state of the world (Tanner Jr and Swets, 1954). In this task, the "signal" corresponded to an incorrectly filled medication. The AI's ability to recommend rejecting an incorrect medication reflected its signal detection capability. The interaction between the true state (correct vs. incorrect medication) and the AI's recommendation (accept vs. reject) yielded four possible outcomes: rejecting an incorrect medication (hit), approving an incorrect medication (miss), rejecting a correct medication (false alarm), and approving a correct medication (correct rejection), as illustrated in Figure 2.2.

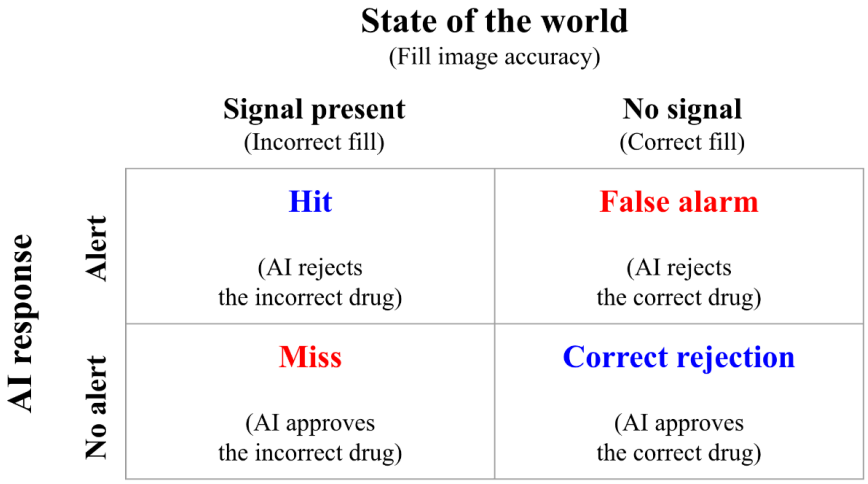


Figure 2.2: Four potential AI performance patterns according to Signal Detection Theory.

Benchmarking prior literature (Reilley et al., 2003), the base rate of incorrectly filled medications was set at 24%, meaning that 24% of trials contained an error. The AI's overall accuracy was set at 82%, ensuring that the system would be perceived as helpful while still producing a sufficient number of misses and false alarms for analysis (Wickens and Dixon, 2007). Given these parameters, the 100 trials consisted of 60 correct rejections (AI approving a correctly filled medication), 22 hits (AI rejecting an incorrectly filled medication), 2 misses (AI approving an incorrectly filled medication), and 16 false alarms (AI rejecting a correctly filled medication). After each trial, participants received feedback indicating whether their accept/reject decision was correct and whether the medication had been correctly filled (e.g., "Your decision was correct. The medication was correctly filled."). Participants then reported their trust in the AI system using a visual analog scale anchored from "Not at all trust" to "Completely trust" (Appendix B.1) (Yang et al., 2023a; Manzey et al., 2012; Yang et al., 2017).

2.2.5 Experimental Design

The study employed a mixed experimental design. The between-subjects factor was AI aid type, defined by the presence or absence of an uncertainty visualization. The within-subjects factor consisted of the four possible AI outcomes: rejecting an incorrect medication, rejecting a correct medication, approving an incorrect medication, and approving a correct medication (see Figure 2.2).

Participants were randomly assigned to one of the two AI conditions, with half using the AI aid without the uncertainty plot and the other half using the uncertainty-augmented version. Each participant completed two blocks of 100 trials. In one block, participants performed the task with the assigned AI aid; in the other block, they completed the task manually without AI assistance. The order of these blocks was counterbalanced to control for order effects. Consistent with prior research (Lee and Moray, 1992), the trial sequence within each 100-trial block was fixed.

Because the primary objective of this study was to examine pharmacists' trust in AI, analyses focused exclusively on the AI-assisted blocks. Data from the manual block was excluded, as they did not involve interaction with the AI system.

2.2.6 Measures

Trust Propensity. Prior to the experiment, participants' general propensity to trust was assessed using the six-item scale developed by Merritt et al. (2013). Trust propensity is defined as a stable, trait-like inclination to trust or distrust others (Merritt and Ilgen, 2008). In the context of automation, propensity to trust machines reflects an individual's general tendency to trust systems, rather than trust directed toward a specific device or application.

End Trust. End trust, $Trust(100)$ is the participant's final trust rating after interacting with the AI help scenario.

Average Trust. Average trust denotes the mean of moment-to-moment trust ratings collected throughout the experiment:

$$\text{Average Trust} = \sum_{i=2}^{100} \text{Trust}(i) \tag{2.1}$$

Trust Adjustment. After each trial i, participants reported their $trust(i)$ in the AI. We calculate a trust adjustment as:

$$\text{Trust adjustment}(i) = \text{Trust}(i) - \text{Trust}(i-1), \quad \text{for } i = 2, 3, \ldots, 100 \tag{2.2}$$

Since the moment-to-moment trust is reported after each trial, only 99 trust adjustments are obtained from each participant.

2.2.7 Experimental procedure

The experiment was conducted remotely with participants who met the inclusion criteria. Interested individuals were first screened via phone to confirm eligibility. Prior to the experimental session, each participant attended a brief virtual meeting with a member of the research team to verify that their physical environment, particularly lighting conditions, was appropriate for the study. Participants were then directed to the Labvanced platform on their personal computer, where they viewed a fifteen-minute instructional video explaining the simulated medication verification task and the testbed interface. They were informed that their objective was to determine whether an image of a filled medication bottle matched a designated reference image. The tutorial also introduced the two versions of the AI aid.

Before beginning the verification task, participants completed a demographics questionnaire and a trust propensity survey (Merritt et al., 2013) (Appendix A.1). They also completed calibration procedures for the eye-tracking software. Participants then performed two blocks of 100 verification trials each: one block with the assigned AI aid (either with or without the uncertainty visualization) and one block manually without AI assistance. The order of the two blocks was counterbalanced across participants. After completing all 200 trials, participants filled out a post-experimental survey and were invited to provide optional open-ended feedback (Muir and Moray, 1996) (Appendix A.2). All participants completed each block within the 60-minute time limit.

Upon finishing the experimental tasks and surveys, participants were invited to a 30-minute debriefing interview session with a member of the research team. During this session, six concepts relevant to automation evaluation—observability, predictability, directing attention, exploring the solution space, adaptability, and calibrated trust—were introduced (McDermott et al., 2018). For each concept, an example scenario was presented, after which participants were asked to reflect on how the concept applied to the AI system used in the study (Appendix A.3).

2.2.8 Statistical Analysis

The analyses focused on participants' trust propensity, final (end) trust, and trust adjustments while interacting with the two types of AI aids. We began with descriptive statistics to characterize participants' baseline trust propensity. To test the directional hypothesis that the uncertainty-augmented AI aid would result in higher end trust and higher average

trust, we conducted one-tailed t-tests.

To examine how trust changed following each of the four AI performance patterns, we fitted mixed-effects linear models with random intercepts. Regression-based t-tests were used to compare mean trust levels between the two AI aid conditions for each performance pattern. Degrees of freedom were estimated using the Kenward-Roger approximation. Mixed-effects generalized linear models were implemented using the lme4 package in R (Bates et al., 2014). Statistical significance was evaluated at the $\alpha = 0.05$ level, and all analyses were conducted using R statistical software, version 4.2.2 (R Core Team, 2019).

2.3 Results

2.3.1 Trust Propensity

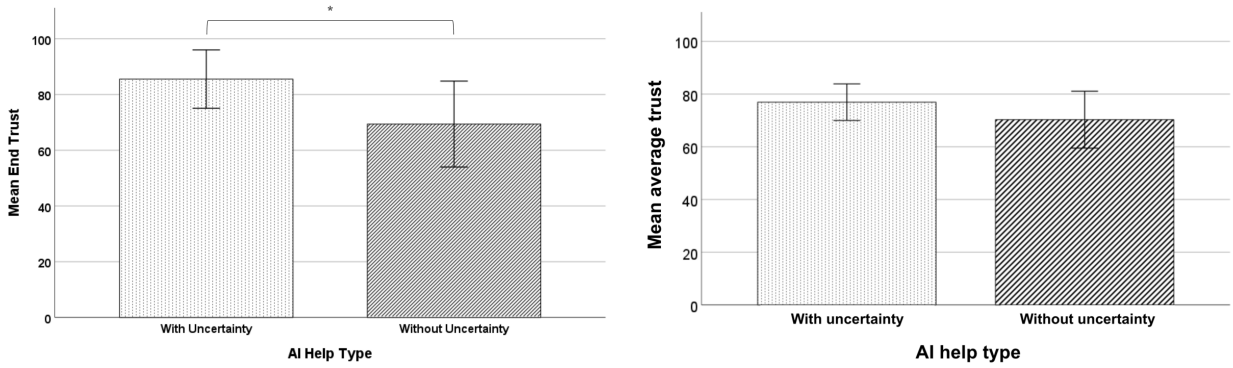
Participants had an average trust propensity score of 72 out of 100 (SD = 18.08), indicating the participants generally had a positive attitude towards trusting automated technologies. There was no significant difference between the two AI aids ($t(28) = -0.854$, $p = .2$, Cohen's $d = .312$).

2.3.2 End-Trust Towards AI Aid

The one-tailed t-test indicated that participants trusted the AI aid with the uncertainty plot significantly more than the AI aid without the plot at the end of the experiment ($t(28) = -1.854$, $p < .05$, Cohen's $d = -.677$) (Figure 2.3a).

2.3.3 Average Trust Towards AI Aid

Participants showed a slightly higher average trust in the AI aid in the with-uncertainty condition (mean=76.92, SD=13.42) than in the without-uncertainty condition (mean=70.29, SD=20.88) (see Figure 2.3b). However, the difference did not reach statistical significance ($t(28) = -1.036$, $p = .155$, Cohen's $d = -.378$).



(a) Mean End trust by AI Help Type

(b) Mean average trust by AI Help Type

Figure 2.3: Mean end (left) and average (right) trust by AI aid help type (with or without the uncertainty plot). The error bars represent a 95% confidence interval.

2.3.4 Trust Adjustment

Figure 2.4 shows the trust adjustment after participants experienced the four AI performance patterns. When AI approved the correct drug, there was a significantly greater trust increment when participants used the AI aid with the uncertainty plot compared to the one without ($t(78.98) = 3.93$, $p < .001$, Cohen's $d = .214$); When AI approved the incorrect drug, there was a significantly greater trust decrement using the AI aid with the uncertainty plot ($t(2939.72) = -4.78$, $p < .001$, Cohen's $d = .712$). Interestingly, when AI rejected the incorrect drug, we observed a decrement of trust for participants using the AI aid with the uncertainty plot ($t(509.77) = -3.96$, $p < .001$, Cohen's $d = .312$). When AI rejected the correct drug, both AI help types showed a decrease in trust, and there was no statistical difference between them ($t(856.57) = -0.68$, $p = .494$, Cohen's $d = .045$). Overall, participants using the AI aid with the uncertainty plot displayed a large magnitude of trust adjustment. In addition, we observed a significant "negativity bias" in that the magnitude of trust adjustment when AI made an error (i.e., AI approves the incorrect drug or AI rejects the correct drug) was significantly larger than the magnitude of trust adjustment when AI provided correct recommendations (generalized linear model test, $z = -11.30$, $p < .001$).

To examine variation in pharmacists' trust behavior across different AI performance patterns categorized by signal detection theory (SDT), we initially combined the data from both the with and without uncertainty help scenarios. Trust decreased significantly more when AI approved the incorrect drug compared to when AI rejected the correct drug ($t(509) = -4.687$, $p < .001$, Cohen's $d = .475$). This trend was observed in the with uncertainty AI help scenario ($t(254) = -4.91$, $p < .001$, Cohen's $d = .758$). However, in the without uncertainty AI

help scenario, although there was a greater decrease in trust when AI approved the incorrect drug than when AI rejected the correct drug, the difference was not statistically significant ($t(254) = -1.14$, $p = .255$, Cohen's $d = .014$).

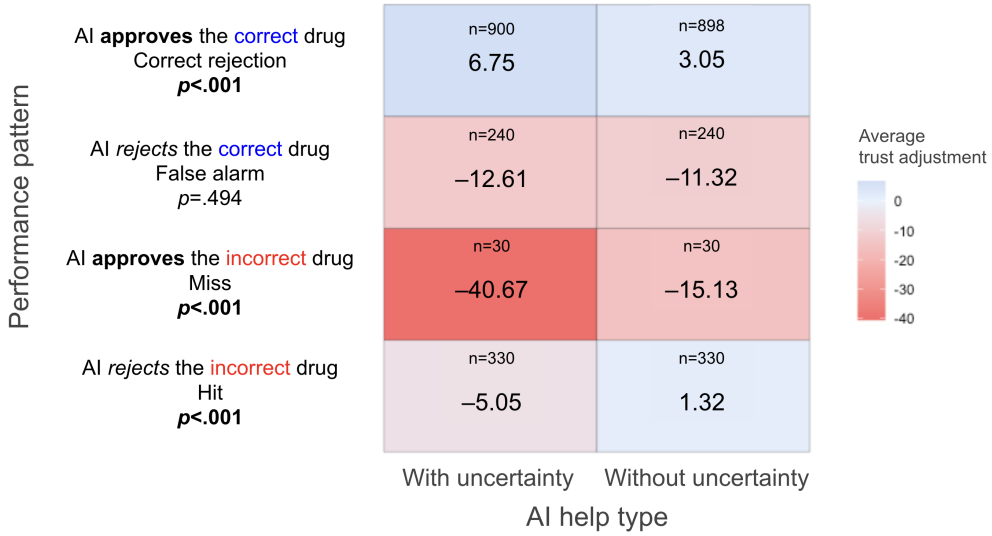


Figure 2.4: Trust adjustment for different AI performance patterns. Red shades represent negative trust adjustment, and blue shades represent positive trust adjustment.

As we measured participants' trust towards AI continuously, we calculated the autocorrelation between the trust ratings. Autocorrelation measures the relationship between a variable's current value and its past values in a time series. Figure 2.5 shows the mean autocorrelation as a function of time separation between the ratings. For both AI aids, the correlation decreased as the time separation increased. The AI aid with the uncertainty plot had a lower mean autocorrelation compared to the aid without the uncertainty plot.

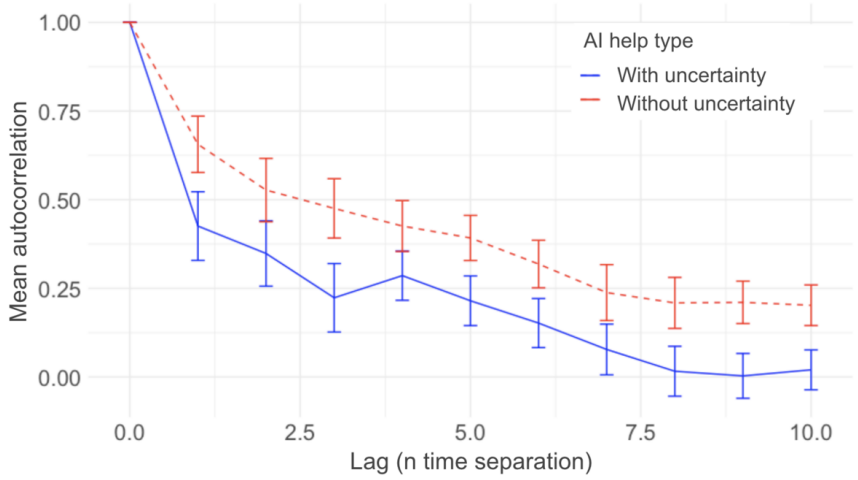


Figure 2.5: Autocorrelation of trust as a function of time separation. The blue solid line represents AI aid with the uncertainty plot, and the red dashed line represents the AI aid without the uncertainty plot. The error bars represent +/- 2 standard errors.

2.4 Discussion

2.4.1 Principal Findings

This study examined pharmacists' trust in automated pill verification technology and investigated how presenting AI uncertainty information influences trust. Overall, the results indicate that pharmacists demonstrate a generally favorable disposition toward trusting automation. Moreover, incorporating uncertainty information into the AI interface significantly increased pharmacists' end-of-task trust in the system (see Figure 2.3a).

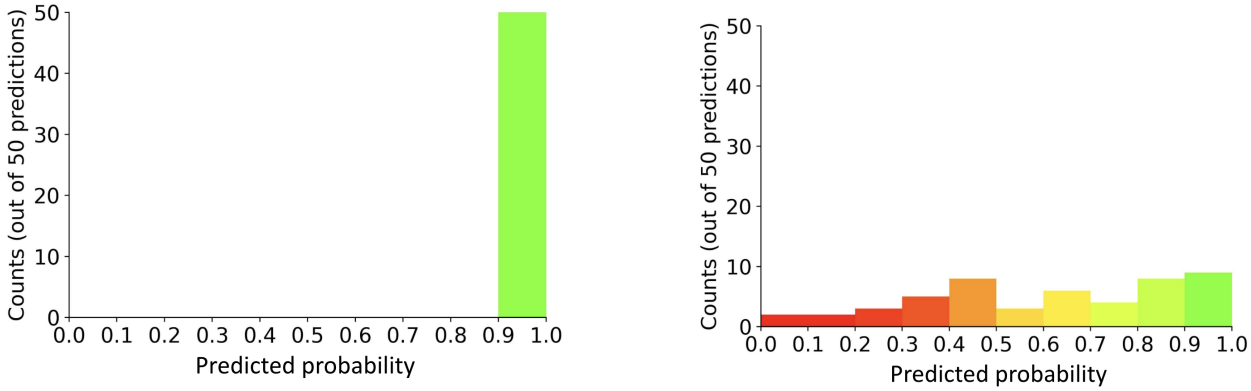
2.4.2 Comparison to Prior Work

Propensity to trust automation refers to an individual's general tendency to rely on automated or autonomous systems, shaped by prior experiences and expectations about future performance (Merritt et al., 2013; Merritt and Ilgen, 2008). Prior research has documented variability in trust propensity across populations. For example, Merritt et al. (2013) reported a mean trust propensity of 3.56 (SD = 0.6) on a 7-point scale among 69 college students (average age = 25). More recent work by Montag et al. (2023), surveying 289 adults aged between 18 and 70, reported a mean of 4.98 (SD = 1.06) on a 7-point Likert scale. Similarly, Miller et al. (2021) found a mean score of 4.97 (SD = 1.21) among 28 participants aged 18-60. In another study, Yang et al. (2022) reported average scores equivalent to 5.08, 4.86, and 4.86 (on a 7-point scale) across three adult groups.

Consistent with these findings (Montag et al., 2023; Miller et al., 2021; Yang et al., 2022; De Visser et al., 2018), pharmacists in our study exhibited a relatively high baseline propensity to trust automation, averaging 72 on a 100-point scale (equivalent to 5.03 on a 7-point Likert scale). This favorable disposition may reflect their routine interaction with automated technologies such as barcode scanners and pill counters in daily practice (Larios Delgado et al., 2019; Bates, 2000). As expected, no significant differences in trust propensity were observed between the two AI aid conditions due to random assignment.

With respect to end trust, pharmacists using the AI aid with the uncertainty visualization reported significantly higher trust scores. This enhancement likely reflects increased transparency achieved through the histogram displaying the distribution of 50 predicted probabilities. Although AI technologies offer substantial performance benefits, they are often perceived as "black boxes," which can undermine trust and team effectiveness (De Visser et al., 2018; Lyons et al., 2017; Luo et al., 2022). The higher end-trust scores observed in the uncertainty condition suggest that revealing aspects of the AI's internal decision-making process can foster stronger trust in automation. Participant feedback supports this interpretation. For example, one pharmacist noted that visible uncertainty prompted greater vigilance ("P14: As soon as the uncertainty plot became red and yellow, I slowed down, which could be helpful because sometimes slowing down when there is uncertainty, just knowing there's uncertainty, is enough."), while another reported increased confidence when the plot indicated high certainty ("P24: if the uncertainty plot was all green bar and AI thought it was doing a 100% accurate job, then it was easier to make my decision.").

Examining moment-to-moment trust dynamics revealed distinct patterns. When the AI correctly approved a filled medication (correct rejection), trust increased in both conditions, with larger increments observed when uncertainty information was present. In contrast, when the AI approved the incorrect medication (miss), trust declined significantly in both conditions, with a larger decrement when uncertainty information was displayed. This stronger decline may have resulted from the histogram conveying high confidence in an erroneous recommendation (illustrated in Figure 2.6a), leading participants to perceive the error as a "confident" failure. Prior research on likelihood alarms suggests that highly confident automation errors produce larger immediate declines in trust (Yang et al., 2017; Sorkin et al., 1988).



(a) Uncertainty plot with narrower IQR

(b) Uncertainty plot with wider IQR

Figure 2.6: Uncertainty plot with narrower (left) and wider (right) interquartile range (IQR). IQR is a measure of statistical dispersion, or how spread out the data points are.

Consistent with prior SDT-based trust research (Davenport and Bustamante, 2010; Stanton et al., 2009), misses (approving incorrect medication) produced larger trust decrements than false alarms (rejecting correct medication). However, this difference reached statistical significance only in the uncertainty condition, further suggesting that visible confidence intensifies the impact of automation errors.

This ambiguity might unintentionally cause the human participants to doubt the capability of the AI aid with the uncertainty plot, resulting in a decrease in their trust, as evidenced by P18's statement: "When the checkboxes were not all green, and the histogram had a bunch of colors (variability), I was even less trusting of the AI tool." This perspective is also supported by P25, who noted: "When the uncertainty plot had some red, it made me double check and decreased my confidence for sure."

Interestingly, when the AI correctly rejected an incorrectly filled bottle, trust increased in the condition without uncertainty information, but decreased when the uncertainty visualization was present. This divergence may reflect the influence of the uncertainty plots on participants' interpretations of the AI's recommendation. When the AI correctly approved a medication, the accompanying uncertainty plots consistently displayed a solid green bar, indicating high confidence (Figure 2.6a). In contrast, when the AI rejected an incorrect medication, the uncertainty distributions were often more dispersed, with a wider interquartile range (Figure 2.7). Specifically, 16 of the 22 such cases showed broader spreads with mixed color bars (Figure 2.6b), signaling greater variability in the model's predictions. This visual ambiguity may have unintentionally reduced participants' confidence in the AI's capability, even when its recommendation was correct. Participant feedback supports this interpretation. For example, P18 noted that "When the checkboxes were not all green, and

the histogram had a bunch of colors (variability), I was even less trusting of the AI tool," while P25 reported "When the uncertainty plot had some red, it made me double check and decreased my confidence for sure."

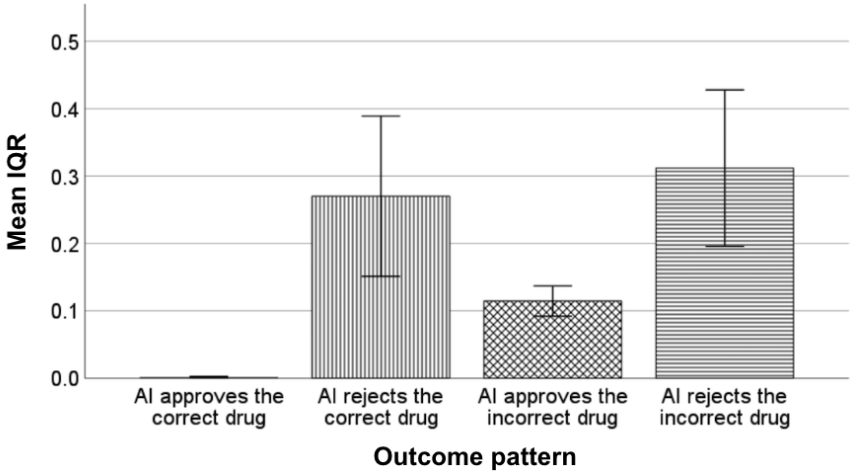


Figure 2.7: Mean interquartile range (IQR) by outcome pattern. The error bars represent a 95% confidence interval.

When the AI rejected a correctly filled medication (false alarm), trust decreased in both conditions, with no significant difference between AI aid types. In real-world practice, such rejections would likely prompt pharmacists to re-inspect prescriptions, potentially increasing workload and fatigue, which may negatively affect performance over time (Carayon and Gurses, 2008; Watterson et al., 2021). Nevertheless, some participants perceived false alarms as acceptable or even desirable, viewing the AI as a precautionary safeguard rather than a final decision-maker, highlighting: "I didn't lose as much trust when AI rejected the correct drug. I feel like AI should be there as a cautionary tool."

The observed trends in trust adjustment patterns further indicated greater variability when uncertainty information was present. Autocorrelation analyses (Figure 2.5) showed lower temporal dependence between successive trust ratings in the uncertainty condition, suggesting that pharmacists updated their trust more dynamically based on current trial information rather than relying on prior impressions. This pattern highlights a potential advantage of transparent AI displays that encourage real-time calibration rather than static reliance.

Across both AI conditions, negative automation outcomes had a stronger impact on trust than positive outcomes, consistent with established negativity bias findings in automation research (Manzey et al., 2012; Yang et al., 2022).

2.4.3 Limitations and Future Directions

We recognize several limitations of this study and outline directions for future research. First, as an early investigation in this area, we did not systematically manipulate or control the interquartile range of the uncertainty visualization (Figure 2.7). Future studies should experimentally vary the distributional characteristics presented in the uncertainty plot to examine how different probability spreads influence pharmacists' trust. Manipulating interquartile ranges across distinct outcome patterns would enable a clearer understanding of how performance patterns independently shape trust, without potential confounding effects from visualization variability.

Second, the uncertainty visualization presented only the top one of the fifty predicted probability distributions for the selected NDC. Future research could enrich the display by incorporating additional contextual or comparative information to improve interpretability. A key challenge identified in this study was that participants with limited familiarity with statistical graphics found the histogram difficult to interpret. As one participant suggested simplifying the uncertainty presentation because "it would be a little bit too much for some people not as comfortable with statistics or technology". However, other feedback suggested that familiarity and exposure increased comfort and trust over time (P30: "I started skeptical because it's something I'm not familiar with, but as I got more examples of it, my trust built up quickly."). These observations indicate that both improved visualization design and structured training may enhance usability. Accordingly, future work should explore alternative visualization strategies that communicate AI uncertainty in a more intuitive and accessible manner while maintaining sufficient informational depth for expert users.

Third, this exploratory study was constrained by a relatively small sample size, which may have reduced statistical power for detecting effects related to trust propensity, end trust, and average trust. This limitation likely contributed to the absence of statistical significance in some comparisons. However, a significant difference in end trust was detected despite the limited sample, suggesting a substantial effect. Additionally, to preserve the perceived usefulness of the AI while ensuring the presence of both misses and false alarms, only two miss trials and sixteen false alarm trials were included, resulting in an overall AI accuracy of 82%. These relatively small cell counts may have further limited statistical sensitivity. Future investigations should recruit larger samples and consider expanded trial distributions to improve robustness and generalizability.

Fourth, the present study focused exclusively on trust and trust adjustment, without incorporating analyses of objective performance measures such as accuracy or response time. Although examining trust in isolation is consistent with prior automation research (Tanner Jr

and Swets, 1954), a more integrated analysis linking trust to behavioral performance would provide deeper insight into trust calibration and its consequences for decision quality.

2.5 Conclusion

Dispensing errors remain a major contributor to adverse drug events, leading to substantial healthcare costs and patient harm. Despite advances in automated pill verification technologies, pharmacists' trust in these systems has not been comprehensively examined. This research represents an initial step toward understanding how pharmacists interact with and evaluate automated verification tools, contributing to the broader integration of AI systems into healthcare practice.

Our findings indicate that pharmacists generally exhibit a positive baseline disposition toward trusting automation, likely reflecting their routine interaction with automated tools in daily workflows. Moreover, presenting uncertainty information alongside AI recommendations significantly enhanced end-of-task trust, underscoring the value of transparency in AI system design. Trust dynamics varied as a function of AI performance. When the AI correctly approved a medication, the uncertainty-augmented interface produced larger increases in trust. Conversely, when the AI incorrectly approved a medication, the presence of uncertainty information resulted in a larger decline in trust. Notably, when the AI correctly rejected an incorrect medication, trust increased in the absence of uncertainty information but decreased when the uncertainty plot was displayed. Finally, a pronounced negativity bias emerged: decreases in trust following AI errors were greater than increases in trust following correct AI decisions.

CHAPTER 3

Beyond Binary Decisions: Evaluating the Effects of AI Error Type on Trust and Performance in AI-Assisted Tasks

3.1 Introduction

AI systems are increasingly integrated into safety-critical decision-making processes (Cheikh et al., 2022; Azevedo-Sa et al., 2020; Robert et al., 2020; Bhat et al., 2024b). The safe implementation of these technologies depends on effective human-AI interaction, and trust in automation has been recognized as a key determinant of reliance, compliance, and overall human-automation performance (Lee and See, 2004; Hoff and Bashir, 2015). A majority of prior research relied on Signal Detection Theory (SDT) to model automation performance and evaluate human interaction outcomes (Sorkin and Woods, 1985; Du et al., 2020; Neyedli et al., 2011; Chancey et al., 2017). While SDT has proven effective in modeling traditional diagnostic tasks, its binary structure presents theoretical limitations in the context of modern AI systems.

Specifically, the traditional binary SDT framework collapses various system error types into a single "incorrect" category. In contrast, contemporary AI models, particularly those based on machine learning, often operate in multi-class classification environments, where multiple possible classification states are considered simultaneously. For example, in the hypothetical scenario presented in Chapter 1 (section 1.1), both Cases D and E would be categorized as "incorrect" classification under the SDT framework. However, these cases differ in their implications. Case E involves an AI error with potentially catastrophic consequences, as the system incorrectly confirms a wrongly filled medication as correct (Table 3.1, last row). On the other hand, although Case D also involves an incorrect classification, this

This chapter has been published as a journal article: Kim, J. Y., Lester, C., & Yang, X. J. (2025). Beyond Binary Decisions: Evaluating the Effects of AI Error Type on Trust and Performance in AI-Assisted Tasks. Human Factors, 00187208251326795.

Table 3.1: Description of the five cases and two outcomes. Bold text indicates potential incorrect outcomes resulting from automation errors.

Case	Pattern	Prescribed Medication	Filled Medication	AI Prediction	Final Outcome
A	RC	X	X	X	R: Dispense medication W: Error-correction refill
B	RI	X	X	Y/Z	R: Dispense medication W: **Error-correction refill**
C	WC	X	Y	Y	R: Error-correction refill W: Dispense medication
D	WI	X	Y	Z	R: Error-correction refill W: **Dispense medication**
E	WI_{ref}	X	Y	X	R: Error-correction refill W: **Dispense medication**

error may be beneficial from a utilitarian perspective, as the incorrect classification differs from the prescribed medication and triggers a safety alert (Table 3.1, second-to-last row).

For clarity, the five cases outlined in the hypothetical scenario are represented using two-letter notation (see Figure 1.1). The first letter, R or W, indicates whether the participant's initial decision was right or wrong, respectively. The second letter, C or I or I_{ref}, denotes the AI's prediction, conditional on the participants' initial performance.

The two types of incorrect predictions, I and I_{ref}, are differentiated to capture meaningful distinctions within multi-class error patterns. Specifically, the "ref" subscript signifies that the AI falsely confirms the incorrectly filled medication as the prescribed (i.e., reference) medication. In addition, the final task outcome can likewise be categorized as R or W, depending on whether the final decision is right or wrong. Altogether, this framework produces ten distinct outcomes, derived from the combination of five patterns (cases) and two possible final outcomes for each pattern (Table 3.1).

This chapter has several objectives. Its primary aim is to investigate how AI predictions influence users following an initial wrong decision, with particular emphasis on three patterns: WC, WI, and WI_{ref}.

Both WI and WI_{ref} describe situations in which the wrong medication is dispensed and subsequently misclassified by the AI system. The critical difference lies in the outcome of the AI's misclassification. In the WI pattern, the AI incorrectly classifies medication Y as Z, prompting re-examination of the dispensed medication and potentially preventing harm. By contrast, in the WI_{ref} pattern, the AI misidentifies medication Y as the prescribed medication X, thereby falsely reinforcing the operator's initial error and providing false reassurance. Based on this distinction, we propose the following hypotheses:

H1: Pattern $\boldsymbol{WI_{ref}}$ will lead to the worst performance, followed by pattern $\boldsymbol{WI}$,

then pattern $\boldsymbol{WC}$.

With respect to trust adjustment, we hypothesize that both WI and WI_{ref} will produce trust decrements relative to the trust increment expected under pattern WC. Furthermore, we anticipate that the decline in trust will be significantly greater for pattern WI_{ref} than for pattern WI.

> **H2**: Pattern $\boldsymbol{WI_{ref}}$ will have the lowest trust decrement, followed by pattern $\boldsymbol{WI}$, then pattern $\boldsymbol{WC}$. We expect this trend to be observed regardless of the final outcome of the task (Note: a larger trust increment is considered a smaller trust decrement).

Regarding the response time to reach a final decision after receiving the AI's prediction, we hypothesize an interaction between the three patterns and the final outcome. Because pattern WI_{ref} misleads the operator by confirming the incorrect initial decision, participants who are deceived under this pattern are expected to respond relatively quickly. In contrast, we anticipate the longest response times under pattern WI, as participants must detect a discrepancy between the AI's incorrect prediction, the reference image, and their initial response, and then engage in error correction before reaching a final decision.

> **H3a**: It would take shorter times for participants in pattern $\boldsymbol{WI_{ref}}$ to commit errors, compared to patterns $\boldsymbol{WI}$ and $\boldsymbol{WC}$.
>
> **H3b**: It would take longer times for participants in pattern $\boldsymbol{WI_{ref}}$ to recognize errors, compared to patterns $\boldsymbol{WC}$ and $\boldsymbol{WI}$.

Beyond these primary hypotheses, this chapter also seeks to further investigate three theoretical properties of trust dynamics: continuity (**H4**), negativity bias (**H5**), and stabilization (**H6**).

> **H4**: Trust at the current moment i is significantly associated with trust at the prior moment $i-1$.
>
> **H5**: The magnitude of trust decrement following incorrect predictions ($\boldsymbol{RI}$, $\boldsymbol{WI}$, $\boldsymbol{WI_{ref}}$) will be greater than the magnitude of trust increment due to correct predictions ($\boldsymbol{RC}$, $\boldsymbol{WC}$).
>
> **H6**: The magnitude of trust adjustment will decrease as a function of time for both the correct and the incorrect AI predictions.

3.2 Methods

3.2.1 Ethical Considerations

This research adhered to the ethical standards of the American Psychological Association and was exempt from Institutional Review Board oversight at the University of Michigan (*HUM00230326*). Informed consent was obtained from all participants prior to participation.

3.2.2 Recruitment and Participants

Thirty-five undergraduate and graduate students from the University of Michigan (mean age = 22.43 years, SD = 3.27) participated in the study. All participants had normal or corrected-to-normal vision. Compensation consisted of a $10 base payment, with an additional performance-based bonus of up to $10.

Participants completed a three-dimensional spatial visualization and recognition task (Vandenberg and Kuse, 1978; Kim et al., 2024, 2023) with AI assistance. In each trial, a reference image was presented from a set of 30 images. Participants were then shown five answer options and asked to select the option that matched the reference. After making their initial choice, the simulated AI generated a prediction regarding the participant-selected image. Participants subsequently indicated whether their initial selection was correct or not.

This task structure parallels the medication dispensing workflow. In pharmacy practice, a pharmacist first receives a prescription (e.g., vitamin C) and retrieves a corresponding medication from a large inventory. The selected medication is then verified, either by a technician or an AI system, by comparing it against the prescription. Based on this verification, the pharmacist decides whether to dispense the medication or to correct the error by retrieving the appropriate item.

3.2.3 Experimental apparatus and stimuli

3.2.3.1 Stimuli development

The study employed a simulated Mental Rotation Task (MRT), with stimuli developed following the methodology of Shepard and Metzler (1971). Thirty three-dimensional (3D) objects were constructed, half composed of 10 cubes and half composed of 8 cubes. For each object, a two-dimensional reference image was generated at a fixed orientation (see Figure 3.1). Each 3D object was then rotated in 45-degree increments along both the horizontal and vertical axes, producing 14 additional rotated images per object. These rotated versions

served as correct alternatives to the corresponding reference image. In total, the stimulus set consisted of 30 reference images, each paired with 14 valid rotated alternatives.

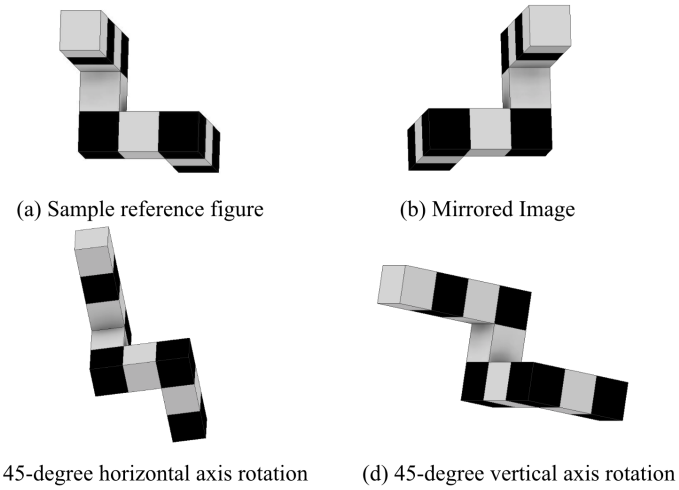


Figure 3.1: Illustration of stimuli development. After creating the (a) reference images, one (b) mirrored image, fourteen 45-degree (c) horizontal axis rotated figures, and fourteen 45-degree (d) vertical axis rotated figures were created for each reference image.

The difficulty of the stimuli was determined through a separate study using Qualtrics, and the participants were recruited via Amazon Mechanical Turk (MTurk). Following Shepard and Metzler (1971)'s mental rotation study, participants viewed multiple image pairs. Each left image was a randomly chosen reference image, and the right image was either a correct alternative, a distractor, or the reference image itself (Figure 3.2). Distractors were randomly chosen from the remaining 29 reference images, which included 14 images with the same cube amount as the reference image (i.e., 8 or 10) and 15 images with a different cube amount. Participants had 10 seconds to identify whether the image pairs were the same or different. Data from participants who failed to match the validation pairs (reference-reference image pairs) correctly were discarded. The mean hit rate determined stimuli difficulty.

Stimulus difficulty was determined through a separate pretest administered via Qualtrics, with participants recruited from Amazon Mechanical Turk (MTurk). Consistent with Shepard and Metzler (1971), participants viewed pairs of images in which the left image was a randomly selected reference and the right image was either a correct alternative, a distractor, or the identical reference image (Figure 3.2). Distractors were selected from the remaining 29 reference images, including 14 with the same number of cubes (8 or 10) and 15 with a different cube amount. Participants were given 10 seconds to judge whether each image pair depicted the same or different objects. Data from participants who failed validation trials (reference-reference pairs) were excluded. Stimulus difficulty was determined using mean hit

rates.

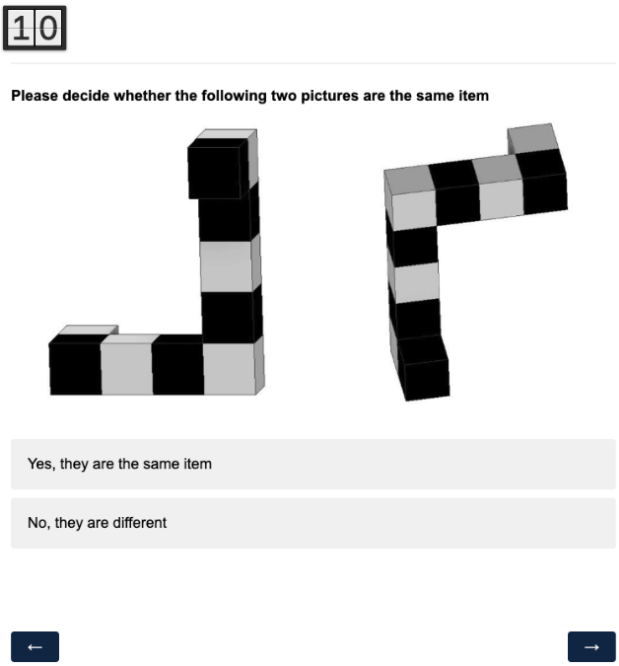


Figure 3.2: For the beta test, participants were asked to determine whether the image pairs are the same items in different rotations (by clicking "Yes, they are the same item") or different items (by clicking "No, they are different") within 10 seconds.

3.2.3.2 Experimental Task

The experimental task was adapted from the three-dimensional spatial visualization test developed by Vandenberg and Kuse (1978) and implemented using the Python Tkinter package. An overview of the task flow is presented in Figure 3.3.

Prior to beginning the experiment, participants completed a demographics questionnaire (age and gender) and a trust propensity survey assessing their general tendency to trust automation (Merritt et al., 2013).

During the experiment, participants completed 60 MRT trials presented in random order. Each trial consisted of multiple stages. First, participants viewed a reference image alongside five alternatives (Figure 3.4a). One alternative was a correct rotated version of the reference image. Two distractors were randomly selected mirrored rotations of the reference object, and the remaining two were rotated images derived from other reference objects. Participants were instructed to identify the option representing the same 3D object as the reference image as accurately as possible within 15 seconds. The time limit was established through pilot testing. If no response was made within 15 seconds, a message indicated that no selection had been recorded (Figure 3.4b), and participants proceeded to the next trial.

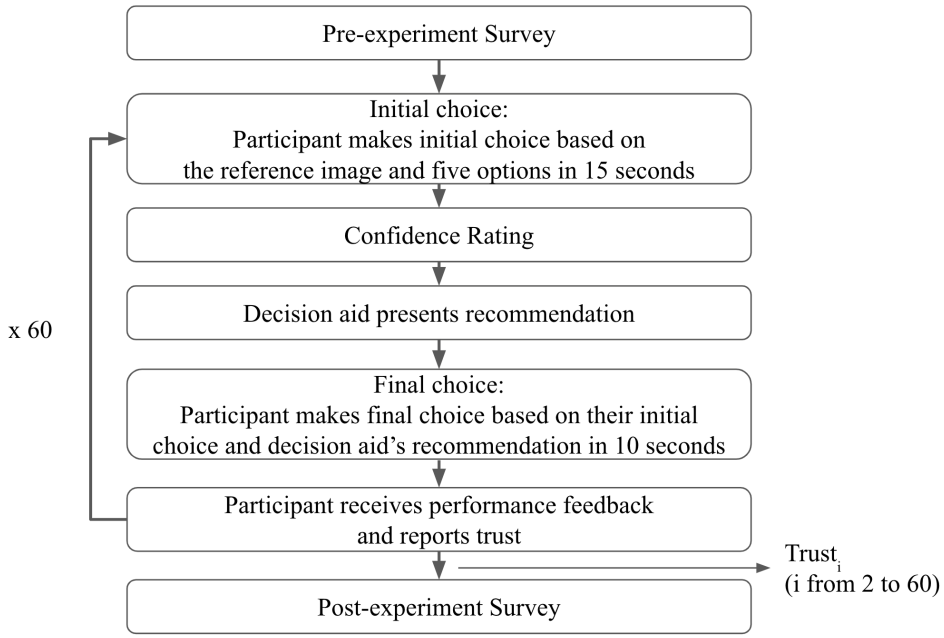
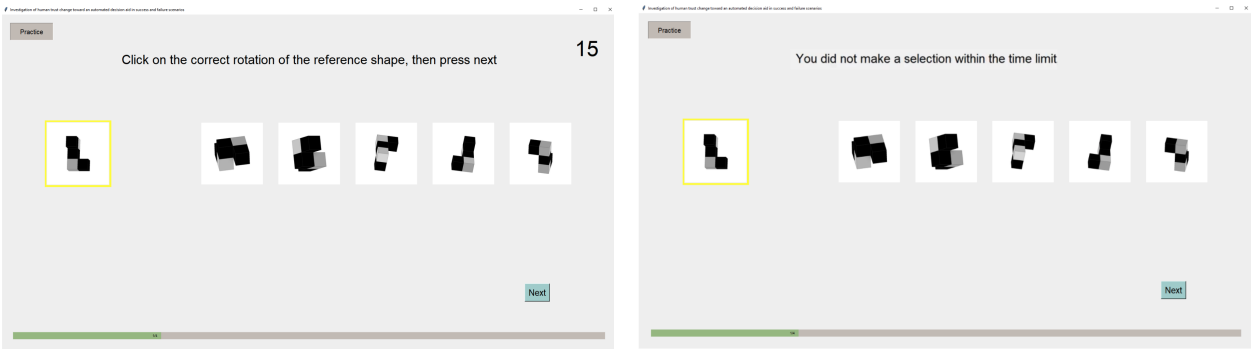


Figure 3.3: Flowchart of the experiment



(a) Initial Answer Choice

(b) No Selection Initial Answer Choice

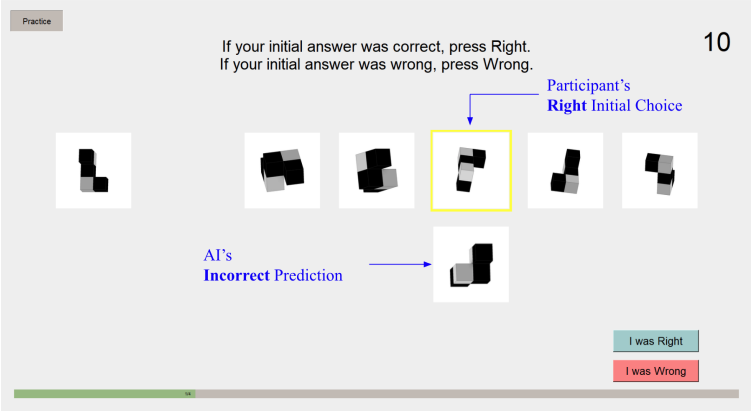
Figure 3.4: Participants were asked to make their initial answer choice within 15 seconds.

After making their initial selection, participants rated their confidence using a visual analog scale ranging from "Not confident at all" to "Absolutely confident" (Figure B.2).

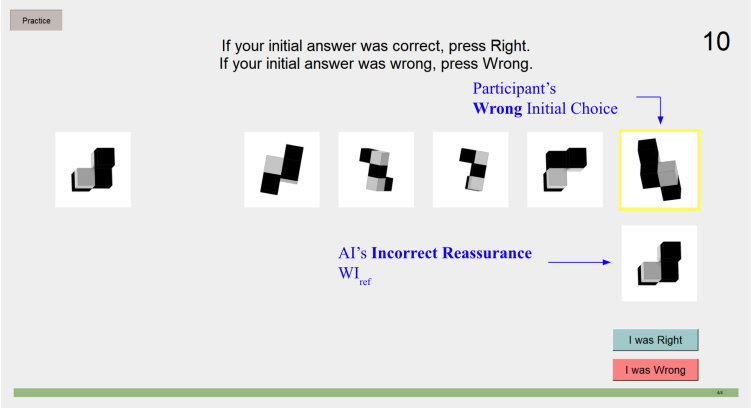
Participants were then presented with the AI's prediction, displayed beneath their original selection to illustrate the AI's interpretation of the participant-selected shape. For example, in Figure 3.5a, the participant selected the third option, but the AI incorrectly identified that selection and displayed a different reference image (corresponding to Case B described in the Introduction). In Figure 3.5b, the participant selected the fifth option, and the AI misclassified it as the correct alternative (corresponding to Case E described in the

Introduction).

After reviewing the AI's prediction, participants had 10 seconds to decide whether to maintain their initial response ("I was right") or revise it ("I was wrong"). This response window was also determined through pilot testing. If no selection was made within the time limit, a notification similar to that shown in Figure 3.4b appeared, and participants proceeded to the subsequent trial.



(a) Right initial choice, incorrect AI prediction (Case B, Pattern RI).



(b) Wrong initial choice, incorrect AI prediction (Case E, Pattern WI_Ref).

Figure 3.5: Participants were presented with the AI system's recognition and chose between sticking with or rejecting their initial choice within 10 seconds. If no selection was made, they were skipped to the next trial.

Next, participants were shown feedback on their performance and the validity of AI's prediction (Figure 3.6).



(a) Correct AI prediction

(b) Wrong AI prediction

Figure 3.6: Participants were presented with the performance feedback page, showing their performance on the initial and final answer choices and the validity of the AI's prediction.

Participants rated their trust in and perceived reliability of the AI system on a visual analog scale (Appendix B.3). Following Yang et al. (2023b)'s approach, participants were asked to rate their trust after each trial. The trust scale ranged from the leftmost anchor labeled "I don't trust the decision aid at all" to the rightmost anchor "I absolutely trust the decision aid." The perceived reliability rating was on a 0 to 100 scale.

3.2.4 Experimental Design

The experiment employed a within-subjects design in which the independent variable was performance pattern (Table 3.1). The presence of each pattern and corresponding outcome depended on participants' responses, as neither initial nor final decisions could be experimentally manipulated.

To generate the various performance patterns, the AI system was programmed with predetermined recognition probabilities. Specifically, the AI correctly identified the selected shape 70% of the time, regardless of whether the participant's initial response was correct. When a correct AI recognition followed a correct initial response, pattern Pattern *RC* (Case A) occurred. When a correct AI recognition followed an incorrect initial response, pattern *WC* (Case C) occurred.

The AI misclassified the participant's selected shape as a non-reference alternative 20% of the time. This produced pattern *RI* (Case B) if the participant's initial response was correct and pattern *WI* (Case D) if the initial response was incorrect. In addition, 10% of the time, the AI identified the participant's selected shape as the reference image. This resulted in pattern pattern *RC* (Case A) when the initial response was correct and pattern WI_{ref} (Case E) when the initial response was incorrect. Consequently, the effective reliability of the AI

ranged between 70% and 80%, depending on participants' initial decisions. This level exceeds the commonly cited baseline threshold of approximately 70% required for automation to be perceived as beneficial (Wickens and Dixon, 2007). Participants were informed that the AI's recommendations were imperfect, though no explicit reliability rate was disclosed.

3.2.5 Measures

The primary dependent measures were as follows.

Confidence. After completing the initial selection in each trial, participants rated their confidence using a Visual Analog Scale ranging from 0 to 100. The left anchor was labeled "I am not confident in my answer at all," and the rightmost anchor was labeled "I absolutely trust the decision aid."

Trust Adjustment. After each trial i, participants reported their $trust(i)$ in the decision aid. Trust adjustment was calculated as:

$$\text{Trust adjustment}(i) = \text{Trust}(i) - \text{Trust}(i-1), \quad \text{for } i = 2, 3, \ldots, 60 \tag{3.1}$$

Since the moment-to-moment trust is reported after each trial, 59 trust adjustments are collected from each participant.

Performance. Final task performance served as another dependent variable and was calculated as the percentage of correct final decisions within each pattern for each participant (Figure 3.5a).

Reaction Time. Reaction time was measured in seconds from the onset of the AI's prediction display until participants selected either "I was right" or "I was wrong" (Figure 3.5a).

3.2.6 Experimental procedure

After providing informed consent and completing online demographic and trust propensity questionnaires (Merritt et al., 2013) (Appendix A.1), participants watched an instructional video describing the experimental procedure. They were reminded that the AI system was imperfect and that its predictions might be either accurate or inaccurate. Each participant completed 60 trials. Following the task, participants filled out a post-experiment survey (Merritt et al., 2013) (Appendix A.2).

3.2.7 Statistical Analysis

To test the hypotheses, linear mixed-effects models were constructed using the lme4 package in R (Bates et al., 2014), with participant included as a random effect. Mixed-effects modeling was chosen due to its suitability for repeated-measures data and its robustness to missing observations. Model development followed an iterative, hierarchical approach (Field et al., 2012), beginning with simpler models and progressively adding predictors. Likelihood ratio tests (LRTs) were used to compare nested models, and more parsimonious models were retained when additional complexity did not significantly improve fit. Statistical significance was evaluated at α =0.05.

The mixed-effects model equation is:

$$y_{ij} = (\beta_0 + u_{0j}) + \beta_1 \cdot \text{CaseB}_{ij} + \beta_2 \cdot \text{CaseC}_{ij} + \beta_3 \cdot \text{CaseD}_{ij} + \beta_4 \cdot \text{CaseE}_{ij} + \epsilon_{ij}$$

Where:

y_{ij}**:** The outcome variable for participant j under case pattern i.

β_0**:** The fixed effect intercept.

u_{0j}**:** The random intercept for the jth participant.

$CaseB_{ij}, CaseC_{ij}, CaseD_{ij}, CaseE_{ij}$**:** Dummy variables for the case patterns, comparing each case to the reference case (i.e., Case A).

$\beta_1, \beta_2, \beta_3, \beta_4$**:** The fixed effect coefficients for the case patterns (B, C, D, and E) relative to the reference case A.

ϵ_{ij}**:** The residual error for participant j in case pattern i.

Post hoc comparisons were conducted using t-tests with degrees of freedom estimated via the Kenward-Roger method (Kenward and Roger, 1997). This approach adjusts the variance-covariance matrix of fixed effects using a Taylor series approximation and is widely recommended in mixed-effects modeling to reduce Type I error rates and produce more accurate p-values, particularly across varying sample sizes (McNeish and Stapleton, 2016; Luke, 2017; Kuznetsova et al., 2017).

3.3 Results

Upon completion of all trials, the frequency of each performance pattern and corresponding outcome was calculated. Because pattern occurrence was determined retrospectively based on participants' responses, not all participants necessarily exhibited every pattern. Table 3.2

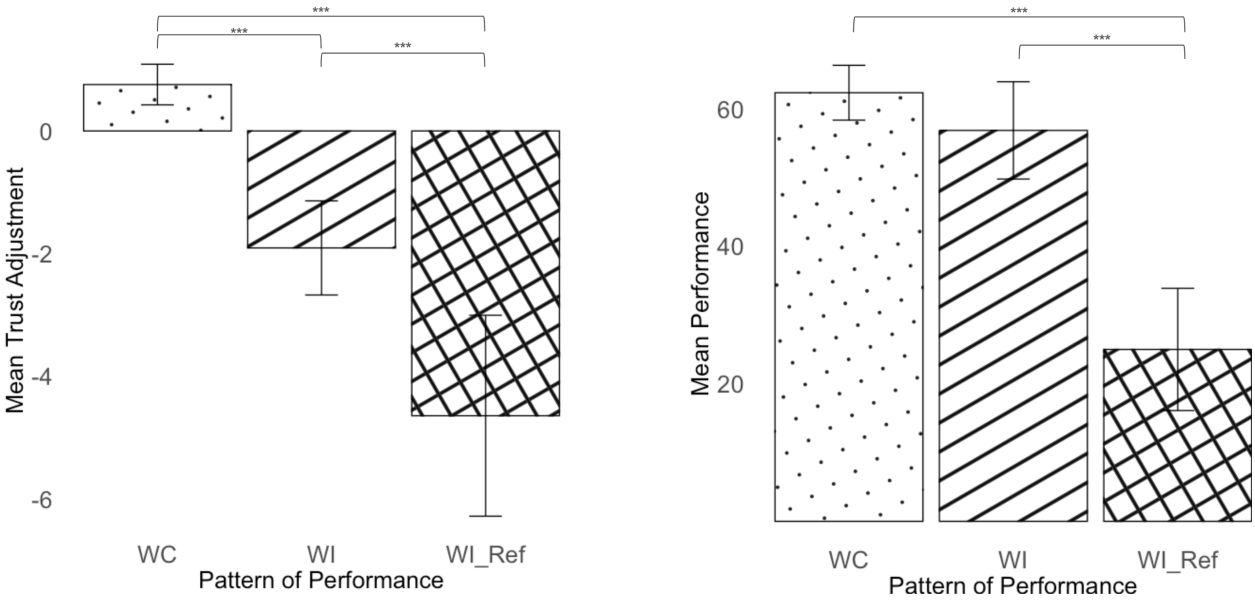
presents the frequency distribution of performance patterns along with descriptive statistics for all measured variables.

Confidence. Participants' confidence ratings were analyzed as an indicator of data validity. We anticipated that confidence would be higher when participants' initial responses were correct. Accordingly, performance patterns were grouped into two categories: right human initial choice (RC, RI) and wrong human initial choice (WC, WI, WI_{ref}). The confidence rating revealed significant differences between the right ($Mean = 72.74, SD = 1.98$) and the wrong ($Mean = 56.77, SD = 2.00$) initial choices ($t(1962.72) = -20.20, p < .001$).

3.3.1 Comparing Patterns WC, WI, and WI_{ref}

Statistical significance indicators are displayed in the figures to denote statistical significance between performance patterns.

Trust adjustment. A significant effect of performance pattern on trust adjustment was observed ($F(2, 875.68) = 61.2, p < .001$). Pattern WC produced a significantly greater increase in trust compared to both pattern WI ($t(875) = 6.57, p < .001$) and pattern WI_{ref} ($t(878) = 9.96, p < .001$). Pattern WI_{ref} resulted in a significantly larger decrease in trust than pattern WI ($t(876) = -4.46, p < .001$) (Figure 3.7a).



(a) Trust Adjustment (b) Performance

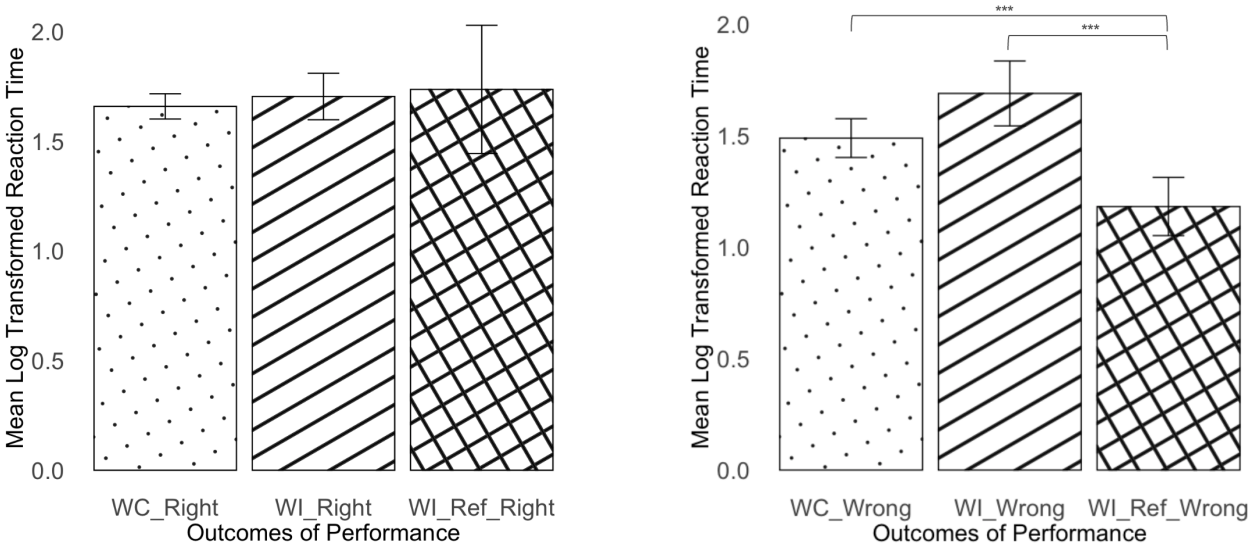
Figure 3.7: Trust adjustment (left) and Performance (right) by patterns of performance. The error bars represent +/- 2 standard errors.

Table 3.2: Mean and sd of confidence rating, trust adjustment, final performance, final reaction time, number of patterns, and number of participants grouped by patterns and outcomes of performance. (*Note: Performance is based on the final answer choice. Therefore, for each pattern, there is only one output.*)

Patterns of Performance		**Confidence Rating Mean (SD)**		**Trust Adjustment Mean (SD)**		**Final Performance (%) Mean (SD)**	**Final Reaction Time (s) Mean (SD)**		**N Patterns (n)**		**N Participants (n)**	
RC	R	74.58 (23.28)	75.48 (22.73)	1.17 (4.33)	1.21 (4.36)	96.96 (17.17)	3.64 (2.00)	3.54 (1.88)	889	862	35	35
	W		45.81 (22.84)		-0.07 (3.02)			6.70 (3.00)		27		13
RI	R	72.53 (24.53)	78.45 (21.57)	-3.67 (6.06)	-3.65 (5.89)	78.61 (41.11)	5.78 (2.47)	5.82 (2.53)	201	158	35	33
	W		50.77 (22.58)		-3.77 (6.71)			5.63 (2.26)		43		23
WC	R	50.41 (22.11)	44.93 (21.75)	0.76 (4.03)	1.01 (3.39)	62.33 (48.50)	5.71 (2.71)	5.96 (2.62)	592	369	35	35
	W		59.46 (19.63)		0.34 (4.90)			5.30 (2.83)		223		33
WI	R	51.68 (23.32)	43.38 (21.53)	-1.91 (5.39)	-2.23 (5.32)	56.85 (49.65)	6.34 (2.88)	6.25 (2.71)	197	112	35	32
	W		62.62 (21.07)		-1.48 (5.47)			6.45 (3.10)		85		30
WI_Ref	R	53.71 (26.16)	42.29 (23.23)	-4.64 (8.03)	-4.46 (7.82)	25.00 (43.53)	4.56 (2.86)	6.79 (3.15)	96	24	34	15
	W		57.51 (26.12)		-4.71 (8.15)			3.82 (2.35)		72		31

Performance. Performance patterns also significantly influenced final task accuracy ($F(2, 858.11) = 29.43, p < .001$). Post-hoc analyses indicated that pattern WI_{ref} had a significantly worse performance compared to WC ($t(856.62) = -7.67, p < .001$) and to WI ($t(856.47) = -5.73, p < .001$) (Figure 3.7b).

Reaction Time. Reaction time data were log-transformed prior to analysis. Both performance patterns ($F(2, 852.35) = 26.82, p < .001$) and outcomes ($F(1, 862.59) = 24.64, p < .001$) significantly affected reaction time. A significant interaction between pattern and outcome was also observed ($F(2, 855.15) = 10.38, p < .001$). Given this interaction, simple effects analyses were conducted within each outcome condition. When the final outcome was right (Figure 3.8a), there was no significant difference between the three patterns ($F(2, 475.49) = .35, p = .71$); When the final outcome was wrong (Figure 3.8b), patterns significantly affected reaction time ($F(2, 363.47) = 22.09, p < .001$). Post-hoc comparisons revealed WI_{ref} resulted in significantly quicker reaction time than WC ($t(362.36) = 5.82, p < .001$) and WI ($t(358.65) = 6.14, p < .001$).



(a) Right Outcomes

(b) Wrong Outcomes

Figure 3.8: Reaction Time by patterns of performance. The error bars represent +/- 2 standard errors.

3.3.2 Evaluating Continuity, Negativity Bias, and Stabilization

To further examine trust dynamics, performance patterns were grouped into correct AI predictions (i.e., patterns RC and WC) and incorrect AI predictions (i.e., patterns RI, WI, and WI_{ref}).

Continuity. Comparing correct AI predictions to incorrect AI predictions shows that AI success led to an increase in trust. Conversely, AI prediction failures led to a decline in trust. In addition, trust at the present moment t is highly correlated with trust at the previous moment $t-1$. Autocorrelation analyses (Figure 3.9) demonstrated that trust ratings were highly correlated between adjacent trials, with the magnitude of autocorrelation gradually diminishing as the time lag increased.

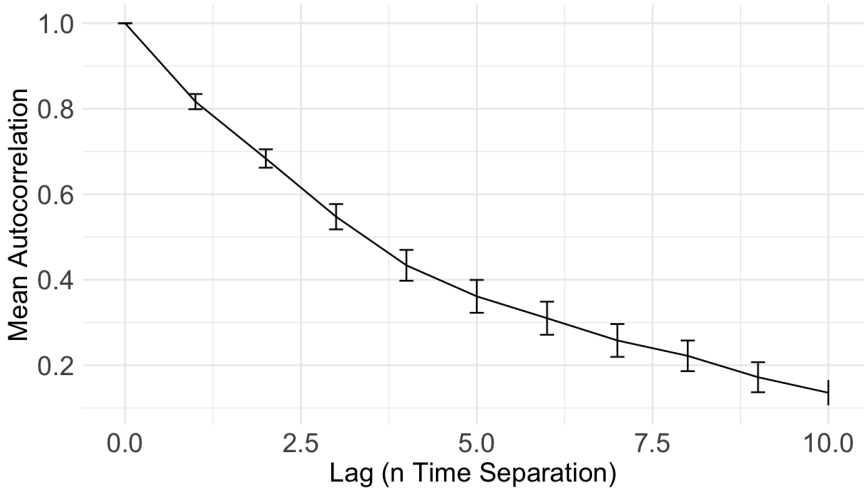


Figure 3.9: Autocorrelation of trust as a function of time separation. The error bars represent +/- 2 standard errors.

Negativity bias. To assess negativity bias, we compared the magnitude of trust increases following correct AI predictions with the magnitude of trust decreases following incorrect predictions. The results revealed a significant asymmetry: the mean trust increment after correct predictions ($Mean = 1.01, SD = 0.13$) was substantially smaller than the mean trust decrement after incorrect predictions ($Mean = -3.16, SD = 0.22$) ($t(1968.00) = -16.58, p < .001$).

Stabilization. Finally, to evaluate the stabilization property of trust, a linear mixed-effects model was estimated with trust adjustment as a function of trial progression. The results indicated a significant reduction in the magnitude of trust adjustments over time for both correct AI predictions ($\beta = -0.035,\ t(1448) = -4.58,\ p < .001$) and the incorrect AI predictions ($\beta = -0.203,\ t(469.27) = -3.50,\ p < .001$), suggesting that trust became increasingly stable as the experiment progressed.

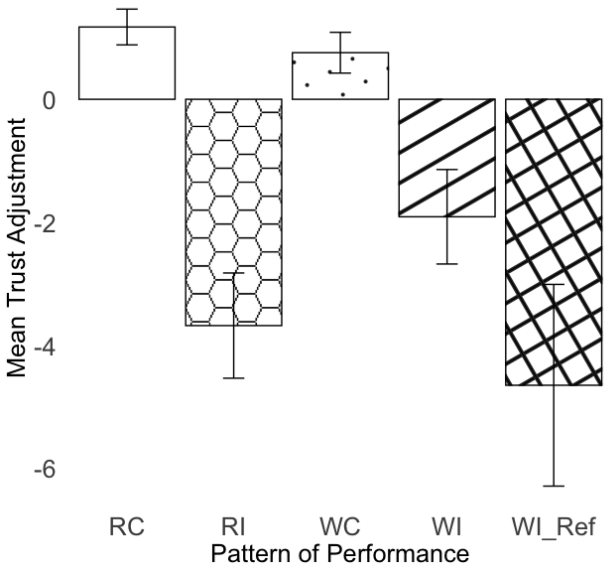


Figure 3.10: Trust adjustment by patterns of performance. The error bars represent +/- 2 standard errors.

3.4 Discussion

In this section, we interpret the findings in relation to the chapter objectives and hypotheses.

The primary focus of this study was the comparison among patterns WC, WI, and WI_{ref}. The results partially support **H1**, which predicted that performance would be highest under WC, followed by pattern WI, and lowest under pattern WI_{ref}. Consistent with this prediction, performance under WI_{ref} was significantly worse than under both patterns WC and WI. Out of the 96 occurrences of WI_{ref}, 72 resulted in incorrect final decisions, underscoring the substantial risk associated with AI systems that falsely confirm an initially incorrect human judgment. When automation provides erroneous reassurance, it can strongly reinforce human error. However, contrary to our expectations, performance did not significantly differ between WC and WI. This finding suggests that certain AI errors, specifically those that prompt re-evaluation, may yield comparable utilitarian benefits to correct AI detections. This counterintuitive result emerges only when moving beyond the binary framework of signal detection theory (SDT) to consider multi-class error structures.

Prior research on human-automation/autonomy interaction highlights risk as a key determinant of trust, reliance, and performance (Hoff and Bashir, 2015). Risk is typically defined as the product of event probability and consequence magnitude (Sheridan, 2008; Slovic and Lichtenstein, 1968). Mathematically, Sheridan (2008) expressed risk as the "product of the probability P_E of an initial reference event E and the sum of products of probability $P_i|E$ of

each consequence occurring given that E occurs, as well as the cost of each consequence C_i." Therefore, $R = P_E \times \sum_i (P_i|E \times C_i)$, where $Consequence_E = \sum_i (P_i|E \times C_i)$. Applying the equation to the present study, P_E is the probability of each case (i.e., P_{case}), and $P_i|E$ can be considered the probability of final outcome (i.e., $P_{success}|case$ and $P_{failure}|case$).

In our experiment, participants were encouraged to perform the task as accurately as possible. Because error costs were not manipulated in this experiment, the values of $C_{success}$ and $C_{failure}$ can be treated as constants. Empirically, the findings related to H1 can be summarized as $P_{success}|\,WC \approx P_{success}|\,WI > P_{success}|\,WI_{ref}$. Furthermore, assuming $C_{success} = 1$ and $C_{failure} = -1$, the average consequences are positives for WC and WI ($Consequence_{WC} = 0.24$, $Consequence_{WC} = 0.14$), but negative for $Consequence_{WI_{ref}} = -0.5$, further illustrating the particularly detrimental nature of false reassurance.

H2 is fully supported. Pattern *WC* produced the largest trust increment, followed by *WI*, with WI_{ref} producing the greatest trust decrement. Consistent with prior research (Lee and Moray, 1992; Yang et al., 2017; Manzey et al., 2012; Yang et al., 2023b), automation success increased trust, whereas automation failure reduced trust. At the same time, our findings revealed two partially competing theories of how people interpret automation performance. On the one hand, if automation performance is assessed based prediction accuracy, we would expect to see a trust increment for pattern *WC*, and a trust decrement for both patterns *WI* and WI_{ref}. Additionally, the amount of trust decrements should be similar for both cases because the automation made a wrong prediction. Alternatively, if users evaluate automation based on the "utility" of its prediction, then we would expect to see decrease in trust for pattern WI_{ref} and increases in trust for both patterns *WC* and *WI*, as both patterns prompt a re-evaluation of the initial selection and may prevent error. The amount of trust increment between the two patterns should also be similar because the final performance of the two patterns was comparable.

Our results suggest that neither perspective alone is sufficient. Instead, the two forces – the AI's performance and the "utility" of an AI prediction – together determine trust adjustment. In binary SDT contexts, the automation's prediction performance and its "utility" are aligned. However, in multi-class settings, the prediction performance and the "utility" of *WI* are in opposite directions, causing the intriguing finding. Specifically, the direction of trust adjustment, increment or decrement, is largely determined by the AI performance. Consistent with prior literature (Lee and Moray, 1992; Yang et al., 2017; Manzey et al., 2012; Yang et al., 2023b), a wrong prediction leads to a trust decrement. However, the magnitude of trust decrement is further influenced by the "utility": a wrong prediction leading to good outcomes has a smaller trust decrement.

Regarding reaction time, we hypothesized an interaction between performance patterns

and final outcomes. The results supported **H3a**: when participants were misled by erroneous AI reassurance (WI_{ref}) they responded significantly faster than under WC and WI. However, **H3b** was not supported. When participants successfully detected the AI's false reassurance, response times did not significantly differ across patterns. This lack of significance could be due to the small sample size, and future research is needed to examine this further. We also observed evidence consistent with a speed-accuracy trade-off: faster decisions were associated with higher error rates. In applied settings, the relative prioritization of speed versus accuracy depends on the payoff structure. For example, in high-stakes medical contexts where errors may have catastrophic consequences, accuracy would likely outweigh time efficiency.

Beyond the primary hypotheses, we examined three properties of trust dynamics: continuity (**H4**), negativity bias (**H5**), and stabilization (**H6**). Continuity and negativity bias have been consistently documented (Lee and Moray, 1992; Manzey et al., 2012; Yang et al., 2016). The stabilization property was recently reported in the more recent studies of Yang et al. (2023a) and Guo and Yang (2021). Our study provide further validation of this property: as participants repeatedly interacted with the AI system, the magnitude of trust adjustments diminished over time, indicating increasing stability. It is important to note that these properties are derived from aggregated data and therefore reflect average behavioral tendencies at the population level. Individual trajectories may vary, but collectively the results reinforce a dynamic and adaptive view of trust in human-AI interaction.

3.5 Conclusion

Most prior research on trust in and reliance on automation and AI has relied on signal detection theory (SDT) to model system performance. Although SDT provides a useful framework, it reduces the state of the world to a binary distinction: signal present versus signal absent. Such simplification does not capture the complexity of many real-world AI applications, which increasingly involve multi-class classification. Modern AI systems routinely differentiate among multiple categories rather than two alternatives. Examples include medical image diagnosis (e.g., classifying chest X-rays into normal, pneumonia, or lung cancer), object recognition in autonomous driving (e.g., recognizing various objects encountered on the road), and product categorization in e-commerce platforms (e.g., classifying products as electronics, clothing, home appliances, or books) (Rahman et al., 2008; Gao et al., 2018; Kozareva, 2015). To our knowledge, this study is among the first to examine trust and reliance in human-AI interaction beyond the binary SDT framework.

Our findings demonstrate systematic differences in performance, response time, and trust

across the three patterns: WC, WI, and WI_{ref}.

These results underscore the importance of distinguishing among error types in automated systems. In particular, when the AI falsely reassured participants by confirming an incorrect initial judgment (WI_{ref}), operators were more likely to respond quickly and ultimately commit errors. This pattern exposes a critical design risk: AI systems that provide misleading confirmation can substantially compromise both safety and effectiveness. By contrast, certain incorrect AI predictions that trigger re-evaluation, even though objectively wrong, may function as safeguards by prompting additional verification. In this sense, not all AI errors are equally harmful; some may paradoxically enhance system-level reliability. Moreover, the study provides additional support for three core properties of trust dynamics, continuity, negativity bias, and stabilization, demonstrating how trust evolves incrementally through moment-to-moment interactions with AI systems.

Several limitations should be acknowledged. First, the study did not manipulate the cost of success or failure (i.e., C_i in the risk equation), which often varies in safety-critical domains. For example, Aldhwaihi et al. (2016) found that although most dispensing errors caused no harm, some errors had high consequences that could potentially cause harm or even death, depending on the dosage and patients' vulnerability. Future research should incorporate varying cost structures to better approximate real-world risk conditions and to examine how varying consequences moderate responses to different AI error types. Future research should also be conducted for ecological validity, particularly in safety-critical settings. Second, although the study was conceptually motivated by medical dispensing scenarios, the experimental paradigm relied on a mental rotation and recognition task. Future work could extend this paradigm to other cognitive domains, such as memory, judgment, or complex decision-making, to test the robustness of the findings across different forms of human information processing.

Third, the relatively small number of WI_{ref} cases (n=96) may have limited statistical power, particularly with respect to the non-significant findings related to H3b. Future studies should consider replication with larger samples.

CHAPTER 4

Evaluating the Effects of Consequences of Error on Trust in AI-Aided Task: An Eye Tracking Study

4.1 Introduction

The cost of error frequently differ across safety-critical settings. For instance, dispensing errors arise when patients receive an incorrect medication or dosage. Although most dispensing errors do not lead to harm, some carry severe consequences, potentially resulting in injury or death depending on the dosage and the patient's level of vulnerability (Aldhwaihi et al., 2016).

Perceived risk, commonly conceptualized as the combination of event likelihood and consequence severity, has been identified as a key external factor influencing trust in automation (Hoff and Bashir, 2015). Because uncertainty is inherent in most operational environments, and particularly pronounced in safety-critical domains, risk may substantially shape trust and reliance behaviors. Similarly, Satterfield et al. (2017) highlighted the importance of risk, though its effects on trust in automation remain insufficiently examined and empirically mixed. For example, Ezer et al. (2008) found that participants relied less on automated aids in counting tasks when errors carried higher consequences. In contrast, Lyons and Stokes (2012) reported that in a route-planning task, participants reduced reliance on human assistance under high-consequence conditions and instead exhibited greater trust in automation.

Building on these findings, the present study examined how varying error consequences influence trust in AI across different error patterns. This design enabled a more nuanced investigation of how trust adapts not only to automation performance but also to consequence severity.

This chapter has been published as a peer-reviewed conference proceeding: Kim, J. Y., Kautz, C., & Yang, X. J. (2025a). Evaluating the Effects of Consequences of Error on Trust in AI-Aided Task: An Eye Tracking Study. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting, volume 69, pages 207-213. SAGE Publications Sage CA: Los Angeles, CA

4.2 Methods

4.2.1 Ethical Considerations

The study adhered to the American Psychological Association's ethical guidelines and received approval from the University of Michigan Institutional Review Board (HUM00249405).

4.2.2 Recruitment and Participants

Thirty-five undergraduate and graduate students from the University of Michigan (mean age = 21.57, SD = 2.33) participated in the experiment. All participants were at least 18 years old, reported no color vision deficiencies, and had normal or corrected-to-normal vision. Because the study incorporated eye-tracking measures, glasses were not permitted; however, contact lenses without color tints were allowed. Participants received a $10 base payment and could earn up to an additional $10 based on task performance.

4.2.3 Experimental Testbed and Stimuli

The experimental testbed was displayed on a 23.8-inch Dell monitor with a resolution of 1920 × 1080 pixels.

4.2.3.1 Mental Rotation Task

Participants completed 60 trials of a mental rotation task (MRT) supported by an AI aid (Kim et al., 2025a). In each trial, they first viewed a reference object and selected one of five alternatives within 15 seconds (Figure 4.1).

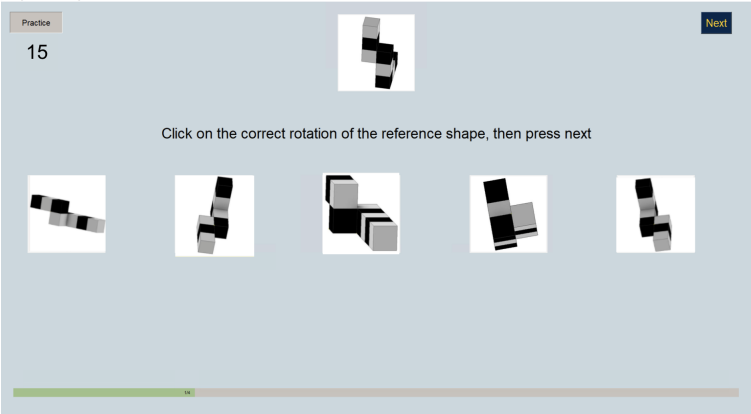
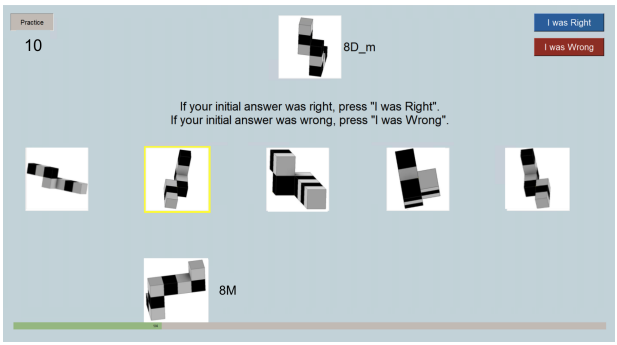


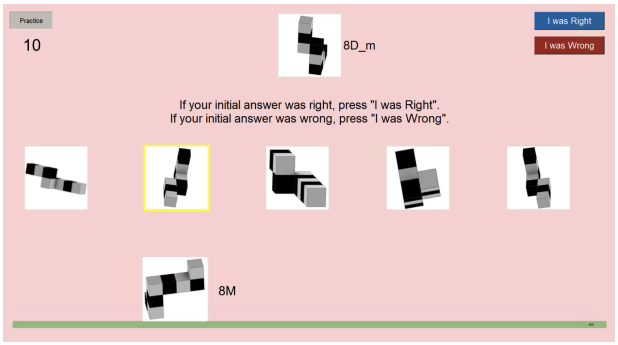
Figure 4.1: Participants were presented with the reference shape and five answer selections, and were instructed to make an initial choice. The low-consequence trials appeared on blue backgrounds.

Next, during a 10-second verification phase, participants reviewed the AI's suggestion and decided whether their initial choice was correct. The AI displayed its prediction of the participant's selected option beneath the initial answer (labeled "8M" in Figure 4.2a).

The experiment included 30 low-consequence trials (blue background; Figure 4.2a) and 30 high-consequence trials (red background; Figure 4.2b), presented in randomized order. Distinct background colors were used to provide a clear and salient visual cue, allowing participants to easily distinguish between consequence levels associated with each trial.



(a) Low Consequence RI



(b) High Consequence RI

Figure 4.2: Pattern *RI* (Right-Incorrect) in low- and high-consequence appeared on blue and red backgrounds, respectively.

After each trial, participants received feedback regarding both their own performance and the AI's prediction accuracy (Figure 4.3). In high-consequence trials, incorrect decisions resulted in double the point loss compared to low-consequence trials. No points were awarded for correct responses in either condition.

Stimulus difficulty across consequence levels was validated using data from 70 participants in prior studies (Kim et al., 2023, 2024). A paired-samples t-test confirmed no significant difference between low-consequence stimuli ($Mean(sd) = .559(.178)$) and high-consequence stimuli ($Mean(sd) = .549(.168)$) ($t(69) = 0.738, p = .46$).

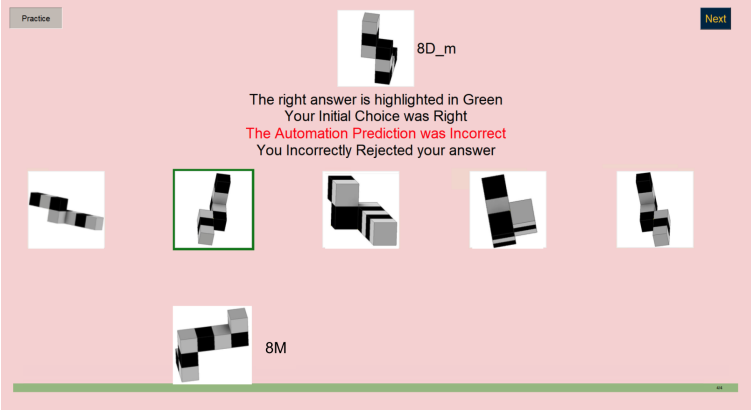


Figure 4.3: Performance feedback of high-consequence pattern RI. In high-consequence scenarios, initial selection, final selection, and performance feedback appeared on red backgrounds.

4.2.3.2 Eye tracking

Eye tracking data were collected using the Tobii Pro Glasses 2 and the Tobii Pro Lab software at a sampling rate of 50 Hz. Three areas of interest (AOIs) were defined (Figure 4.4): the reference AOI (reference stimulus and label), the answer-choice AOI (five response options), and the AI-prediction AOI (AI image and label).

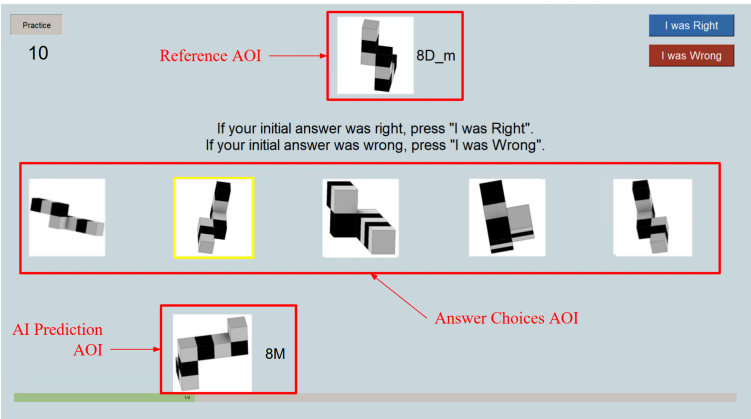


Figure 4.4: Eye tracking areas of interests (AOIs).

4.2.4 Experimental Design

In both consequence conditions, participants interacted with an imperfect AI aid whose reliability ranged between 70% and 80%, consistent with prior findings suggesting that automation below approximately 70% reliability may impose greater costs than benefits (Wickens and Dixon, 2007). The AI aid aided participants in assessing the accuracy of their initial answer selections by providing its prediction of the initial selection (presented below the participants' initial selection in Figure 4.2a, with its label "8M"). However, participants were responsible for making the correct final decision on whether their initial answers were right or wrong.

Each consequence condition level included 30 trials: 21 with correct AI predictions, 6 with incorrect predictions, and 3 in which AI validity depended on the participant's initial response. In these latter trials, a correct initial response resulted in a correct AI prediction (pattern RC), whereas an incorrect initial response resulted in false reassurance (pattern WI_{ref}).

4.2.5 Measures

The experiment utilized within-subject design with two independent variables: consequence levels (high versus low) and performance patterns (Table 4.1). Each pattern indicates the combination of right or wrong initial actions by participants (indicated by the first letter) and the AI's correct or incorrect prediction (indicated by the second letter), which could not be manipulated (Kim et al., 2025b).

Occurrence of Error Patterns. The occurrence of error patterns were determined by the combination of participants' initial responses and the correctness of AI.

Trust Adjustment. After each trial i, participants reported their $trust(i)$ in the decision aid, on a 0-100 visual analog scale (Appendix B.3). We calculate a trust adjustment as:

$$\text{Trust adjustment}(i) = \text{Trust}(i) - \text{Trust}(i-1), \quad \text{for } i = 2, 3, \ldots, 60 \tag{4.1}$$

Since the moment-to-moment trust is reported after each trial, only 29 trust adjustments are obtained from each participant for each consequence level.

Performance. Another dependent variable was the final performance of the experimental task, which was calculated as the percentage of correct verification performance (determining the validity of the initial answers) for each pattern for each participant (Figure 4.2a).

Reaction Time. The reaction time was measured in seconds from when the AI prediction

appeared until the participant pressed either the "I was right" or "I was wrong" button (Figure 4.2a).

Eye Tracking Metrics. Eye tracking metrics included three measurements within three pre-defined AOIs: average fixation duration (mean time in milliseconds the participants' eyes remained relatively still), the total fixation duration (cumulative time in milliseconds of all fixations), and the fixation count (the number of fixations).

4.2.6 Experimental procedure

Participants were recruited through email invitations. Upon arrival at the lab, they reviewed and signed an informed consent form. They first completed a pre-experiment questionnaire assessing trust propensity (Merritt et al., 2013) (Appendix A.1), then watched a five-minute instructional video explaining the experimental procedures and received training to familiarize themselves with the test environment. The training phase included four practice trials of the mental rotation task (MRT). Following training, participants completed 60 experimental MRT trials. After finishing the task, they completed a post-experiment questionnaire (Muir and Moray, 1996) (Appendix A.2) and were compensated according to their performance. The entire session lasted approximately 50 minutes.

4.2.7 Statistical Analysis

To determine whether error pattern frequencies differed across consequence conditions, a Pearson's chi-square test was conducted. For all other dependent variables, mixed-effects generalized linear models (GLMs) were estimated using the lme4 package (Bates et al., 2014), with participant ID included as a random intercept. Separate mixed linear models were constructed for each error pattern, and consequence levels were compared using Welch's two-sample t-tests. Statistical significance was defined as $p < .05$, with $p < .1$ interpreted as marginal significance. All analyses were conducted using R statistical software, version 4.2.2 (R Core Team, 2013).

4.3 Results

4.3.1 Occurrence of Error Patterns Across Consequence Levels

The chi-square analysis indicated a significant association between consequence level and the distribution of error patterns ($\chi^2(4) = 349.82, p < .001$). Notably, patterns WC, WI, and WI_{ref} were absent in the high-consequence condition because participants did not commit

initial response errors in that condition (see Table 4.1). Although some variation in pattern frequency across conditions was expected, the complete absence of incorrect initial responses under high-consequence trials was unanticipated. Importantly, prior validation data from two earlier studies (n = 70) confirmed that stimulus difficulty was equivalent across consequence conditions. Paired-sample t-tests showed no significant difference between low-consequence ($Mean(sd) = .559(.178)$) and high-consequence ($Mean(sd) = .549(.168)$) stimuli ($t(69) = 0.738, p = .46$) (Kim et al., 2023, 2024).

Table 4.1: Results table summarizing dependent measures by error patterns.

Performance Pattern	Consequence Level	Occurrence	Trust Change	Performance	Reaction Time
RC	High	424	2.33 (4.51)	.972 (.057)	4.00 (1.61)
	Low	420	3.21 (6.84)	.977 (.069)	4.13 (1.77)
RI	High	101	-6.82 (11.06)	.674 (.368)	6.00 (2.00)
	Low	109	-4.44 (6.45)	.685 (.413)	5.57 (2.07)
WC	High	0	-	-	-
	Low	324	0.99 (1.75)	.649 (.295)	6.07 (1.58)
WI	High	0	-	-	-
	Low	89	-3.90 (9.36)	.704 (.343)	5.99 (2.29)
WI_{ref}	High	0	-	-	-
	Low	45	-6.31 (11.74)	.191 (.309)	5.44 (2.70)

4.3.2 Within-Pattern Effects of Consequence

Patterns WC, WI, and WI_{ref} were absent in the high-consequence scenario and could not be compared across consequence conditions. Therefore, the subsequent analyses focused on patterns RC and RI. Additionally, data from six participants were excluded due to eye-tracking malfunctions.

Pattern Right-Incorrect (RI). For pattern RI (when AI predictions were incorrect, following participants' right initial actions), participants showed significantly longer average fixation duration on answer choices ($F(1, 149.06) = 4.07, p < .05$), average fixation duration on reference ($F(1, 149.32) = 5.30, p < .05$), and total fixation duration on reference ($F(1, 150.38) = 6.45, p < .05$) during high-consequence trials compared to during low-consequence trials (Table 4.2).

The eye-tracking data corresponded closely with observed changes in trust and reaction time. A marginal effect of consequence level on trust adjustment was found ($F(1, 145.84) = 3.47, p = .065$), with slightly larger decreases in trust during high-

consequence trials ($Mean(SD) = -6.89(13.81)$) compared to low-consequence trials ($Mean(SD) = -3.71(6.94)$). Similarly, reaction time showed a marginal difference between consequence conditions ($F(1, 155.06) = 3.43, p = .066$), with participants responding marginally slower during high-consequence trials ($Mean(SD) = 5.95(2.48)$) versus low-consequence trials ($Mean(SD) = 5.58(2.61)$). No significant differences were observed in overall performance ($F(1, 155.42) = 0.41, p = .524$) (Table 4.1).

Pattern Right-Correct (RC). For pattern RC, where AI predictions were correct following participants' correct initial responses, participants exhibited shorter average fixation durations on the AI prediction ($F(1, 642.73) = 9.10, p < .01$), shorter total fixation durations ($F(1, 641.58) = 13.05, p < .001$), and fewer total fixations ($F(1, 642.55) = 8.21, p < .01$) in high-consequence trials compared to low-consequence trials (Table 4.2).

When AI predictions were correct, trust change ($F(1, 660.23) = 1.22, p = 0.270$), reaction time ($F(1, 676.48) = 1.47, p = 0.142$), and performance ($F(1, 689.72) = 1.67, p = 0.100$) did not significantly differ between high- and low-consequence conditions (Table 4.1).

4.4 Discussion

Although our main aim was to examine how consequence levels affect trust across error patterns, we could not analyze wrong initial actions (patterns RI, WI, and WI_{ref}) under high-consequence conditions because participants did not commit such errors (Table 4.1). The primary distinction between consequence levels was the background color: blue for low-consequence trials (Figure 4.2a) and red for high-consequence trials (Figure 4.2b). Interestingly, no initial errors occurred in high-consequence (red background) trials. To maintain AI reliability above 70% and avoid participant fatigue, the number of WI trials was limited to six, and WI_{ref} trials to three per consequence level. The complete absence of wrong initial responses was unexpected, given prior studies demonstrated balanced task difficulty between high- and low-consequence conditions (Kim et al., 2023, 2024). This suggests that when AI systems flag potentially high-consequence situations, users become more vigilant, reducing initial errors and enhancing human-AI collaboration.

Prior work on the effect of error consequences on trust in automation has yielded mixed results. Some studies report increased reliance on automation in high-consequence contexts, while others find the opposite (Ezer et al., 2008; Lyons and Stokes, 2012; Perkins et al., 2010; Chancey et al., 2017). Our findings support the latter: participants relied less on AI when potential errors had higher consequences. Eye-tracking measures reinforced this, showing reduced attention to AI predictions under high-consequence trials, regardless of whether the AI prediction was correct. Specifically, correct AI predictions resulted in

Table 4.2: Results table summarizing eye tracking measures by error patterns.

Performance Pattern	Consequence Level	Fixation Average Duration			Fixation Total Duration			Fixation Count		
		Reference	Answer Choice	AI Prediction	Reference	Answer Choice	AI Prediction	Reference	Answer Choice	AI Prediction
RC	High	249.34 (91.76)	256.54 (83.77)	230.95 (93.44)	921.14 (583.13)	545.16 (312.20)	597.03 (338.43)	2.07 (0.88)	3.07 (1.64)	1.98 (0.86)
	Low	249.51 (78.60)	249.28 (99.90)	266.99 (100.00)	1019.33 (798.86)	641.84 (412.87)	618.30 (377.61)	2.10 (1.00)	3.24 (1.98)	2.10 (1.05)
RI	High	322.34 (136.12)	302.36 (133.67)	297.51 (131.55)	1447.06 (956.56)	841.82 (498.40)	1096.58 (718.59)	3.06 (1.45)	4.23 (2.24)	2.71 (1.25)
	Low	281.73 (138.12)	261.82 (126.71)	280.49 (126.64)	1304.96 (988.06)	729.67 (463.73)	949.88 (644.64)	2.84 (1.46)	4.10 (2.02)	2.23 (1.22)
WC	High	-	-	-	-	-	-	-	-	-
	Low	294.50 (87.71)	283.76 (100.47)	267.89 (95.35)	1393.71 (760.75)	766.67 (352.76)	977.17 (486.44)	3.05 (1.25)	4.35 (1.89)	2.52 (0.91)
WI	High	-	-	-	-	-	-	-	-	-
	Low	301.29 (136.24)	296.65 (152.58)	298.55 (149.10)	1342.15 (929.13)	857.44 (513.80)	932.69 (721.07)	2.61 (1.72)	3.77 (2.54)	2.74 (1.44)
WI_Ref	High	-	-	-	-	-	-	-	-	-
	Low	285.83 (120.19)	232.10 (132.12)	243.75 (113.25)	692.53 (133.28)	654.28 (600.12)	638.75 (399.30)	2.20 (1.29)	3.28 (2.15)	2.25 (1.66)

shorter average and total fixation durations and fewer fixations in high-consequence trials, while incorrect predictions prompted longer fixations on answer choices and reference images. Although these measures do not directly quantify AI use, increased attention to manual verification areas implies participants relied less on AI when stakes were high. Trust changes mirrored these patterns, with marginally larger decreases following incorrect AI predictions in high-consequence trials, while correct predictions showed no significant difference across consequence levels.

Kim et al. (2025b) demonstrated that modeling AI performance beyond binary SDT classification uncovers error patterns that significantly affect trust. They found the largest trust decline occurred following AI's incorrect reassurance (pattern WI_{ref}), when participants were led to confirm a wrong initial choice. In our experiment, trust decreased more ($F(1, 969) = 2.86, p < .05$) for the WI_{ref} pattern ($Mean(sd) = -6.31(11.74)$) compared to the RI pattern ($Mean(sd) = -4.44(6.45)$) in the low-consequence condition (Table 4.1). There was a marginally greater trust drop for the RI pattern under high-consequence conditions. This trend suggests that incorrect reassurance (pattern WI_{ref}) in high-consequence scenarios may lead to even greater decline in trust, potentially increasing the likelihood of AI disuse. Due to the limited number of incorrect AI trials, we were unable to assess the effects of consequences on trust across all error patterns. Future work should consider increasing the number of trials to examine how consequence levels influence trust in AI, particularly in error patterns involving initial human errors (patterns WC, WI, and WI_{ref}), which could provide valuable insights into trust dynamics in high-risk domains.

The mental rotation task (MRT) was selected as a domain-independent, generalized task. While MRT does not reflect the domain specificity or real-world consequences of safety-critical environments, we were able to successfully manipulate perceived consequences through a point-based reward system: high-consequence trials penalized errors more severely, reducing bonus eligibility. Although monetary consequences are not equivalent to real-world risks, they were sufficient to induce behavioral differences. Participants were more attentive in high-consequence trials (no initial errors) and relied less on AI guidance, indicating that consequence influences trust even in non-safety-critical tasks. These findings suggest that varying consequence levels influence trust in a non-safety-critical testbed. Relating these results to real-world contexts, in safety-critical domains, operators may similarly reduce reliance on AI following errors, especially under high-stakes conditions (patterns RI, WI, and WI_{ref}, Table 4.1).

Future research should explicitly investigate how consequence shapes trust dynamics in domain-specific tasks and populations. Researchers are encouraged to classify human-AI performance into detailed multi-class error patterns, such as dividing incorrect AI error into

RI (AI's incorrect detection of the right human initial action), WI (AI's incorrect detection of the wrong human initial action), and WI_{ref} (false reassurance), to better understand trust responses to AI performance across varying consequence levels, particularly after initial human errors. Such insights could inform the design of reliable AI systems in high-risk environments.

4.5 Conclusion

This study examined how varying consequence levels of errors influence human trust in AI across different error patterns. Participants demonstrated reduced reliance on AI under high-consequence conditions, irrespective of AI accuracy. Notably, participants avoided initial errors in high-consequence trials, suggesting that AI capable of identifying and signaling high-stakes situations can increase vigilance and reduce mistakes.

In low-consequence trials, trust decreased more substantially for pattern WI_{ref} than for pattern RI, while high-consequence pattern RI showed a marginally greater trust decrement than low-consequence pattern RI. These trends indicate that incorrect reassurance (WI_{ref}) in high-consequence scenarios may lead to even greater decline in trust, and need to be further investigated.

Future research should explore consequence-driven trust dynamics using domain-specific tasks and user populations. classifying human-AI interactions into detailed error patterns (i.e., RI, WI, WI_{ref}) understand trust adjustments in response to AI performance across varying consequence levels. These insights could guide the development of safer and more reliable AI systems in high-risk domains.

CHAPTER 5

Evaluating the Effects of Stages of Automation (SoA) on Performance and Trust in AI-Assisted Tasks

5.1 Introduction

Section 1.2.2 discusses the stages of automation, with particular emphasis on the distinctions between the Decision Selection (DS) stage and the Information Analysis (IA) stage. Prior findings suggest DS automation may offer greater benefits than IA automation, especially under accurate conditions (Tatasciore et al., 2020; Wickens et al., 2007; Rovira et al., 2007; Picco et al., 2025; Gegoff et al., 2025; Wang et al., 2025). However, across both single and multi-task contexts, the majority of prior studies report that while IA and DS automation offer similar benefits when accurate, the cost of DS automation emerges when automation fails (Crocoll and Coury, 1990; Rovira et al., 2002; Sarter and Schroeder, 2001; Sethumadhavan, 2009; Rovira et al., 2007). These studies have classified automation accuracy in binary terms (correct or incorrect) using the SDT framework (Tanner Jr and Swets, 1954; Sorkin and Woods, 1985; Hautus et al., 2021). With the SDT framework, performance patterns WI_{ref} and WI are treated equivalently as incorrect AI performance.

However, in a multi-class framework, these two incorrect classifications lead to varying trust and performance outcomes. Chapter 3 examined the effects of multi-class performance patterns on human trust in AI systems using a mental rotation task (MRT) experiment. Their results showed that pattern WI led to a smaller decrement in trust, better verification performance, and longer reaction times compared to pattern WI_{ref}. These findings demonstrate that trust adjustment is shaped by both the AI accuracy and the utility of the AI's classification. Although both outcomes reflect AI's incorrect underlying classification, they differ in their "utility" for supporting potentially safe human-AI performance (Kim et al., 2025b).

As a result, the same underlying AI classification error can have different implications for human performance and trust depending on the stages of automation (SoA). At Stage

2, pattern *WI* constitutes an incorrect classification that implicitly prompts a safety check. At Stage 3, although the underlying AI classification remains incorrect, the AI's explicit recommendation to re-check is correct from a decision-support perspective (pattern *WC*∗). Therefore, an incorrect Stage 2 classification becomes a correct Stage 3 interaction-level prediction, explicitly revealing the unintended safeguard scenario.

The case of unintended safeguard pattern (*WI*) can only emerge in multi-class classification scenarios, and has been overlooked in prior trust in automation/autonomy/AI systems research, which predominantly relied on SDT-based binary performance representation (Kim et al., 2025b). This research, therefore, aims to examine how varying stages of automation (SoA) influence human trust and performance in AI-assisted tasks, with particular attention to the cases in which unintended safeguards arise (the last row of Table 5.1). The following section reviews related literature on levels and stages of automation.

Table 5.1: Description of the five performance patterns based on human initial selection, AI classification, and AI advice. At Stage 2, pattern *WI* constitutes an incorrect classification that inadvertently prompts a safety check. At Stage 3, although the underlying AI classification remains incorrect, the AI recommendation to re-check is correct from a decision-support perspective, explicitly highlighting the unintended safeguard.

Reference (Prescription)	Human Initial Selection	AI Classification	Stage-2-AI Performance Pattern	AI Advice	Stage-3-AI Performance Pattern
Shape "X"	Shape "X"	Shape "X"	RC	"dispense"	RC
		Shape "Y" or "Z"	RI	"refill"	RI
	Shape "Y"	Shape "Y"	WC	"refill"	WC
		Shape "X"	WI_{ref}	*"dispense"*	WI_{ref}
		Shape "Z"	WI	***"refill"***	**WC***

Acronym descriptions: (*RC*=Right-Correct, *RI*=Right-Incorrect, *WC*=Wrong-Correct, *WI*=Wrong-Incorrect, *WC*∗=Wrong-Correct due to unintended safeguards, WI_{ref}=Wrong-Incorrect-Matches-Reference).

5.1.1 Proposition

In this section, we mathematically demonstrate that Stage-3-AI (decision selection support) yields greater interaction-level accuracy compared to Stage-2-AI (information analysis support). After that, we present our hypotheses.

Proposition 1 *Let X indicate the prescribed medication. Let $t \in [0, 1]$ denote the probability that the pharmacist correctly fills medication X. Consequently, $1 - t$ denotes the probability that the pharmacist fills the prescription incorrectly with some other medication, Y, where $Y \neq X$ (indicates an incorrect action). Once the bottle is filled, the AI system scans it and classifies the medication into one of three labels: X (the prescribed medication), Y (the*

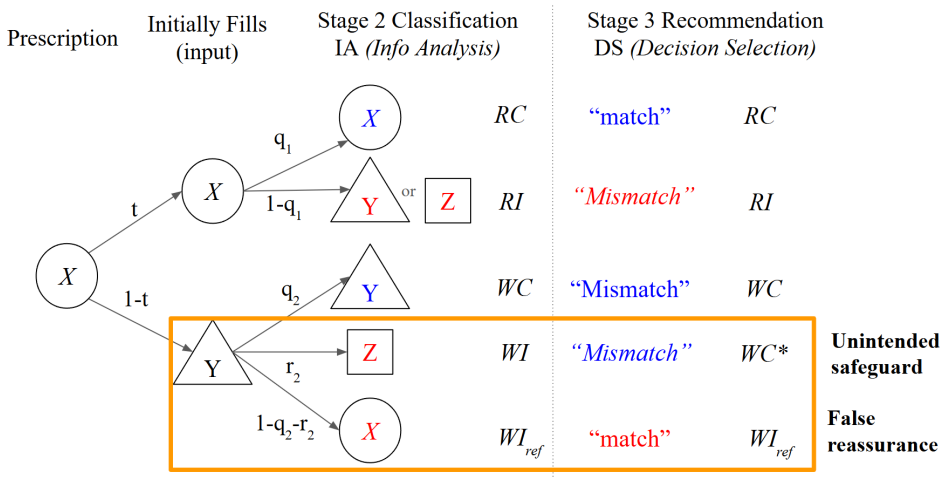


Figure 5.1: AI-assisted medication dispensing procedure is displayed with the probabilities. The red text indicates incorrect classifications and advice. Note that the last row of AI incorrectly predicting Y as Z (pattern WI, Stage-2-AI) results in a correct advice to "refill" (pattern $WC*$, Stage-3-AI).

specific incorrectly filled medication), or Z (any other unrelated medication that is neither X nor Y).

Let:

- $q \in [0, 1]$ *denote the probability that the AI correctly identifies the filled medication;*
- $r \in [0, 1]$ *denote the probability that the AI classifies the filled medication as Z;*
- $1 - q - r$ *denote the residual probability that the AI misidentifies the filled medication as the incorrectly filled medication (i.e., AI predicts Y when X was filled, or X when Y was filled).*

We define the AI's overall reliability as the probability that it produces correct outcomes. This depends on the system's Stage of Automation (SoA)*, i.e., at Stage-2-AI, AI identifies the state of the world; at Stage-3-AI, AI recommends the action:*

With Stage-2-AI, the AI is considered correct only if it correctly identifies the filled medication. Then the overall correctness is:

$$P_{Correct,\ Stage\text{-}2\text{-}AI} = tq + (1 - t)q = q.$$

With Stage-3-AI, the AI is considered correct if it recommends the correct action, either "dispense" or "refill". Then the overall correctness is:

$$P_{Correct,\ Stage\text{-}3\text{-}AI} = tq + (1 - t)(q + r) = q + r(1 - t).$$

Comparing the two expressions, we obtain:

$$P_{Correct,\ Stage\text{-}3\text{-}AI} \geq P_{Correct,\ Stage\text{-}2\text{-}AI},$$

since:

$$q + r(1-t) \geq q \quad because\ r(1-t) \geq 0.$$

Thus, the overall correctness probability is always greater under the Stage-3-AI condition than under the Stage-2-AI condition.

5.1.2 Hypotheses

To examine the proposition, we conducted an in-person, controlled laboratory experiment using a mental rotation task (MRT).

We wish to replicate results from previous studies showing greater performance decrements with an increasing SoA under unreliable conditions (Rovira et al., 2002; Sarter and Schroeder, 2001; Sethumadhavan, 2009; Crocoll and Coury, 1990), and a shorter reaction time with an increasing SoA under all conditions (Rovira et al., 2002; Gegoff et al., 2025; Wickens et al., 2007). Therefore, we hypothesize that:

> **H1a** When the AI is correct in both classification and decision recommendation, Stage-3 AI will yield better task performance and greater trust increment compared to Stage-2-AI, reflecting the benefits of decision-selection support.
>
> **H1b**: When AI is incorrect in both classification and decision recommendation, Stage-3-AI will yield worse task performance and greater trust decrement compared to Stage-2-AI.
>
> **H1c**: Regardless of AI correctness, Stage-3-AI will yield faster reaction times than Stage-2-AI

We are particularly interested in the case of *unintended safeguard*, which occurs when a misidentification at Stage-2 AI leads to a correct recommendation at Stage-3 AI. We hypothesize that:

> **H2a**: When the case of unintended safeguard occurs, Stage-3-AI will lead to better task performance and faster response time compared to Stage-2-AI.
>
> **H2b**: When the case of unintended safeguard occurs, Stage-3-AI will lead to trust increment and Stage-2-AI will lead to trust decrement.

As shown in our proposition, the overall accuracy of Stage-3-AI would be greater than or equal to that of Stage-2-AI. Prior research has shown that automation accuracy strongly influences operator dependence (Chancey et al., 2017; Dixon and Wickens, 2006), which would positively affect overall human-system performance. Therefore, we hypothesize:

> **H3a**: Stage-3-AI will lead to better overall task performance and faster overall reaction time than Stage-2-AI.
>
> **H3b**: Stage-3-AI will result in higher perceived usability, greater trust, and higher perceived reliability compared to Stage-2-AI.

5.2 Methods

5.2.1 Ethical Considerations

This research complied with the ethical guidelines established by the American Psychological Association and qualified for exemption from the Institutional Review Board oversight at the University of Michigan (*HUM00230326*).

5.2.2 Participants

Ninety adults with a mean age of 22.18 years (SD = 2.88) took part in the experiment. All participants had normal or corrected-to-normal vision. They were compensated with a base payment of $15, with the opportunity to earn up to an additional $10 based on their performance.

5.2.3 Experimental Task

This study utilized a simulated AI-assisted Mental Rotation Task developed by Kim et al. (2025b). The experiment consisted of two 30-trial blocks with different stages (Stage-2-AI and Stage-3-AI). The flowchart of the experimental task is presented in Figure 5.2.

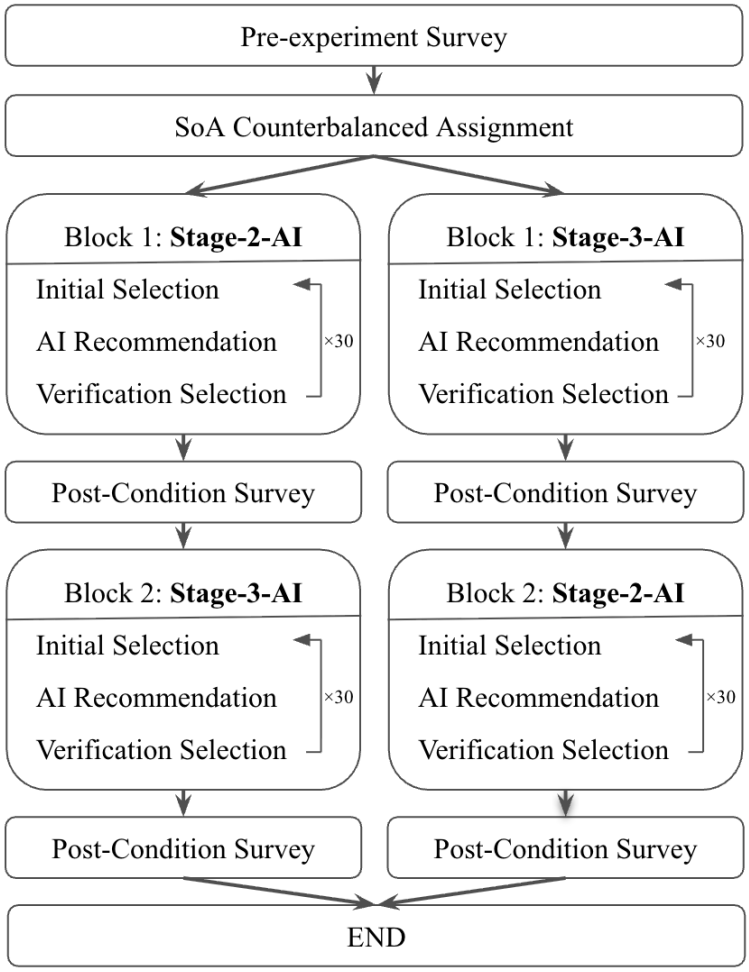


Figure 5.2: Flowchart of the SoA experiment.

Initially, participants were presented with a reference image along with five options: one correct alternative and four distractors (see Figure 5.3). The correct selection was structurally identical to the reference image, but was rotated in the horizontal and/or vertical axes. Participants were instructed to accurately identify the image representing the same 3D object as the reference within 15 seconds. When participants failed to make a selection within the allocated time, the interface displayed the message "You did not make a selection within the time limit." The participants had to click "Next" to proceed to the subsequent trial.

Following their initial selection, participants provided their confidence level using a visual analog scale to provide their self-perceived certainty about their initial response. The scale ranged from "Not confident at all" on the left extreme to "Absolutely confident" on the right extreme.

After making their initial selection, participants received the AI system's recommendation

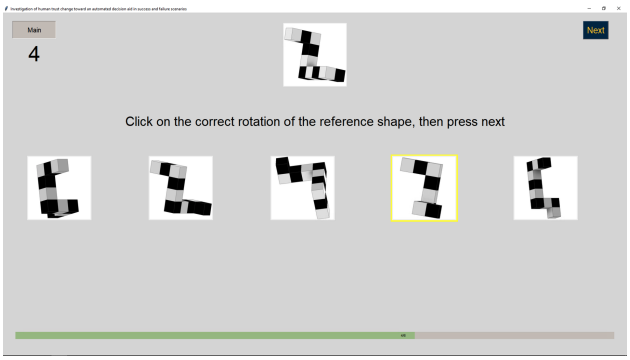
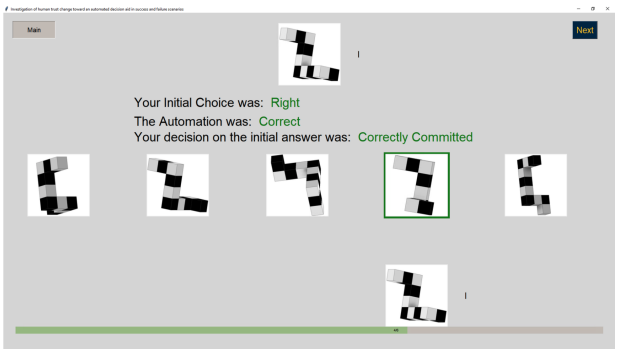


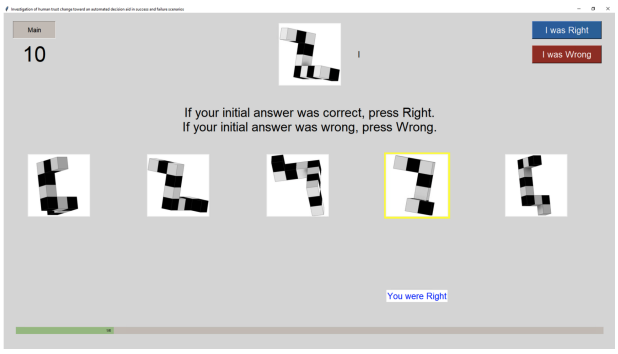
Figure 5.3: Participants selected an initial answer within 15 seconds. Their initial selection was highlighted in yellow.

in one of two forms. The Stage-2-AI predicted the shape that the participant selected and presented a rotation closely matching the reference image orientation (Figure 5.4a). This design allowed participants to visually compare their selection with the reference, implicitly helping the decision-making process. In contrast, Stage-3-AI provided explicit advice. If the AI's classification of the participants' initial selection aligned with the reference image, it advised the participant to "accept" their initial answer (Figure 5.4b). Conversely, if there was a mismatch, the AI recommended to "reject" the initial selection.

After reviewing the AI's prediction, participants were given 10 seconds to make a final verification decision. They could either confirm their initial answer by selecting "I was right" or flag their initial answer by clicking "I was wrong." This time-constrained decision-making process added an element of pressure to the task. If participants failed to make a selection within the allocated 10 seconds, they were skipped to the next trial.



(a) Stage-2-AI Prediction



(b) Stage-3-AI Advice

Figure 5.4: AI advice was presented in Stage-2-AI or Stage-3-AI format. The Stage-2-AI displayed its prediction, oriented to closely match the angle of the reference image. The Stage-3-AI provided explicit advice.

Upon completing each trial, participants were presented with performance feedback, which included assessments of the participant's initial selection accuracy and the AI's accuracy (see Figures 5.5a and 5.5b).

Following the performance feedback, participants used two visual analog scales to report their trust in the AI (ranging from "I don't trust the decision aid at all" on the far left to "I absolutely trust the decision aid" on the far right) and their perception of the AI's reliability (ranging from 0 to 100).



(a) Stage-2-AI Feedback

(b) Stage-3-AI Feedback

Figure 5.5: Feedback was presented following the final verification selection. Feedback included the accuracy of participants' initial selection (right or wrong), the accuracy of AI prediction (correct or incorrect), and the accuracy of participants' final selection (correctly committed or incorrectly committed).

5.2.4 Experimental Design

The experiment utilized a 2x3 within-subjects design. The independent variables were stages of automation (Stage-2-AI or Stage-3-AI) and outcome type. Outcome type was defined based on the correctness of the Stage-2 classification and the resulting Stage-3 recommendation, and was categorized as simultaneously correct, simultaneously incorrect, and unintended safeguard (Table 5.2). The order of Stage-2-AI and Stage-3-AI was counterbalanced across participants to control for order effects.

5.2.5 Measures

We collected a range of performance, behavioral, and subjective measures. Regarding performance and behavioral measures, we measured participants' change in performance from initial to final responses, as well as participants' reaction time.

In terms of subjective measures, after each trial i, participants reported their perceived

Table 5.2: Outcome type was defined based on the correctness of the Stage-2 classification and the resulting Stage-3 recommendation, and was categorized as simultaneously correct, simultaneously incorrect, and unintended safeguard.

Outcome Type	Correct Stage-2 Classification	Correct Stage-3 Recommendation
Simultaneous Correct	Yes	Yes
Simultaneous Incorrect	No	No
Unintended safeguard	No	Yes

reliability of and their dynamic trust in their automation, *Perceived reliability*(i) and $Trust(i)$. Based on them, we calculated each participant's trust adjustment and their average perceived reliability of the automation. After participants completed a 30-trial block with each SoA, they reported their trust in the automation and the perceived usability of the automation.

ΔPerformance. For each trial, participants' initial performance was measured based on the accuracy of their initial response: 1 indicates right and 0 wrong (see Figure 5.3). Similarly, their final performance was determined by whether they correctly selected "I was right" or "I was wrong" (see Figure 5.5b). We defined ΔPerformance as a categorical variable showing the difference between the final and initial performance, where $\Delta\text{Performance} \in \{-1 = decrement, 0 = same, 1 = increment\}$.

$$\Delta\text{Performance}(i) = \text{Final Performance}(i) - \text{Initial Performance}(i), \quad i = 1, 2, 3, \ldots, 30$$

Reaction Time. For each trial, reaction time was measured in seconds from the moment the AI prediction appeared to when the participant pressed either the "I was right" or "I was wrong" button (see Figure 5.4b). A logarithmic transformation was applied to the reaction time data prior to statistical analysis to reduce skewness and improve normality.

Trust Adjustment. After each trial i, participants reported their $trust(i)$ in the AI decision aid. We calculate a trust adjustment as:

$$Trust\ adjustment(i) = Trust(i) - Trust(i-1), \text{ where } i = 2, 3, ..., 30$$

The moment-to-moment trust was reported after each trial. Therefore, 29 trust adjustments were obtained from each participant in each SoA.

Post-Condition Perceived Reliability Rating. As part of the post-condition questionnaire, participants rated perceived reliability on a visual analog scale (0-100) after each SoA interaction.

Post-Condition Trust Rating. After experiencing each SoA, participants completed a 9-item trust questionnaire (Muir and Moray, 1996).

System Usability Scale Score. After experiencing each SoA, participants completed the System Usability Scale (SUS) questions (Brooke et al., 1996).

5.2.6 Experimental procedure

Prior to the experiment, participants completed a pre-experiment survey, including demographic questions and a trust propensity survey to assess their general tendency to trust AI (Merritt et al., 2013). They then watched a video explaining the experimental task and were informed that the AI recommendations may be correct or incorrect. Participants were assigned to either the Stage-2-AI-First or Stage-3-AI-First group using a counterbalanced design. After each block of 30 trials, participants completed a post-condition survey.

5.2.7 Statistical Analysis

We built mixed-effects models to analyze data on Δperformance, reaction time, and trust adjustment, with outcome type (Table 5.2) and Stage of Automation (SoA) as fixed effects and participants modeled as random intercepts. Degrees of freedom (df) were estimated using the Kenward-Roger method, which adjusts the variance-covariance matrix of the fixed effects via a Taylor series expansion before computing df (Kenward and Roger, 1997). This approach is widely used in mixed-effects modeling because it reduces Type I error rates, yields more accurate p-values, and remains reliable under small samples, unbalanced designs, and complex random-effects structures (McNeish and Stapleton, 2016; Luke, 2017; Kuznetsova et al., 2017). Since Δperformance is an ordinal categorical variable, we used mixed-effects ordinal logistic regression, specifically, cumulative link mixed models (CLMMs) were implemented via the clmm function in the ordinal package in R (Christensen, 2019, 2018). Reaction time and trust adjustment, both continuous variables, were analyzed using linear mixed-effects models (LMMs), implemented via the lme4 package in R (Bates et al., 2014).

We adopted an iterative approach to construct models. Following standard approaches for mixed-effects modeling, models were initially specified with a simple structure and progressively extended to include additional fixed and random effects (Field et al., 2012). Model comparisons were conducted using likelihood ratio tests (LRTs), and the simpler model was kept when no significant improvement in fit was observed. Statistical significance was determined at $\alpha = .05$, and marginal significance was noted for $p < .1$.

To analyze post-condition survey data (i.e., SUS score, trust rating, and perceived reliability rating), we conducted paired-sample t-tests.

5.3 Results

The results section is structured in subsections that progressively broaden the scope of analysis. We first examine differences between SoA when unintended safeguards occur. Next, we analyze SoA differences within each AI accuracy (correct vs incorrect). Finally, we compare outcomes based on the SoA (Stage-2-AI vs. Stage-3-AI).

Initial reaction time values (measured during the stage shown in Figure 5.3) exceeding three standard deviations of the mean across the whole dataset were classified as outliers and excluded from all analyses. A log scale transformation was applied to the final reaction time data for the statistical analyses. For ease of interpretation, descriptive statistics (means and standard deviations) are reported in seconds.

Tables 5.3 and 5.4 summarize descriptive statistics for performance, reaction time, and trust adjustment. Marginal averages for each SoA level are presented in Table 5.5.

5.3.1 Comparing between SoA when AI is correct and incorrect

In this section, we present results when AI is simultaneously correct or incorrect in both classification and decision recommendation.

Table 5.3: Counts of ΔPerformance categories and means (with standard deviations) of reaction time and trust adjustment, grouped by SoA and AI Correctness.

SoA	Binary Representation			ΔPerformance	Count (n)	Total Count (n)	Reaction Time (s)	Trust Adjustment
	Initial Response	AI Correctness	Final Response					
Stage 3 (Correct Recommendation)	1	1	0	-1	24	1236	3.23 (2.09)	1.32 (4.94)
	1	1	1	0	1212			
	0	1	0	0	164	735	5.38 (2.83)	1.57 (6.32)
	0	1	1	1	571			
Stage 2 (Correct Classification)	1	1	0	-1	41	1201	4.00 (2.20)	1.63 (6.45)
	1	1	1	0	1160			
	0	1	0	0	283	738	6.13 (2.63)	1.34 (5.35)
	0	1	1	1	455			
Stage 3 (Incorrect Recommendation)	1	0	0	-1	78	323	5.45 (2.46)	-6.14 (9.84)
	1	0	1	0	245			
	0	0	0	0	80	100	4.43 (2.99)	-4.92 (7.46)
	0	0	1	0	232			
Stage 2 (Incorrect Classification)	1	0	0	-1	60	292	6.27 (2.33)	-5.89 (10.64)
	1	0	1	0	232			
	0	0	0	0	82	110	5.11 (2.50)	-5.00 (11.60)
	0	0	1	1	28			

Note that Count (n) refers to the number of observations within each correctness category (final response right (1) or wrong (0)), whereas total count refers to the sum of right and wrong observations within a given pattern.

ΔPerformance. When AI was correct, the effect of SoA was significant ($\beta = .308$, $SE =$

.072, $z = 4.26$, $p < .001$), indicating that participants had 1.36 times greater odds (95% $CI = [1.18, 1.57]$) of improving Δperformance using Stage-3-AI compared to using Stage-2-AI. When AI was incorrect, the effect of SoA was marginally significant ($\beta = -.325$, $SE = .170$, $z = -1.904$, $p = .057$), indicating that Stage-3-AI marginally reduced odds of improving Δperformance by .72 (95% $CI = [.52, 1.01]$) compared to Stage-2-AI (Table 5.3).

Reaction Time. When AI aid was correct, participants showed significantly shorter reaction time when using Stage-3-AI (Mean (SD) = 4.03s (2.61)) compared to using Stage-2-AI (Mean (SD) = 4.79s (2.59)) ($t(3783.02) = -13.42$, $p < .001$, Cohen's $d = -.432$). When AI aid was incorrect, shorter reaction time was also observed in the Stage-3-AI condition (Mean (SD) = 5.21s (2.63)) compared to the Stage-2-AI condition (Mean (SD) = 5.96s (2.43)) ($t(752.61) = -5.11$, $p < .001$, Cohen's $d = -.365$) (Table 5.3).

Trust adjustment. When AI aid was correct, no difference in trust adjustment was found between Stage-3-AI (Mean(SD) = 1.41(5.49)) and Stage-2-AI (Mean(SD) = 1.52(6.06)) conditions ($t(3622.42) = -0.650$, $p = .516$, Cohen's $d = -.02$). When AI aid was incorrect, no difference in trust adjustment was observed in the Stage-3-AI condition (Mean(SD) = -5.85(9.35)) compared to the Stage-2-AI condition (Mean(SD) = -5.65(10.90)) ($t(756.59) = -0.940$, $p = .348$, Cohen's $d = -.067$) (Table 5.3).

5.3.2 Comparing between SoA when Unintended Safeguards occur

Table 5.4: Counts of ΔPerformance categories and means (with standard deviations) of reaction time and trust adjustment, grouped by Final Response Accuracy and SoA, for unintended safeguard.

Outcome Type	SoA	Binary Representation			ΔPerformance	Count (n)	Total Count (n)	Reaction Time (s)	Trust Adjustment
		Initial Response	Automation Correctness	Final Response					
Inadvertent Safeguard	Stage 3	0	1	0	0	40	187	5.29 (2.91)	1.16 (6.03)
		0	1	1	1	147			
	Stage 2	0	0	0	0	91	221	6.43 (2.75)	-2.38 (9.84)
		0	0	1	1	130			

Note that Count (n) refers to the number of observations within each correctness category (final response right (1) or wrong (0)), whereas total count refers to the sum of right and wrong observations within a given pattern.

ΔPerformance. We conducted a mixed-effects ordinal logistic regression to examine whether SoA predicted changes in ΔPerformance (-1 = decrement, 0 = same, 1 = increment). The model revealed a statistically significant effect of SoA, $\beta = 1.00$, $SE = .242$, $z = 4.15$, $p < .001$. The odds ratio (OR) was 2.75 (95% $CI = [1.71, 4.44]$), indicating that

participants using Stage-3-AI had 2.75 times the odds of being in a higher ΔPerformance category (e.g., 1 vs 0, or 0 vs -1) compared to those using the Stage-2-AI (Table 5.4).

Reaction Time. A pairwise comparison of unintended safeguard outcomes indicated that Reaction time in Stage-3-AI (Mean (SD) = 5.29s (2.91)) was significantly shorter than in Stage-2-AI (Mean (SD) = 6.43s (2.75)) ($t(351.21) = -5.04$, $p < .001$, Cohen's $d = -.536$) (Table 5.4).

Trust adjustment. A pairwise comparison of unintended safeguard outcomes indicated that trust adjustment in Stage-2-AI was significantly different than the trust adjustment in Stage-3-AI ($t(385.39) = 4.47$, $p < .001$, Cohen's $d = .457$) (Table 5.4).

Trust decrement shown in pattern WI of Stage-2-AI (Mean(SD) = -2.38 (9.84)) was significantly below 0 ($t(175) = -4.09$, $p < .001$, Cohen's $d = -0.31$). In contrast, trust increment shown in pattern $WC*$ of Stage-3-AI (Mean(SD) = 1.16 (6.03)) was marginally above 0 ($t(208) = 1.89$, $p = .06$, Cohen's $d = 0.15$).

5.3.3 Comparing Between SoA (Stage-2-AI vs. Stage-3-AI)

Overall ΔPerformance. The effect of SoA on overall Δperformance was significant ($\beta = .211, SE = .061, z = 3.48, p < .001$), indicating that participants had 1.24 times the odds (95% $CI = [1.10, 1.39]$) of improving Δperformance using Stage-3-AI compared to using Stage-2-AI (Table 5.5).

Overall Reaction Time. Significantly shorter reaction time was observed in the Stage-3-AI (Mean (SD) = 4.31s (2.68)) compared to in the Stage-2-AI (Mean (SD) = 5.12s (2.64)) ($t(89) = 5.02$, $p < .001$, Cohen's $d = .53$) (Table 5.5).

Table 5.5: Counts of ΔPerformance categories and means (with standard deviations) of reaction time and trust adjustment, grouped by SoA. The table presents the marginal average.

SoA	ΔPerformance	Count (n)	Total Count (n)	Reaction Time (s)	Trust Adjustment
Stage 3	-1	102	2581	4.31 (2.68)	.196 (6.85)
	0	1741			
	1	738			
Stage 2	-1	101	2562	5.11 (2.64)	.054 (7.87)
	0	1848			
	1	613			

Note that Count (n) refers to the number of observations within each ΔPerformance category, whereas total count refers to the sum of all observations within each Stage.

System Usability Scale (SUS) Score. A significantly higher SUS score was observed after participants interacted with the Stage-3-AI (Mean(SD) = 75.35 (12.27)) compared to

using the Stage-2-AI (Mean(SD) = 71.37 (15.42)) ($t(89) = 2.64$, $p < .01$, Cohen's $d = .27$) (Figure 5.6).

Post-Condition Trust Rating. A significantly higher mean post condition trust rating was observed after participants used the Stage-3-AI (Mean(SD) = 64.35 (15.87)) compared to using the Stage-2-AI (Mean(SD) = 60.26 (17.17)) ($t(89) = 1.99$, $p < .05$, Cohen's $d = .21$) (Figure 5.6).

Post-Condition Perceived Reliability Rating. A significantly higher post-condition perceived reliability rating was observed after participants interacted with the Stage-3-AI (Mean(SD) = 79.72 (10.50)) compared to using the Stage-2-AI (Mean(SD) = 74.36 (15.97)) ($t(89) = 3.13$, $p < .01$, Cohen's $d = .33$) (Figure 5.6).

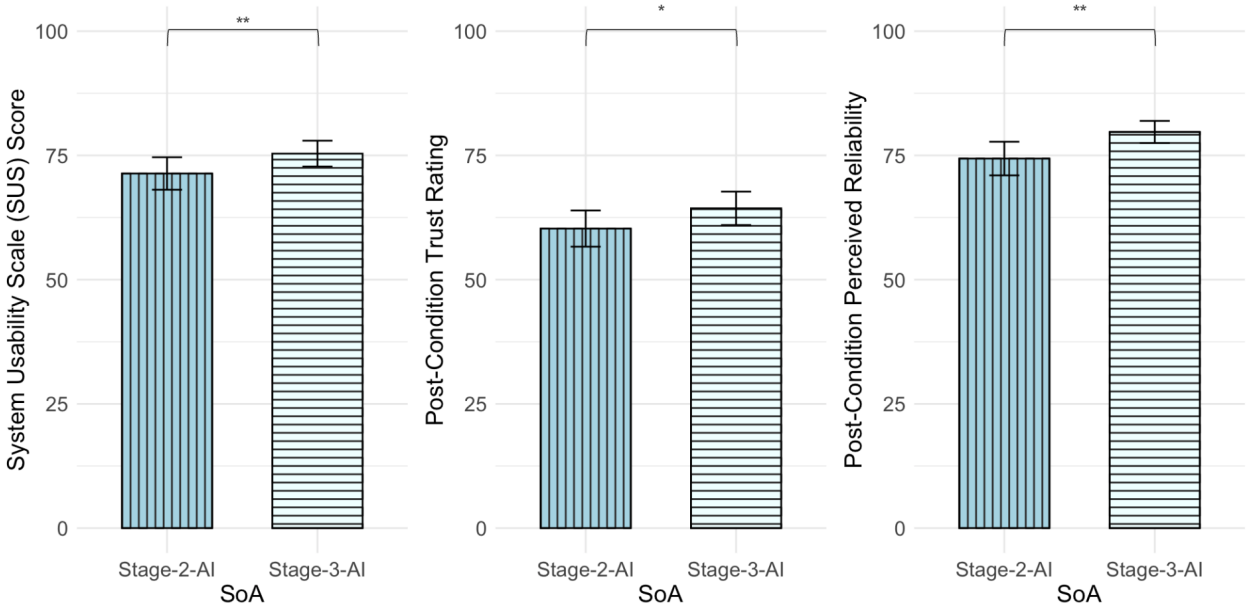


Figure 5.6: Comparing system usability score (SUS) rating, post-condition trust rating, and post-condition perceived reliability between the two SoA

5.4 Discussion

In this section, we discuss the findings in relation to our aims and hypotheses.

Prior literature on stages of automation (SoA) cautions that higher SoA automation should be implemented carefully in safety-critical systems, because the automation failures at higher stages are more likely to be missed by human operators than failures at lower stages (Rovira et al., 2002; Sarter and Schroeder, 2001; Sethumadhavan, 2009; Crocoll and Coury, 1990). When unintended safeguards were excluded, consistent with established findings, par-

ticipants using Stage-3-AI experienced greater performance decrements when AI provided incorrect recommendations, reflected by a marginally greater decrease in ΔPerformance (**H1b** partially supported). In contrast, when AI recommendations were correct, Stage-3-AI yielded greater ΔPerformance improvements than correct classification in Stage-2-AI (**H1a** partially supported). Regardless of AI accuracy, trust adjustment remained stable across the two SoA conditions. Reaction times were consistently faster with Stage-3-AI than Stage-2-AI (**H1c** supported), replicating prior evidence that higher stages of automation can reduce cognitive workload by supporting decision selection (Rovira et al., 2002; Wickens et al., 2007; Gegoff et al., 2025).

Expanding upon the foundational SoA research that modeled automation errors using binary classifications (i.e., correct vs. incorrect) (Crocoll and Coury, 1990; Rovira et al., 2002; Sarter and Schroeder, 2001; Sethumadhavan, 2009; Rovira et al., 2007; Tatasciore et al., 2020; Wickens et al., 2007; Rovira et al., 2007; Picco et al., 2025; Gegoff et al., 2025; Wang et al., 2025), we introduce multi-class AI performance patterns, where unintended safeguards (pattern *WI*) can emerge, cases in which an incorrect AI classification produces an inadvertently correct recommendation (Kim et al., 2025b). When focusing on the unintended safeguard cases, Stage-3-AI resulted in significantly greater odds of Δ performance improvement and faster reaction time compared to Stage-2-AI (**H2a** fully supported), despite identical underlying classification accuracy. Moreover, trust increased under Stage-3-AI but decreased under Stage-2-AI had a trust decrement (**H2b** supported). Notably, although unintended safeguards in Stage-2-AI classification implicitly signaled a mismatch between the participants' initial selection and the AI's classification, participants did not consistently leverage this information to improve task performance. In contrast, when the same underlying AI classification was translated into an explicit decision-level guidance in Stage-3-AI recommendation, participants demonstrated significantly greater task performance improvements and larger trust increases. These results suggest that the "functional utility" of unintended safeguards is amplified when correct interaction-level predictions are conveyed as actionable advice (Stage-2-AI), rather than as raw underlying classification information in Stage-2-AI, where users must infer implications.

Studies have indicated that AI accuracy strongly influences automation dependence, that higher automation reliability promotes greater reliance and compliance (Dixon and Wickens, 2006; Chancey et al., 2017). The underlying AI classification accuracy in this study exceeded the commonly cited 70% usefulness threshold (Wickens and Dixon, 2007). The presence of unintended safeguards, where the incorrect classification in Stage-2-AI becomes correct advice in Stage-3-AI, further increased the interaction-level accuracy of Stage-3-AI, contributing to greater overall task performance, faster overall reaction times, greater system

usability, higher post-condition trust, and higher post-condition perceived reliability relative to Stage-2-AI (**H3** supported).

We proved with ***Proposition 1*** that the overall correct probability of Stage-3-AI, $q+r(1-t)$, is always greater than or equal to that of Stage-2-AI, q. The task performance advantage of Stage-3-AI becomes more pronounced as $r(1-t)$ increases. This probability is driven by two factors: wrong initial input (probability $1-t$, where Y is selected instead of X) and the AI's error of misclassifying the input as a Z (probability r), which is unrelated to both X and Y (see Figure 5.1). The hypothetical scenario in the introduction involved three simplified "classes" of medications, X, Y, and Z. In practice, the likelihood of unintended safeguards increases as the number of possible classes grows, expanding the margin for both input and classification errors. For example, an analysis of 10,399,324 electronic prescriptions from 2019 to 2021 reported 353,002 unique medications (Lester et al., 2022). With such a large number of classes, both the probability of wrong input $(1-t)$ and the probability of misclassification (r) may increase, as there are substantially more alternatives for both selecting the wrong medication and misclassifying it. In this setting, the incorrect input could be mapped not only to Z, but to a large set of other medications (n=353,000), increasing the likelihood that some misclassifications may still lead to operationally acceptable recommendations.

These benefits are likely to generalize to other complex real-world domains characterized by multi-class classification, such as TSA luggage screening, medical diagnosis, manufacturing, defense, and large language model auditing (Walter et al., 2024; Sundaram and Zeid, 2023; Souissi et al., 2019; Rashid et al., 2023; Rastogi et al., 2023). In such contexts, certain misclassifications may be operationally equivalent, allowing system-level safety to be preserved when they result in correct or safe decision recommendations. It is important to note, however, that the performance advantages of Stage-3 systems arise specifically when unintended safeguards are present. While these structural buffers can enhance the accuracy of final decision recommendations, they may also contribute to the perceived "black-box" nature of AI systems. Evidence suggests that when both Stage-2 and Stage-3 information are available, human operators tend to prioritize the correctness of Stage-3 outputs over incorrect Stage-2 classification (Gegoff et al., 2025). Accordingly, future research should explicitly investigate multi-class classification scenarios across Stage 2, Stage 3, and integrated Stage 2&3 configurations in a variety of domains. Examining how unintended safeguards influence performance across these configurations can provide deeper insight into the roles of transparency and explainability in systems where underlying model-level errors do not necessarily lead to incorrect decision recommendations.

5.4.1 Limitations

This study has several limitations. First, the current study was conducted over a relatively small number of trials (30 trials), which limits insights into how operator behavior and performance change over time. Future research could benefit from longer-duration designs to examine whether the benefits of Stage-3-AI remain as users adapt to and gain expertise with AI during extended collaboration.

Second, the experimental design prioritized accuracy over speed. The incentive structure rewarded correct verification outcomes rather than quicker responses, encouraging careful decision-making within the time constraint. However, real-world operational contexts vary widely in the degree and nature of time pressure they impose. In environments with higher temporal incentives or lower AI reliability, the balance between speed and accuracy may differ. Future research should systematically manipulate time pressure and AI reliability to examine how speed-accuracy trade-offs interact with SoA and unintended safeguards.

Third, while the findings provide general insights, the applicability to domain-specific and real-world contexts, especially in safety-critical environments, remains to be validated. Future studies should examine SoA in domain-specific simulations and multi-class classification tasks to assess whether the observed difference between incorrect classification in Stage-2-AI and correct recommendation in Stage-3-AI generalizes across complex operational settings. Additionally, extending this work to multi-task environments would clarify the influence of unintended safeguards under higher cognitive workloads and competing task demands.

5.5 Conclusion

Existing research on stages of automation (SoA) suggests that even though increasing SoA could benefit performance when reliable, failures result in greater negative consequences. Similar to prior literature, our findings also indicate that Stage-3-AI led to poorer task performance when AI recommendations were incorrect compared to Stage-2-AI.

However, we further broke down the binary dichotomy (accuracy classification of correct versus incorrect) into multi-classification cases and proved with ***Proposition 1*** that an AI system operating at a higher stage will yield greater interaction-level accuracy. This greater accuracy is driven by the unintended safeguards, in which incorrect classifications in Stage-2-AI become correct interaction-level advice in Stage-3-AI. The increased AI accuracy of Stage-3-AI led to a higher overall verification task performance, followed by quicker reaction time, a greater system usability score, a higher post-condition trust rating, and a higher post-condition perceived reliability.

Such benefits may extend to domains in which unintended safeguards can emerge during human-AI interaction in multi-class decision contexts, including TSA luggage screening, medical diagnosis, manufacturing, defense, and large language model auditing (Walter et al., 2024; Sundaram and Zeid, 2023; Souissi et al., 2019; Rashid et al., 2023; Rastogi et al., 2023).

CHAPTER 6

Conclusion and Future Directions

6.1 Summary, Intellectual Merit, and Practical Impact

This dissertation advances understanding of human trust and performance in AI-assisted decision-making by systematically examining how AI uncertainty, multi-class error categorization framework, severity of error consequence, and stages of automation shape human-AI interaction. Across four complementary research aims, this work moves beyond traditional binary performance categorization models of automation performance to capture the complexity inherent in multi-class classification scenarios of real-world, safety-critical AI-assisted decision environments.

6.1.1 Connecting to Existing Literature

Chapter 2 contributes to the literature on trust in AI-assisted healthcare technologies by examining pharmacists' trust in automated pill verification systems. Prior research in human-automation interaction has shown that transparency and uncertainty information can influence user trust and reliance on automated systems (Lee and See, 2004; Hoff and Bashir, 2015; Chatzimparmpas et al., 2020a). However, empirical work examining these factors in pharmacy workflows has been limited. Chapter 2 addressed this gap by investigating how pharmacists perceive and interact with an AI-assisted pill verification system, particularly when uncertainty information is presented. The findings suggest that pharmacists generally demonstrate a high disposition to trust automated systems, and that transparent presentation of AI uncertainty can influence trust calibration. This aim extends the human-AI trust literature into a domain where AI system errors may directly affect patient safety.

Chapter 3 expands the literature on human-AI trust by examining automation performance beyond the commonly used binary framework of signal detection theory (SDT) (Tanner and Swets, 1954; Hautus et al., 2021). The majority of the existing research in trust

in automation/autonomy/AI systems modeled automation outcomes as either correct or incorrect predictions (Wang et al., 2009; Du et al., 2020; Neyedli et al., 2011). While this binary categorization has supported many theoretical developments in human-automation interaction (Bhat et al., 2022; Yang et al., 2017; Chancey et al., 2017), it does not reflect the complexity of modern AI systems that often operate in multi-class classification environments. Chapter 3 addressed this limitation by identifying and analyzing distinct error patterns that emerge when AI systems interact with human decision-makers in multi-class settings. The results demonstrate that different AI error patterns can produce substantially different effects on human performance, response time, and trust adjustment. By revealing how specific types of AI errors influence human behavior, this study contributes to a more nuanced understanding of trust dynamics in contemporary AI-assisted decision environments.

Chapter 4 examines how the consequences of errors influence trust and reliance in human-AI collaboration. Prior literature has recognized error consequence as an important factor affecting trust in automation (Hoff and Bashir, 2015). However, empirical findings on how consequence severity affects reliance on automation have been mixed, and relatively few studies have explored this issue in the context of AI-assisted decision making (Ezer et al., 2008; Lyons and Stokes, 2012; Satterfield et al., 2017). This chapter addressed this gap by experimentally manipulating the perceived consequences of errors and examining how users adjusted their trust, reliance, and visual attention when interacting with AI support. The results suggest that higher perceived consequences lead users to rely less on AI assistance and engage in more independent verification behaviors. Chapter 4 is especially relevant to the pharmacy domain because, although most dispensing errors do not lead to harm, some carry severe consequences, potentially resulting in injury or death depending on the dosage and the patient's level of vulnerability (Aldhwaihi et al., 2016). An important consideration is pharmacists' ability to recognize medications that carry higher clinical risks. In real-world practice, pharmacists are generally aware that certain drugs require heightened vigilance. These are known as high-alert medications, defined as medications that bear a heightened risk of causing significant patient harm when used in error (Engels and Ciarkowski, 2015; Cohen et al., 2007). This domain knowledge allows pharmacists to implicitly distinguish between higher- and lower-risk dispensing situations even without explicit visual cues. In Chapter 4, consequence levels were communicated through simplified visual indicators to ensure clear manipulation. However, in real-world environments, pharmacists rely on their professional training and experience to identify situations that warrant additional verification and caution. Understanding how pharmacists adjust their trust and reliance on AI systems across medications with varying levels of clinical risk remains an important direction for

future research.

The stages of automation (SoA) framework have provided theoretical foundations for understanding how automation supports human information processing (Parasuraman et al., 2000; Wickens et al., 2010). Chapter 5 contributes to the literature on SoA by examining how trust and user behavior vary across Stage-2 and Stage-3 systems. Prior studies suggest that although Stage-3 systems may offer performance advantages over Stage-2 systems when both systems operate accurately, the cost of automation failure is typically greater for Stage-3 systems (Crocoll and Coury, 1990; Rovira et al., 2002; Sarter and Schroeder, 2001; Sethumadhavan, 2009; Rovira et al., 2007). These studies generally classify automation accuracy in binary terms (correct or incorrect) using the signal detection theory (SDT) framework (Tanner Jr and Swets, 1954; Sorkin and Woods, 1985; Hautus et al., 2021). However, when automation performance is considered within a multi-class framework, unintended safeguards can occur in which an incorrect classification at the Stage-2 level results in a correct recommendation at the Stage-3 level. Building on this perspective, Chapter 5 demonstrated through a mathematical proposition that Stage-3 systems can achieve AI reliability that is greater than or equal to that of Stage-2 systems, depending on the presence of these unintended safeguards. The empirical results further showed that this increased reliability can lead to greater performance and trust benefits for Stage-3 systems. These findings suggest that trust adjustment is shaped not only by the underlying accuracy of AI classifications but also by the practical utility of the AI's recommendation. Although both outcomes may stem from incorrect underlying classifications, they differ in their utility for supporting safe human-AI performance (Kim et al., 2025b). It is important to note, however, that the performance benefits of Stage-3 systems arise specifically when unintended safeguards are present. While this structural buffer can enhance the operational accuracy of decision recommendations, it may also contribute to the perceived "black-box" nature of AI systems (Von Eschenbach, 2021). In such cases, the correctness of the Stage-3 recommendation may appear "deceptive," as it stems from Stage-2 errors. Evidence suggests that when both Stage-2 and Stage-3 information are available, human operators tend to prioritize the correctness of Stage-3 recommendations even when the underlying Stage-2 classification is incorrect (Gegoff et al., 2025). Future research should therefore explicitly investigate differences among Stage-2, Stage-3, and integrated Stage-2/3 configurations across various domains. By examining how unintended safeguards influence performance and trust across these configurations, researchers can better understand the roles of transparency and explainability in AI systems where model-level errors do not always lead to incorrect decision recommendations.

6.1.2 Intellectual Merit

This dissertation makes several key theoretical contributions to the literature on human-AI interaction and trust in automation. Most notably, it challenges the predominant reliance on binary Signal Detection Theory frameworks to capture automation performance by introducing a beyond-binary, multi-class decision framework that more accurately reflects the complexities of modern AI-assisted decision-making. This framework reveals novel error patterns and trust dynamics that are unnoticed under traditional binary performance classifications, thereby extending foundational theories of trust dynamics (Chapter 3). By integrating severity of error consequences (Chapter 4) and stages of automation (Chapter 5) into this multi-class framework, this thesis advances a more nuanced and dynamic understanding of trust in complex decision environments. Collectively, these contributions provide a unifying theoretical foundation for studying trust and performance in AI systems whose predictions, errors, and impacts are inherently non-binary.

6.1.3 Practical Impact

The findings offer actionable insights for the design and deployment of AI systems in high-risk domains.

Chapter 2 provides the first empirical evidence of pharmacists' trust in automated pill verification technology, demonstrating that transparent presentation of AI uncertainty, designed through focus group studies and grounded in human factors principles, can meaningfully enhance trust in healthcare settings. By incorporating user feedback into the design process and applying established cognitive systems engineering approaches, this work highlights the importance of user-centered AI design that integrates transparent uncertainty communication into clinical AI tools to support safety.

Chapter 3 demonstrates that varying AI error types have distinct effects on human trust and performance, revealing that some errors can mislead operators through false reassurance, while others can paradoxically improve safety by prompting verification. As multi-class classification scenarios become increasingly prevalent in advanced AI systems across domains such as healthcare, autonomous driving, e-commerce, and luggage screening, these findings highlight the importance of examining human-AI performance through the multi-class framework to capture the nuanced differences in trust and performance.

Chapter 4 demonstrates that users adapt their reliance on AI in response to consequence severity, becoming more vigilant in high-risk situations. This suggests that AI systems capable of identifying and signaling high-consequence contexts may help mitigate errors by encouraging appropriate human oversight.

Finally, Chapter 5 informs the selection of appropriate stages of automation by showing that higher stages can yield superior interaction-level performance when evaluated through the multi-class framework, including the presence of unintended safeguards. In safety-critical domains, careful consideration of automation stages is essential to balance performance benefits with potential risks and ensure safe and effective human-AI collaboration.

Together, these findings can guide the development of AI systems that are safer, more transparent, and better aligned with human decision strategies in safety-critical domains, enabling the operators to make more informed and accurate decisions.

6.2 Future Work

As with all research, this dissertation has limitations that highlight avenues for further refinement and future investigation. Addressing these limitations and investigating the identified future research directions will strengthen the validity, generalizability, and practical relevance of the findings. Such efforts will be essential to ensure the safe, effective, and responsible deployment of AI systems in real-world settings, especially within safety-critical domains.

6.2.1 Considerations for Pharmacy Technology Implementation

Chapter 2 represents one of the first exploratory efforts to examine trust in automated pill verification technology. While the findings indicate that pharmacists generally exhibit a high disposition toward trusting automation and that transparent AI uncertainty presentation can further enhance trust, successful real-world integration depends on addressing several key variables.

First, the study did not explicitly vary the clinical risk associated with each medication. In practice, medications differ substantially in their potential to cause patient harm, and future work should examine how trust and reliance change when pharmacists interact with AI systems across medications with varying levels of clinical risk.

Second, the experimental design provided immediate feedback on AI performance following pharmacists' accept-reject decisions. In real-world pharmacy workflows, the consequences of dispensing errors often become apparent only after a delay, sometimes after the medication has already been collected by the patient. Capturing this delayed-feedback dynamic in controlled experimental settings remains a challenge, and future research should develop methods to better capture how delayed error discovery influences trust, learning, and reliance over time.

Third, the role of worker training was not explicitly examined in this study. Pharmacists'

training with automated pill verification systems may critically shape their understanding of system capabilities, limitations, and appropriate use. Future research should investigate how different training strategies interact with transparency and uncertainty presentation to support safe and effective adoption of AI-assisted medication dispensing systems.

6.2.2 Extensions of the Proposed Theoretical Frameworks

Chapters 3, 4, and 5, introduced theoretical frameworks to examine how multi-class decision scenarios, error consequences, and varying stages of AI and automation influence human trust and performance in a generalizable experimental context. While these abstractions enable systematic investigation across general domains, several limitations highlight important directions for future research.

First, future work should examine these dynamics within domain-specific tasks. Multi-class decision scenarios are particularly relevant in domains where AI systems generate multiple possible classifications that are subsequently reduced to lower-dimensional decision recommendations for human operators. Examples include TSA luggage screening, where multiple threat categories inform a binary clear-or-search decision; clinical decision-making based on medical image classification (e.g., X-rays or radiographs); autonomous vehicle object detection, where classification results may inform route adjustments versus proceeding; and e-commerce systems that classify whether packaged items are eligible for shipment. Investigating trust and performance within these domain-specific contexts would enhance ecological validity and support more precise guidance for system design. Moreover, extending this work to examine varying stages of automation within each domain could further clarify which forms of AI assistance are most appropriate given domain-specific risk profiles, task demands, and operational constraints.

Second, this dissertation primarily examined trust from the perspective of the provider, individuals using AI systems to perform tasks on behalf of others, such as pharmacists dispensing medication for patients. As AI systems become increasingly embedded in decision-making processes, it is equally important to consider the perspectives of receivers, including patients, customers, or the general public. For instance, if a dispensing error occurs due to an AI failure that a pharmacist was unable to detect, this may fundamentally alter patients' acceptance of AI-assisted healthcare and their trust in both the technology and the provider. Questions regarding responsibility, accountability, and transparency in such cases remain unresolved and requires further investigation. Future research should explore how outcomes influenced by AI errors are perceived by receivers and how these perceptions shape trust, acceptance, and ethical considerations across domains.

Third, short-term, controlled experimental interactions limit insight into how trust in AI systems evolves over extended periods of real-world use. Trust is dynamic, and changes as users accumulate experience with the system. Future research should therefore incorporate longitudinal designs to examine how trust and reliance develop over time, including how users adapt to recurring AI errors. Longitudinal studies could also capture learning effects or disuse behaviors that may not emerge in short-term experiments, thereby providing a more comprehensive understanding of sustained human-AI interaction and long-term system safety.

APPENDIX A

Questionnaires

A.1 Trust Propensity Scale Items

The Merritt et al. (2013) trust propensity scale items were presented on a 5-point Likert scale, ranging from "strongly disagree" to "strongly agree."

- I usually trust machines until there is a reason not to.
- For the most part, I distrust machines.
- In general, I would rely on a machine to assist me.
- My tendency to trust machines is high.
- It is easy for me to trust machines to do their job.
- I am likely to trust a machine even when I have little knowledge about it.

A.2 Post-Experiment Survey Items

The Muir and Moray (1996) post-experiment survey items were presented on a 5-point Likert scale, ranging from "extremely high" to "not at all."

- To what extent does the AI perform its function properly?
- To what extent can the AI's behavior be predicted from moment to moment?
- To what extent can you count on the AI to do its job?
- To what extent does the AI perform the task it was designed to do in the system?
- To what extent does the AI respond similarly to similar circumstances at different points in time?

- Your degree of faith that the AI will be able to cope with other system states in the future.
- Your degree of trust in the AI to respond accurately.
- Your degree of trust in the AI's display.
- Your overall degree of trust in the AI.

A.3 Automation Evaluation Interview Items

Study team members described 6 concepts of automation evaluation, and provided an example scenario for each concept McDermott et al. (2018). After each description and example, the participants were asked to provide their thoughts on how the concept relates to the medication verification AI system.

- Observability

    1. Did you notice any features that told you how the system was working or making decisions?

    2. Did you notice any features that told you how the system was working or making decisions?

- Predictability

    1. Were there any features that conveyed information about the system's intentions or goals?

    2. What are the aspects of the model that need greater predictability?

- Directing Attention

    1. In what ways did the system appropriately direct your attention?

    2. What could the system do differently to direct your attention more efficiently?

- Exploring Solution Space

    1. How did the system offer alternative suggestions?

    2. How did the system help you problem solve?

    3. How could the system be improved to help you problem solve?

- Adaptability
    1. How could this system be more adaptable?
    2. Are there additional features that would help convey the system's adaptability?
- Calibrated Trust
    1. What features of the system help you determine how much and when to trust it?
    2. How could this system be modified to improve your trust in it?

APPENDIX B

Testbed Supporting Figures

B.1 Study 1 Trust Rating

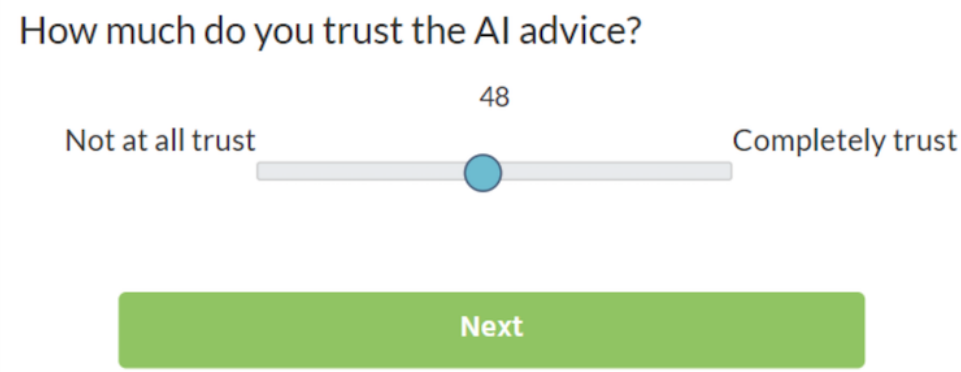


Figure B.1: After each trial i, participants reported their $trust(i)$ in the AI.

B.1.1 MRT Confidence Rating

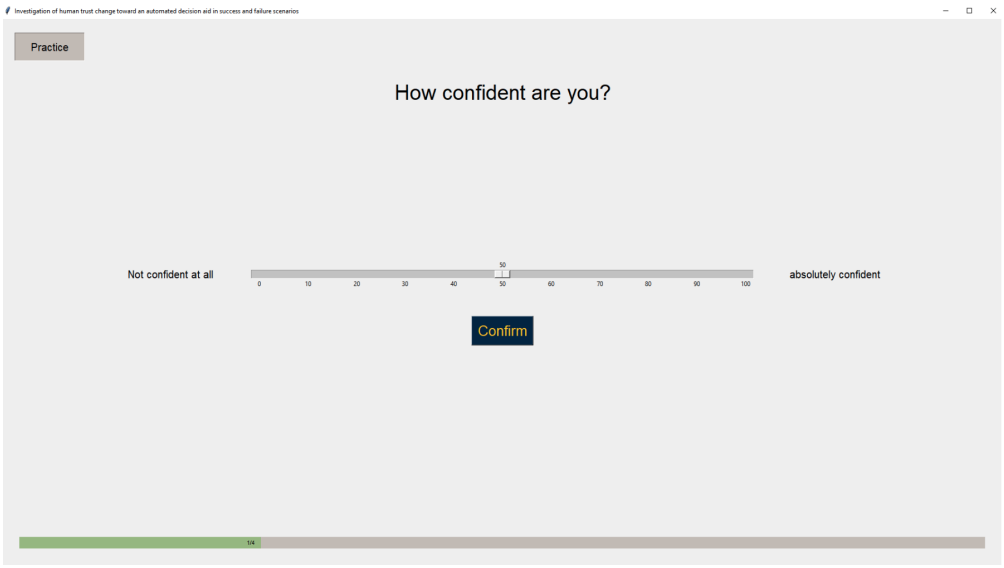


Figure B.2: Participants rated their confidence using a visual analog scale.

B.1.2 MRT Trust Rating

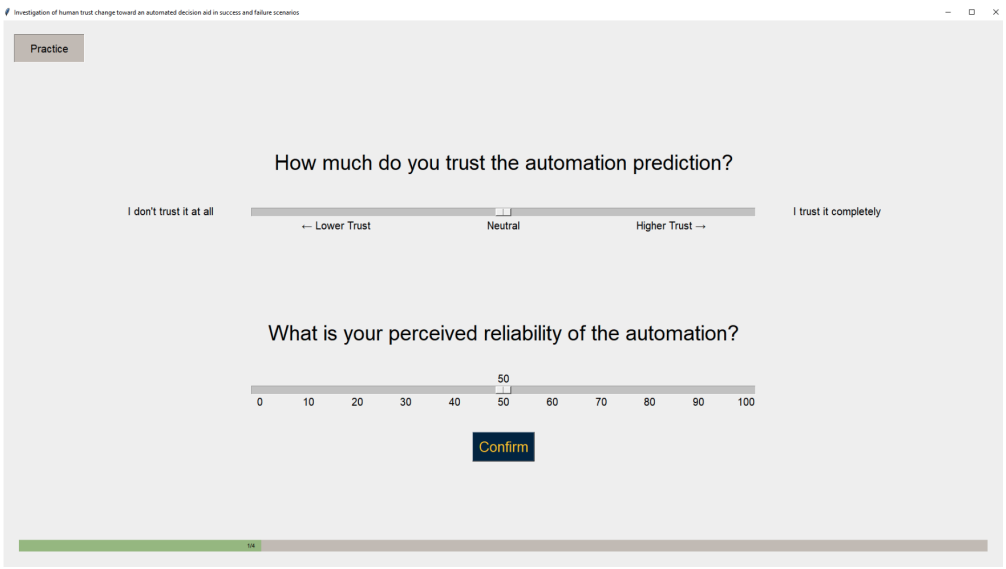


Figure B.3: Participants rated their trust and perceived reliability using a visual analog scale.

BIBLIOGRAPHY

Albayram, Y., Jensen, T., Khan, M. M. H., Fahim, M. A. A., Buck, R., and Coman, E. (2020). Investigating the effects of (empty) promises on human-automation interaction and trust repair. In *Proceedings of the 8th International Conference on Human-Agent Interaction*, pages 6–14.

Aldhwaihi, K., Schifano, F., Pezzolesi, C., and Umaru, N. (2016). A systematic review of the nature of dispensing errors in hospital pharmacies. *Integrated Pharmacy Research and Practice*, pages 1–10.

Asan, O., Bayrak, A. E., and Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. *Journal of Medical Internet Research*, 22(6):e15154.

Ashktorab, Z., Jain, M., Liao, Q. V., and Weisz, J. D. (2019). Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12.

Ayoub, J., Avetisian, L., Yang, X., and Zhou, F. (2023). Real-time trust prediction in conditionally automated driving using physiological measures. *IEEE Transactions on Intelligent Transportation Systems*.

Azevedo-Sa, H., Jayaraman, S. K., Esterwood, C. T., Yang, X. J., Robert Jr, L. P., and Tilbury, D. M. (2020). Comparing the effects of false alarms and misses on humans' trust in (semi) autonomous vehicles. In *Companion of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 113–115.

Baker, A. L., Phillips, E. K., Ullman, D., and Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(4):1–30.

Balchaites, A., McCarthy, S., and Fleming, A. (2022). Exploring the human factors of medication errors in community pharmacy: a mixed methods study. *International Journal of Pharmacy Practice*, 30(Supplement_1):i39–i40.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bates, D. W. (2000). Using information technology to reduce rates of medication errors in hospitals. *BMJ*, 320(7237):788–791.

Bhat, S., Lyons, J. B., Shi, C., and Yang, X. J. (2022). Clustering Trust Dynamics in a Human-Robot Sequential Decision-Making Task. *IEEE Robotics and Automation Letters*, 7(4):8815–8822.

Bhat, S., Lyons, J. B., Shi, C., and Yang, X. J. (2024a). Evaluating the impact of personalized value alignment in human-robot interaction: Insights into trust and team performance outcomes. In *Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction*, pages 32–41.

Bhat, S., Lyons, J. B., Shi, C., and Yang, X. J. (2024b). Evaluating the Impact of Personalized Value Alignment in Human-Robot Interaction: Insights into Trust and Team Performance Outcomes. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, pages 32–41, New York, NY, USA. Association for Computing Machinery.

Breznitz, S. (2013). *Cry wolf: The psychology of false alarms*. Psychology Press.

Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Buchanan, T., Barker, K., Gibson, J., Jiang, B., and Pearson, R. (1991). Illumination and errors in dispensing. *American journal of hospital pharmacy*, 48(10):2137–2145.

Caban, J., Rosebrock, A., and Yoo, T. (2012). Automatic identification of prescription drugs using shape distribution models. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1005–1008. IEEE.

Carayon, P. and Gurses, A. (2008). Nursing workload and patient safety—a human factors engineering perspective. In *Patient safety and quality: An evidence-based handbook for nurses*. Agency for Healthcare Research and Quality (US).

Chancey, E. T., Bliss, J. P., Liechty, M., and Proaps, A. B. (2015). False alarms vs. misses: Subjective trust as a mediator between reliability and alarm reaction measures. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 647–651. SAGE Publications Sage CA: Los Angeles, CA.

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.

Chatzimparmpas, A., Martins, R. M., Jusufi, I., and Kerren, A. (2020a). A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233.

Chatzimparmpas, A., Martins, R. M., Jusufi, I., Kucher, K., Rossi, F., and Kerren, A. (2020b). The state of the art in enhancing trust in machine learning models with the use of visualizations. *Computer Graphics Forum*, 39(3):713–756.

Cheikh, A. B., Gorincour, G., Nivet, H., May, J., Seux, M., Calame, P., Thomson, V., Delabrousse, E., and Crombé, A. (2022). How artificial intelligence improves radiological interpretation in suspected pulmonary embolism. *European radiology*, 32(9):5831–5842.

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., and Srinivasa, S. (2018). Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*, pages 307–315.

Chen, Q., Al Kontar, R., Nouiehed, M., Yang, X. J., and Lester, C. (2024). Rethinking Cost-Sensitive Classification in Deep Learning via Adversarial Data Augmentation. *INFORMS Journal on Data Science*.

Christensen, R. H. B. (2018). *Cumulative Link Models for Ordinal Regression with the R Package ordinal*. CRAN. R package vignette.

Christensen, R. H. B. (2019). ordinal—regression models for ordinal data. *R package version*, 10(2019):54.

Chui, M. A., Look, K. A., and Mott, D. A. (2014). The association of subjective workload dimensions on quality of care and pharmacist quality of work life. *Research in Social and Administrative Pharmacy*, 10(2):328–340.

Chui, M. A. and Mott, D. A. (2012). Community pharmacists' subjective workload and perceived task performance: a human factors approach. *Journal of the American Pharmacists Association*, 52(6):e153–e160.

Chung, H. and Yang, X. J. (2024). Associations between trust dynamics and personal characteristics. In *2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS)*, pages 1–6.

Cohen, M. R., Smetzer, J. L., Tuohy, N. R., and Kilo, C. M. (2007). High-alert medications: safeguarding against errors. *Medication Errors. 2nd ed. Washington (DC): American Pharmaceutical Association*, pages 317–411.

Crocoll, W. M. and Coury, B. G. (1990). Status or recommendation: Selecting the type of information for decision aiding. In *Proceedings of the human factors society annual meeting*, volume 34, pages 1524–1528. SAGE Publications Sage CA: Los Angeles, CA.

Dasgupta, A., Lee, J.-Y., Wilson, R., Lafrance, R. A., Cramer, N., Cook, K., and Payne, S. (2016). Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE transactions on visualization and computer graphics*, 23(1):271–280.

Davenport, R. B. and Bustamante, E. A. (2010). Effects of false-alarm vs. miss-prone automation and likelihood alarm technology on trust, reliance, and compliance in a miss-prone task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 54, pages 1513–1517. SAGE Publications Sage CA: Los Angeles, CA.

De Visser, E. J., Pak, R., and Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, 61(10):1409–1427.

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.

Debono, D. S., Greenfield, D., Travaglia, J. F., et al. (2013). Nurses' workarounds in acute healthcare settings: a scoping review. *BMC health services research*, 13:1–16.

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

Du, N., Huang, K. Y., and Yang, X. J. (2020). Not All Information Is Equal: Effects of Disclosing Different Types of Likelihood Information on Trust, Compliance and Reliance, and Task Performance in Human-Automation Teaming. *Human Factors*, 62(6):987–1001.

Engels, M. J. and Ciarkowski, S. L. (2015). Nursing, pharmacy, and prescriber knowledge and perceptions of high-alert medications in a large, academic medical hospital. *Hospital pharmacy*, 50(4):287–300.

Esterwood, C. and Robert, L. P. (2022). A literature review of trust repair in hri. In *2022 31st IEEE international conference on robot and human interactive communication (ro-man)*, pages 1641–1646. IEEE.

Esterwood, C. and Robert Jr, L. P. (2023). Three strikes and you are out!: The impacts of multiple human–robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior*, 142:107658.

Ezer, N., Fisk, A. D., and Rogers, W. A. (2008). Age-related differences in reliance behavior attributable to costs within a human-decision aid system. *Human factors*, 50(6):853–863.

Fallon, C. K., Bustamante, E. A., and Bliss, J. P. (2005). Enhancing perceived situation awareness and reducing perceived workload with a likelihood alarm display. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 49, pages 1677–1681. SAGE Publications Sage CA: Los Angeles, CA.

Fernandes, M., Walls, L., Munson, S., Hullman, J., and Kay, M. (2018). Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

Field, A., Field, Z., and Miles, J. (2012). *Discovering statistics using R*. sage.

Fisher, D., Popov, I., Drucker, S., and Schraefel, M. (2012). Trust me, i'm partially right: incremental visualization lets analysts explore large datasets faster. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1673–1682.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning*, pages 1050–1059. PMLR.

Gao, H., Cheng, B., Wang, J., Li, K., Zhao, J., and Li, D. (2018). Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics*, 14(9):4224–4231.

Gegoff, I., Tatasciore, M., Bowden, V. K., and Loft, S. (2025). Deciphering automation transparency: do the benefits of transparency differ based on whether decision recommendations are provided? *Human Factors*, page 00187208251318465.

Gidman, W. (2011). Increasing community pharmacy workloads in england: causes and consequences. *International journal of clinical pharmacy*, 33:512–520.

Gombolay, M., Yang, X., Hayes, B., et al. (2018). Robotic assistance in the coordination of patient care. *The International Journal of Robotics Research*, 37(10):1300–1316.

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.

Guzman-Bonilla, E. and Patton, C. E. (2024). Misses or false alarms? how error type affects confidence in automation reliability estimates. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, page 10711813241275093. SAGE Publications Sage CA: Los Angeles, CA.

Hamilton, R. (2004). Pharmacy pill counting vision system.

Hart, J. and Valasek, J. (2009). Methodology for prototyping increased levels of automation for spacecraft rendezvous functions. In *AIAA Infotech@ Aerospace 2007 Conference and Exhibit*, page 2797.

Hautus, M. J., Macmillan, N. A., and Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

Johnson, J. D., Sanchez, J., Fisk, A. D., and Rogers, W. A. (2004). Type of automation failure: The effects on trust and reliance in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 48, pages 2163–2167. SAGE Publications Sage CA: Los Angeles, CA.

Joslyn, S. and LeClerc, J. (2013). Decisions with uncertainty: The glass half full. *Current directions in psychological science*, 22(4):308–315.

Jung, M. F., Sirkin, D., Gür, T. M., and Steinert, M. (2015). Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2201–2210.

Kaplan, A. D., Kessler, T. T., Brill, J. C., and Hancock, P. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2):337–359.

Kay, M., Kola, T., Hullman, J. R., and Munson, S. A. (2016). When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 chi conference on human factors in computing systems*, pages 5092–5103.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, pages 983–997.

Kim, J., Chen, S. T., Lester, C., and Yang, X. J. (2023). Exploring trust and performance in human-automation interaction: Novel perspectives on incorrect reassurances from imperfect automation. AHFE International Conference.

Kim, J. Y., Kautz, C., and Yang, X. J. (2025a). Evaluating the effects of consequences of error on trust in ai-aided task: An eye tracking study. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 69, pages 207–213. SAGE Publications Sage CA: Los Angeles, CA.

Kim, J. Y., Lester, C., and Yang, X. J. (2025b). Beyond binary decisions: Evaluating the effects of ai error type on trust and performance in ai-assisted tasks. *Human Factors*, page 00187208251326795.

Kim, J. Y., Richie, O., and Yang, X. J. (2024). Evaluating the influence of incorrect reassurances on trust in imperfect automated decision aid. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 81–87. SAGE Publications Sage CA: Los Angeles, CA.

Kim, P. H., Ferrin, D. L., Cooper, C. D., and Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of applied psychology*, 89(1):104.

Kim, T. and Song, H. (2021). How should intelligent agents apologize to restore trust? interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, 61:101595.

Kohn, S. C., Momen, A., Wiese, E., Lee, Y.-C., and Shaw, T. H. (2019). The consequences of purposefulness and human-likeness on trust repair attempts made by self-driving vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 222–226. SAGE Publications Sage CA: Los Angeles, CA.

Kok, B. C. and Soh, H. (2020). Trust in robots: Challenges and opportunities. *Current Robotics Reports*, 1(4):297–309.

Koppel, R., Wetterneck, T., Telles, J. L., and Karsh, B.-T. (2008). Workarounds to barcode medication administration systems: their occurrences, causes, and threats to patient safety. *Journal of the American Medical Informatics Association*, 15(4):408–423.

Kozareva, Z. (2015). Everyone likes shopping! multi-class product categorization for e-commerce. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1329–1333.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13).

Lacson, F. C., Wiegmann, D. A., and Madhavan, P. (2005). Effects of attribute and goal framing on automation reliance and compliance. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 49, pages 482–486. SAGE Publications Sage CA: Los Angeles, CA.

Larios Delgado, N., Usuyama, N., Hall, A., et al. (2019). Fast and accurate medication identification. *NPJ digital medicine*, 2(1):10.

Larson, H. and Stirling, L. (2025). Formalizing motion plan legibility using empirical manual takeover data in autonomous spacecraft docking. *IEEE Transactions on Human-Machine Systems*.

Lea, V., Corlett, S., and Rodgers, R. (2012). Workload and its impact on community pharmacists' job satisfaction and stress: a review of the literature. *International Journal of Pharmacy Practice*, 20(4):259–271.

Lee, J. and Kolodge, K. (2020). Exploring trust in self-driving vehicles through text analysis. *Human factors*, 62(2):260–277.

Lee, J. and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270.

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

Lester, C., Li, J., Ding, Y., Rowell, B., Yang, J., and Kontar, R. (2021). Performance evaluation of a prescription medication image classification model: an observational cohort. *NPJ Digital Medicine*, 4(1):118.

Lester, C. A., Flynn, A. J., Marshall, V. D., Rochowiak, S., Rowell, B., and Bagian, J. P. (2022). Comparing the variability of ingredient, strength, and dose form information from electronic prescriptions with rxnorm drug product descriptions. *Journal of the American Medical Informatics Association*, 29(9):1471–1479.

Lewicki, R. J. and Brinsfield, C. (2017). Trust repair. *Annual review of organizational psychology and organizational behavior*, 4(1):287–313.

Li, H., Wickens, C. D., Sarter, N., and Sebok, A. (2014). Stages and levels of automation in support of space teleoperations. *Human factors*, 56(6):1050–1061.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in r. *Behavior research methods*, 49:1494–1502.

Luo, R., Du, N., and Yang, X. (2022). Evaluating effects of enhanced autonomy transparency on trust, dependence, and human-autonomy team performance over time. *International Journal of Human–Computer Interaction*, 38(18-20):1962–1971.

Lyons, J. B., Sadler, G. G., Koltai, K., et al. (2017). Shaping trust through transparent design: theoretical and experimental guidelines. In *Trust in Cognitive-Emotional Information Processing and Management*, pages 127–136. Springer.

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

M. Faas, S., Kraus, J., Schoenhals, A., and Baumann, M. (2021). Calibrating pedestrians' trust in automated vehicles: does an intent display in an external hmi support trust calibration and safe crossing behavior? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Madhavan, P., Wiegmann, D. A., and Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human factors*, 48(2):241–256.

Mahmood, A., Fung, J. W., Won, I., and Huang, C.-M. (2022). Owning mistakes sincerely: Strategies for mitigating AI errors. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–11.

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.

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.

McDermott, P., Dominguez, C., Kasdaglis, N., Ryan, M., Trhan, I., and Nelson, A. (2018). Human-machine teaming systems engineering guide. Technical report, MITRE Corp Bedford MA.

McNeish, D. M. and Stapleton, L. M. (2016). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, 28:295–314.

Mercado, J. E., Rupp, M. A., Chen, J. Y., Barnes, M. J., Barber, D., and Procci, K. (2016). Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human factors*, 58(3):401–415.

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.

Miller, L., Kraus, J., Babel, F., and Baumann, M. (2021). More than a feeling—interrelation of trust layers in human-robot interaction and the role of user dispositions and state anxiety. *Frontiers in psychology*, 12:592711.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

Montag, C., Kraus, J., Baumann, M., and Rozgonjuk, D. (2023). The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. *Computers in Human Behavior Reports*, 11:100315.

Muir, B. M. and Moray, N. (1996). Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460.

Natarajan, M. and Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pages 33–42.

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.

Oldland, A. R., Golightly, L. K., May, S. K., Barber, G. R., and Stolpman, N. M. (2015). Electronic inventory systems and barcode technology: impact on pharmacy technical accuracy and error liability. *Hospital pharmacy*, 50(1):34–41.

Onnasch, L. and Hösterey, S. (2019). Stages of decision automation: Impact on operators' role, awareness and monitoring. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 282–286. SAGE Publications Sage CA: Los Angeles, CA.

Onnasch, L., Wickens, C. D., Li, H., and Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human factors*, 56(3):476–488.

Paoletti, R., Suess, T., Lesko, M., et al. (2007). Using bar-code technology and medication observation methodology for safer medication administration. *American journal of health-system pharmacy*, 64(5):536–543.

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.

Perkins, L., Miller, J. E., Hashemi, A., and Burns, G. (2010). Designing for human-centered systems: Situational risk as a factor of trust in automation. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 54, pages 2130–2134. SAGE Publications Sage CA: Los Angeles, CA.

Peterson, G., Wu, M., and Bergin, J. (1999). Pharmacists' attitudes towards dispensing errors: their causes and prevention. *Journal of clinical pharmacy and therapeutics*, 24(1):57–71.

Picco, A., Stuiver, A., de Winter, J., and de Waard, D. (2025). Information, assessment, or decision: a driving simulator study on the effect of real-time feedback based on information-processing stages. *Ergonomics*, 68(8):1165–1180.

Poon, E. G., Cina, J. L., Churchill, W., et al. (2006). Medication dispensing errors and potential adverse drug events before and after implementing bar code technology in the pharmacy. *Annals of internal medicine*, 145(6):426–434.

Poon, E. G., Keohane, C. A., Yoon, C. S., et al. (2010). Effect of bar-code technology on the safety of medication administration. *New England Journal of Medicine*, 362(18):1698–1707.

Popovich Jr, J. and Jordan, M. (2011). Automated drug discrimination during dispensing.

R Core Team (2013). R: A language and environment for statistical computing. r foundation for statistical computing.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, M. M., Desai, B. C., and Bhattacharya, P. (2008). Medical image retrieval with probabilistic multi-class support vector machine classifiers and adaptive similarity fusion. *Computerized Medical Imaging and Graphics*, 32(2):95–108.

Rashid, A. B., Kausik, A. K., Al Hassan Sunny, A., and Bappy, M. H. (2023). Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. *International journal of intelligent systems*, 2023(1):8676366.

Rastogi, C., Tulio Ribeiro, M., King, N., Nori, H., and Amershi, S. (2023). Supporting human-ai collaboration in auditing llms with llms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 913–926.

Reilley, S., Grasha, A., Matthews, G., and Schafer, J. (2003). Automatic-controlled information processing and error detection in a simulated pharmacy-verification task. *Perceptual and motor skills*, 97(1):151–174.

Reiner, G., Pierce, S., and Flynn, J. (2020). Wrong drug and wrong dose dispensing errors identified in pharmacist professional liability claims. *Journal of the American Pharmacists Association*, 60(5):e50–e56.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., and Alahmad, R. (2020). Designing fair ai for managing employees in organizations: a review, critique, and design agenda. *Human–Computer Interaction*, 35(5-6):545–575.

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.

Rovira, E., Zinni, M., and Parasuraman, R. (2002). Effects of information and decision automation on multi-task performance. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 46, pages 327–331. SAGE Publications Sage CA: Los Angeles, CA.

Rudin, C. and Ustun, B. (2018). Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5):449–466.

Sacha, D., Sedlmair, M., Zhang, L., Lee, J., Weiskopf, D., North, S., and Keim, D. (2016). Human-centered machine learning through interactive visualization. In *24th European Symposium on Artificial Neural Networks*, pages 641–646.

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. and Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human factors*, 43(4):573–583.

Satterfield, K., Baldwin, C., De Visser, E., and Shaw, T. (2017). The influence of risky conditions in trust in autonomous systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 324–328. SAGE Publications Sage CA: Los Angeles, CA.

Schaefer, K., Chen, J., Szalma, J., 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.

Sebo, S. S., Krishnamurthi, P., and Scassellati, B. (2019). "i don't believe you": Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 57–65. IEEE.

Sebok, A. and Wickens, C. D. (2017). Implementing lumberjacks and black swans into model-based tools to support human–automation interaction. *Human factors*, 59(2):189–203.

Seong, Y. and Bisantz, A. M. (2008). The impact of cognitive feedback on judgment performance and trust with decision aids. *International Journal of Industrial Ergonomics*, 38(7-8):608–625.

Sethumadhavan, A. (2009). Effects of automation types on air traffic controller situation awareness and performance. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 53, pages 1–5. Sage Publications Sage CA: Los Angeles, CA.

Shepard, R. N. and Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703.

Sheridan, T. B. (2008). Risk, Human Error, and System Resilience: Fundamental Ideas. 50(3):418–426. Issue: 3 Pages: 418-426.

Sheridan, T. B., Verplank, W. L., and Brooks, T. (1978). Human/computer control of undersea teleoperators. In *NASA. Ames Res. Center The 14th Ann. Conf. on Manual Control*.

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.

Slovic, P. and Lichtenstein, S. (1968). Relative importance of probabilities and payoffs in risk taking. *Journal of Experimental Psychology*, 78(3, Pt.2):1–18.

Soh, H., Xie, Y., Chen, M., and Hsu, D. (2020). Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research*, 39(2-3):233–249.

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.

Souissi, S. B., Abed, M., El Hiki, L., Fortemps, P., and Pirlot, M. (2019). Pars, a system combining semantic technologies with multiple criteria decision aiding for supporting antibiotic prescriptions. *Journal of Biomedical Informatics*, 99:103304.

Stanton, N., Ragsdale, S., and Bustamante, E. (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.

Sundaram, S. and Zeid, A. (2023). Artificial intelligence-based smart quality inspection for manufacturing. *Micromachines*, 14(3):570.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist*, 47(4):522.

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

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

Tatasciore, M., Bowden, V. K., Visser, T. A., Michailovs, S. I., and Loft, S. (2020). The benefits and costs of low and high degree of automation. *Human factors*, 62(6):874–896.

Tsai, C.-C., Kim, J. Y., Chen, Q., Rowell, B., Yang, X. J., Kontar, R., Whitaker, M., and Lester, C. (2025). Effect of artificial intelligence helpfulness and uncertainty on cognitive interactions with pharmacists: Randomized controlled trial. *Journal of Medical Internet Research*, 27:e59946.

Vandenberg, S. G. and Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and motor skills*, 47(2):599–604.

Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy & technology*, 34(4):1607–1622.

Walter, S., Schmutz, J. B., and Grote, G. (2024). A safety approach for improving security-effective coordination strategies at the airport security screening checkpoint. *Human factors*, 66(4):1302–1314.

Wang, E., Gross, C., Tilburt, J., James, B., Nguyen, P., Smaldone, M., et al. (2015). Shared decision making and use of decision aids for localized prostate cancer: perceptions from radiation oncologists and urologists. *JAMA internal medicine*, 175(5):792–799.

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, S., Liu, Y., Wang, X., Liu, Z., You, X., and Li, Y. (2025). Where is the function allocation boundary? the effect of degree of automation on attention allocation and human performance under different reliabilities. *Human Factors*, 67(8):757–775.

Watterson, T., Look, K., Steege, L., and Chui, M. (2021). Operationalizing occupational fatigue in pharmacists: An exploratory factor analysis. *Research in Social and Administrative Pharmacy*, 17(7):1282–1287.

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., Gordon, S. E., Liu, Y., and Lee, J. (2004). *An introduction to human factors engineering*, volume 2. Pearson Prentice Hall Upper Saddle River, NJ.

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

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., and Sarter, N. B. (2010). Stages and levels of automation: An integrated meta-analysis. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 54, pages 389–393. Sage Publications Sage CA: Los Angeles, CA.

Wickens, C. D., Self, B. P., Andre, T. S., Reynolds, T. J., and Small, R. L. (2007). Unusual attitude recoveries with a spatial disorientation icon. *The International Journal of Aviation Psychology*, 17(2):153–165.

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.

Wiegmann, D. A., Rich, A., and Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, 2(4):352–367.

Wischnewski, M., Krämer, N., and Müller, E. (2023). Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Wong, Y., Ng, H., Leung, K., Chan, K., Chan, S., and Loy, C. (2017). Development of fine-grained pill identification algorithm using deep convolutional network. *Journal of biomedical informatics*, 74:130–136.

Wootton, J., Reznack, V., and Hobson, G. (2003). Pharmaceutical pill recognition and verification system.

Xu, A. and Dudek, G. (2015). OPTIMo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *Proceedings of the 2015 ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*, pages 221–228. ACM.

Yang, F., Huang, Z., Scholtz, J., and Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 189–201.

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., Guo, Y., and Schemanske, C. (2023a). From Trust to Trust Dynamics: Combining Empirical and Computational Approaches to Model and Predict Trust Dynamics In Human-Autonomy Interaction. In Duffy, V. G., Landry, S. J., Lee, J. D., and Stanton, N., editors, *Human-Automation Interaction: Transportation*, Automation, Collaboration, & E-Services, pages 253–265. Springer International Publishing, Cham.

Yang, X. J., Schemanske, C., and Searle, C. (2023b). 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.

Yang, X. J., Wickens, C. D., and Hölttä-Otto, K. (2016). How users adjust trust in automation: Contrast effect and hindsight bias. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 60, pages 196–200. SAGE Publications Sage CA: Los Angeles, CA.

Young, D. (2002). Veterans affairs bar-code-scanning system reduces medication errors. *American journal of health-system pharmacy*, 59(7):591–592.

Yu, J., Chen, Z., and Kamata, S.-i. (2014). Pill recognition using imprint information by two-step sampling distance sets. In *2014 22nd International Conference on Pattern Recognition*, pages 3156–3161. IEEE.

Yu, J., Chen, Z., Kamata, S.-i., and Yang, J. (2015). Accurate system for automatic pill recognition using imprint information. *IET Image Processing*, 9(12):1039–1047.

Zheng, Y., Rowell, B., Chen, Q., et al. (2023). Designing human-centered ai to prevent medication dispensing errors: Focus group study with pharmacists. *JMIR Formative Research*, 7(1):e51921.