

**FACTORS THAT AFFECT TRUST AND RELIANCE
ON AN AUTOMATED AID**

A Dissertation
Presented to
The Academic Faculty

By

Julian Sanchez

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology

Georgia Institute of Technology

May 2006

Factors That Affect Trust and Reliance
On An Automated Aid

Approved by:

Dr. Gregory Corso
School of Psychology
Georgia Institute of Technology

Dr. Jerry R. Duncan
John Deere Technology Center
Deere & Company

Dr. Ute M. Fischer
School of Literature, Communication, & Culture
Georgia Institute of Technology

Dr. Arthur D. Fisk, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Wendy A. Rogers
School of Psychology
Georgia Institute of Technology

Date Approved: 03/13/2006

Jillers, you are the reason I am

ACKNOWLEDGEMENTS

To the members of my committee, Greg, Ute, Jerry, Wendy, and Dan, thank you for your time and feedback. To my advisor, Dan, thanks for believing in my ideas, thanks for always pushing me to think about the importance and the relevance of our research, and most of all thanks for caring. To Wendy, thanks for your dedication to the lab and to your students. To all current and past members of the HFA lab that I have had the privilege to work with, thank you for always being there for me. Travis, I will always cherish our work together and use it as a measure of excellence; also, thank you for your friendship. Tim, your great personality is contagious and thanks to you I made it through some of the hard times. Anne, thanks for your support during the last 2 years, you're a great friend.

Mom, Dad, and Alex, you guys are the reason I have been able to travel this road. I have always done my best to make sure you guys are proud of me, I promise to continue to do so; thank you for being with me at all times, thank you for being the best friends a person could ever have. This project would not have been possible without your unconditional support and love.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF TABLES	viii
LIST OF FIGURES	ix
SUMMARY	xii
CHAPTER 1 Introduction.....	1
Reliability.....	2
Type of automation error	3
Does error type affect behavior?	11
Time on task.....	12
Does it matter when errors occur?	12
Age-Related Factors in Human-Automation Interaction.....	15
Human-Automation Interaction: Micro-Level Analysis of Behavior.....	18
Objectives of the Current Investigation	21
CHAPTER 2 Method (Experiment 1).....	23
Overview	23
Participants.....	23
Design	24
Materials	24
Equipment	24
Task.....	25
Collision avoidance task	26
Reliability of the collision avoidance system	30
Tracking task.....	30
Procedure	31
Dependent Variables.....	33
Collision avoidance task	33
Tracking task.....	35
Expected Results.....	35
Method (Experiment 2).....	37
Participants.....	37
Design	38
Materials	39
Equipment	39
Task.....	39
Procedure	39
Dependent variables.....	39

Expected results	40
CHAPTER 3 Results.....	42
Overview of Analyses.....	42
Summary of Results.....	44
Reliance and Trust in Automation: A Macro-Level Analysis (Experiment 2)	45
First spacebar presses.....	45
Spacebar presses during alarms and non-alarms	46
Total time for spacebar presses.....	49
Subjective trust.....	50
Perceived reliability	51
The Effects of Error and Distribution of Errors on Reliance: A Micro-Level Analysis.....	53
Throughout condition.....	55
First half condition.....	59
Second half condition	62
Age-Related Effect on Reliance: A Micro-Level Analysis	66
Throughout condition.....	67
First half condition.....	69
Second half condition	72
The Effects of Agricultural Experience on Reliance and Trust.....	74
Participant characteristics	75
Spacebar presses	75
Micro-level analysis.....	77
Detection of the Exception Error	81
Effects of First and Subsequent Failures on Reliance	82
Tracking Task Performance.....	83
Obstacle Detection Performance.....	86
Number of collision errors.....	86
False negative and positive rates.....	87
CHAPTER 4 Discussion.....	89
Does Error Type Affect Reliance on Automation?.....	89
Do humans adjust optimally to automation error type?.....	91
The Meaning of Alarms	94
Does it Matter When Errors Occur?	94
Does Age Matter in Human-Automation Interaction?.....	97
Does Domain Experience Matter in Human-Automation Interaction?	98
The Construct of Trust.....	99
Practical Applications	101
Design of automated systems.....	101
Future investigations of human-automation interaction	103
Where We Have Been, Where We Are Now, and Where We Should Go	104
Where we have been	105
Where we are now	106
Where we should go.....	108

APPENDIX A: Reliance data from Johnson (2004).....	110
APPENDIX B: Participant characteristics by Error Type (Experiment 1).....	111
APPENDIX C: Distribution of errors	112
APPENDIX D: Subjective trust and perceived reliability questionnaire	118
APPENDIX E: Participant characteristics by Age, Error Type, and Distribution of Errors (Experiment 2)	119
APPENDIX F: ANOVA table (Experiment 2).....	121
APPENDIX G: Correlations table (Experiment 2).....	124
APPENDIX H: ANOVA table (Experiment 1)	125
APPENDIX I: “First error effect”	127
APPENDIX J: Conceptual model of human-automation interaction	131
APPENDIX K: Technical terms and definitions	132
REFERENCES	133

LIST OF TABLES

Table 1. Summary of results from Experiments 1 and 2	44
Table B1. Participant characteristics by Error Type (Experiment 1).....	111
Table C1. Distribution of errors.....	112
Table E1. Participant characteristics by Age and Error Type for the Throughout Condition.....	119
Table E2. Participant characteristics by Age and Error Type for the Throughout Condition.....	119
Table E3. Participant characteristics by Age and Error Type for the Throughout Condition.....	120
Table F1. ANOVA table (Experiment 2).....	121
Table G1. Correlations table (Experiment 2).....	124
Table H1. ANOVA table (Experiment 1).....	125
Table I1. Changes in the number of participants who pressed the spacebar before and after the first automation error	130

LIST OF FIGURES

Figure 1. Illustration of two hypothetical systems with the same average reliability (70%) but different error configurations through time	13
Figure 2. The effects of experience (x-axis) on trust and reliance (y-axis). Adapted from Wickens and Xu (2002)	14
Figure 3. Screenshot of the simulator	26
Figure 4. Immediate feedback about the outcome of actions in the collision avoidance task	27
Figure 5. Non-reliance (first spacebar presses) by Error Type	45
Figure 6. Non-reliance (first spacebar presses during alarms) by Error Type	47
Figure 7. Non-reliance (first spacebar presses during non-alarms) by Error Type	48
Figure 8. Total amount of time the spacebar was pressed by Error Type and Distribution of Errors	49
Figure 9. Subjective trust ratings by Age and Distribution of Errors	51
Figure 10. Perceived reliability ratings by age and Distribution of Errors	53
Figure 11. Percentage of Younger Adults who pressed the spacebar during alarms in the Throughout Condition	56
Figure 12. Percentage of Younger Adults who pressed the spacebar during non-alarms in the Throughout Condition	57
Figure 13. Percentage of Younger Adults who pressed the spacebar during alarms in the First Half Condition	60
Figure 14. Percentage of Younger Adults who pressed the spacebar during non-alarms in the First Half Condition	61
Figure 15. Percentage of Younger Adults who pressed the spacebar during alarms in the Second Half Condition	63
Figure 16. Percentage of Younger Adults who pressed the spacebar during non-alarms in the Second Half Condition	65

Figure 17. Percentage of Older Adults who pressed the spacebar during alarms in the Throughout Condition.....	68
Figure 18. Percentage of Older Adults who pressed the spacebar during non-alarms in the Throughout Condition.....	69
Figure 19. Percentage of Older Adults who pressed the spacebar during alarms in the First Half Condition	70
Figure 20. Percentage of Older Adults who pressed the spacebar during alarms in the First Half Condition	71
Figure 21. Percentage of Older Adults who pressed the spacebar during alarms in the Second Half Condition.....	73
Figure 22. Percentage of Older Adults who pressed the spacebar during alarms in the Second Half Condition.....	74
Figure 23. Non-reliance (first spacebar presses) by agricultural experience.....	76
Figure 24. Percentage of Farmers who pressed the spacebar during alarms	78
Figure 25. Percentage of Farmers who pressed the spacebar during non-alarms	79
Figure 26. Tacking task performance (red events) by Age and Error Type	84
Figure 27. Total number of collision errors by Error Type	86
Figure 28. The relationship between variables specific to the automation, reliance, and system performance	90
Figure A1. Reliance data from Johnson (2004).....	110
Figure I1. Changes in the number of participants who pressed the spacebar before and after each error during alarms in the Throughout Condition	127
Figure I2. Changes in the number of participants who pressed the spacebar before and after each error during non-alarms in the Throughout Condition.....	128
Figure I3. Changes in the number of participants who pressed the spacebar before and after each error during alarms in the First Half Condition.....	128
Figure I4. Changes in the number of participants who pressed the spacebar before and after each error during non-alarms in the First Half Condition	129

Figure I5. Changes in the number of participants who pressed the spacebar before and after each error during alarms in the Second Half Condition	129
Figure I6. Changes in the number of participants who pressed the spacebar before and after each error during non-alarms in the Second Half Condition.....	130
Figure J1. Conceptual model of human-automation interaction.....	131

SUMMARY

Previous research efforts aimed at understanding the relationship between automation reliability and reliance on the automation have mainly focused on a single dimension of reliability, the automation's error rate (Bailey, 2004; Kantowitz, Hanowiski, & Kantowitz, 1997; Sanchez, Fisk, & Rogers, 2004; Wiegmann, Rich, & Zhang, 2001). Efforts to understand the effects of additional dimensions, such as types of errors, have merely provided suggestions about the effects that automation false alarms and misses can have on human behavior (Dixon & Wickens, 2003; 2004). Furthermore, other dimensions of reliability, such as the distribution of errors in time, have been almost completely ignored. The results of this investigation make a critical contribution to the theory of automation use as a function of reliability; specifically, the effects of error type and the distribution of errors on reliance. Another objective of this research was to gain a better understanding of the age-related factors that affect human-automation interaction. To date, it is unclear if age-related differences in automation use stem from cognitive and physical age-related declines or from an aversion by older adults to the use of new technologies (Kantowitz, Becker, & Barlow, 1993). In addition to investigating age-related differences in the use of automation, the effects of domain-experience on human-automation interaction were also explored.

A multi-task simulation of an agricultural vehicle was used in this investigation. The simulator was composed of two main tasks, a collision avoidance task and a tracking task. The collision avoidance task was supported by an "imperfect" automated collision avoidance system and the tracking task was performed manually. The data were

analyzed at both the macro- and micro-levels. The micro-level analyses provided valuable insight into the way in which humans' reliance patterns change over time. The results of this investigation indicated that there are distinct patterns of reliance that develop as a function of error type, which are dependent on the state of the automation (alarms or non-alarms). The different distributions of errors across time had an effect on the estimates of reliability and subjective trust ratings. The recency of errors was negatively related to perceived reliability and trust. As far as age-related factors that affected human-automation interaction, the current results suggest that older adults are able to adjust their behavior according to the characteristics of the automation, although it takes them longer to do so. Furthermore, consistent with previous findings (Johnson, 2004; Sanchez et al., 2004), it appears that older adults are as willing to use automated systems as younger adults, as long as they are reliable enough to reduce workload. The analysis on the effects of domain experience on automation use indicated that those with experience operating agricultural vehicles had different tendencies of reliance. Specifically, participants with experience operating agricultural vehicles were less likely to rely on automated alarms than those without experience. The results of this investigation have important implications in our understanding of how humans adjust their behavior according to the characteristics of an automated system.

CHAPTER 1

INTRODUCTION

During unexpected circumstances, where automated systems are most likely to behave unreliably, the human is still expected to act as a backup to detect failure events and act accordingly to avoid a system error. Therefore, when attempting to increase the performance of a human-automation system, increasing the reliability of the automation only makes up half of the solution; the other half is ensuring that humans perform their part of the task. This means that part of the success of a human-automation system hinges on understanding and being able to make some predictions about the behavior of the human in an automated environment.

Research on human-automation interaction can be approached by understanding the effects of three main types of variables on human behavior. These variables include those that are specific to the automation (e.g., reliability, level of automation), specific to the human (e.g., age, experience) and specific to the environment (e.g., cost of an error, workload demands). Irrespective of the types of variables that are studied, the objectives of human-automation interaction research field are to a) identify the variables that affect human behavior in automated environments; b) identify the degree to which variables affect behavior; c) understand, from a psychological perspective, why specific variables lead to specific changes in behavior; d) inform human performance theories; and e) inform the design of automated systems to improve the effectiveness of systems.

To date, there is still uncertainty surrounding the effects of a number of variables on human behavior within the human-automation system, such as automation reliability.

While numerous studies have been conducted to understand the effects of automation reliability on human behavior, there are still a number of open questions regarding the way humans accumulate evidence about the reliability of the automation and how this affects their behavior. Furthermore, the effects of variables specific to the automation (i.e., changes in reliability over time and error types) on behavior are still open questions. Gaining an understanding of how humans perceive reliability in automated environments and how this perception influences behavior can contribute to our understanding of how humans accumulate and integrate information over time. Furthermore, the effects of variables specific to the human, such as age and domain experience on the use of automation are largely unexplored.

Reliability

A common definition of automation reliability used in human-automation interaction research is “the number of correct operations divided by the total number of operations in which a task is automated” (Xu, Wickens, & Rantanen, 2004, p. 5). This definition is solely based on the percentage of correct actions made by an automated system. In general, there is a positive relationship between automation reliability (the percentage of correct actions it makes) and use of the automation; although evidence also suggests that the relationship is not linear (Lee & See, 2004; Moray, Inagaki, & Itoh, 2000; Sanchez et al., 2004). Similarly, subjective levels of trust and perceived reliability have also been found to be affected by automation reliability (Bailey, 2004; Kantowitz et al., 1997; Sanchez et al., 2004; Wiegmann et al., 2001). Therefore, there is ample

evidence that as automation reliability degrades— as it makes more errors—trust and use of the automation also decline.

Previous research studies that have focused their efforts on investigating the effects of automation reliability on behavior have failed to consider all of the components of reliability. For example, one variable that is commonly overlooked is the distribution of errors in time (i.e., when errors occur). Wickens and Xu (2002) suggested that because of the effect this variable can have on the expectations of the human, it is a critical component of the relationship between automation reliability and human behavior. Another variable that is commonly overlooked is the type of error that an automated system makes. The literature is laden with mixed findings regarding the effects that different types of errors have on human behavior and expectations. However, to date, these findings have not yielded a clear understanding of the psychological implications associated with the occurrence of different types of errors.

Type of Automation Error

The types of errors that an automated system makes depend on the Level of Automation (Endsley & Kaber, 1999). The term Levels of Automation (LOA) is used to describe the level of support that an automated system provides. Parasuraman, Sheridan and Wickens (2000) suggested that there are four levels of automation: information acquisition, information analysis, decision/action selection, and action implementation. Within the LOA of information analysis and decision/actions selection, false alarms and misses are the two types of errors the automation can generate. A false alarm occurs

when the automation generates a warning or alert in the absence of a true signal.

Conversely, during a miss, the automation fails to notify the human of a true signal.

Even though false alarms and misses are both types of errors that are generated within the same LOA, they have considerably different characteristics. By nature, false alarms are more salient because they are accompanied by the stimulus that the automation generates to produce a warning. Furthermore, depending on the operational environment, the cost of each type of error can vary considerably. Some have argued that automation misses tend to have greater consequences than automation false alarms (Bliss & Gilson, 1998), although there are ample examples of systems and situations in which an argument can be made that false alarms are most costly (e.g., a false alarm that leads to the shooting of a civilian airliner). Nevertheless, because the two types of errors are fundamentally different, one might expect that they have different effects on behavior, even when they are equated for cost.

The potentially different effect of false alarms and misses on human behavior is an issue that has received limited attention. In most studies that have investigated the effects of unreliable automation on human behavior, the distinction between false alarms and misses is seldom addressed. Rather, automation false alarms and misses are usually combined in experiments, while the effects of other variables are measured (e.g., Dzindolet, Pierce, Pomranky, Peterson & Beck, 2001; Sanchez et al., 2004; Skitka, Mosier & Burdick, 1999; 2000). While a balanced combination of false alarms and misses provides a valuable contribution to the understanding of how humans interact with systems that are predisposed to making both types of errors, the unique effects of each type of error are likely to be attenuated by the presence of the other type of error.

Other programs of research have only focused on the effects of one type of error. For example, many of the studies that focus on alarm systems typically only use false alarms (e.g., Bliss & Dunn, 2000; Bliss, Dunn & Fuller, 1995; Bliss, Gilson, & Deaton, 1995; Bliss & McAbee, 1995). Bliss and Gilson (1998) argued that investigating the effects of false alarms is critical because they tend to be more prevalent in applied contexts than misses. The most common finding from this area of research is that a high number of false alarms usually results in humans ignoring true alarms (Meyer, Bitan, Shinar & Zmora, 1999). This behavior is often referred to as the Cry-Wolf Effect (Breznitz, 1984).

While the research efforts that have exclusively investigated the effects of false alarms have contributed to the understanding of human-automation interaction, there are some important limitations with a majority of the studies in this area. The main limitation in these studies is that they seldom provide participants with the opportunity to verify the validity of alarms before “agreeing” or “disagreeing” with them. Therefore, the Cry-Wolf Effect, which means humans begin to ignore true alarms given the possibility of false alarms, is the only way that participants are able to objectively express mistrust in the automation. However, in studies in which participants are provided with the opportunity to verify the validity of the automation the Cry-Wolf Effect is seldom observed (Sanchez et al., 2004; Wiegmann et al., 2001). Instead, mistrust is usually manifested by an increase in time and effort spent accessing other sources of information. This behavior of verifying the validity of the automation by checking another source is referred to as *non-reliance* in the current investigation. It is worth noting that in a recent discussion by Meyer (2004), reliance was defined as the lack of a response by the user

during periods when the warning system is idle. However, Meyer's definition is constrained by the same limitations that bound alarm studies; it does not consider the existence of other sources of information that can be used to verify the validity of the automation (for a discussion see Sanchez, 2005).

A very limited amount of research has specifically compared the effects of false alarms and misses on human behavior and overall system performance. In general, the available evidence suggests that the prevalence of each type of error has different effects on behavior. However, the experimental manipulations and measures used to assess the effects of false alarms and misses have only yielded indirect evidence for this phenomenon. For example, Gupta, Bisantz and Singh (2001) compared the effects of false alarms and misses of an adverse condition warning system (ACWS) in a driving simulator. The threshold used to generate an auditory alarm was manipulated to either generate a high or a low number of false alarms relative to the number of misses. True alarms were meant to notify the driver of upcoming skids. It is worth noting that a false alarm in this study was defined as an alarm that sounded much sooner than necessary, while a miss was defined as an alarm that sounded, but was too late to allow the human to react in a safe and timely fashion. Therefore, from a Signal Detection Theory (SDT; Green & Sweets, 1966) perspective there were no true automation false alarms or misses, rather, false alarms were early alarms and misses were late alarms. A pure automation false alarm would have consisted of a skid warning in the absence of conditions that would have led to a skid, and a pure automation miss would have consisted of a failure by the automation to provide a warning in the presence of skidding conditions.

The results of Gupta et al. (2001) indicated that the condition with a higher false alarm (early) rate led to lower levels of subjective trust on the automation than the condition with a high miss (late alarm) rate. This finding is consistent with those who have suggested that the nuisance associated with false alarms leads to lower levels of subjective trust relative to misses (Breznitz, 1984). Furthermore, the driving behavior of participants in the Gupta et al. study was slightly affected by the type of error. In the high miss rate condition, drivers showed less steering deviation than in the high false alarm rate condition. This difference as a function of error type provides some evidence that drivers in the high miss rate condition were aware that if an alarm were to occur, it may be too late to react; therefore, they would benefit from a more conservative driving approach to prevent skidding.

In another driving study, Cotte, Meyer, and Coughlin (2001) manipulated the criterion of a collision detection system, while holding the sensitivity of the system constant. In this study, misses by the automation meant no alarm was generated in the presence of an obstacle, while false alarms occurred when the alarm became active in the absence of an obstacle. Their results indicated that driving behavior was moderated as a function of the automation's criterion. In the condition with high false alarm rates, the average driving speed was significantly faster than in the high miss rate condition. This difference in driving speed increased throughout the experimental session. These results suggest that drivers in the high false alarm rate condition became aware that while the system was likely to generate false alarms it would not miss much; therefore, they could increase their driving speed with a high degree of confidence that if an obstacle appeared the system would not miss it. Conversely, the slower driving behavior by participants in

the high miss rate condition suggests that they adopted a more conservative behavior as a function of the automation's criterion.

Because the actual and perceived costs of false alarms and misses can vary greatly, it is worth noting that neither the Gupta et al. (2001) nor the Cotte et al. (2001) studies reported a payoff structure associated with the task. The payoff structure could considerably influence the effect that each error type has on behavior. For example, if the cost of an error is high (e.g., missile detection system), even with a high false alarm rate the human will likely verify all alarms instead of responding to some and ignoring others. Conversely, if there is not a high cost associated with errors (e.g., browsing to a "non-secure" website), the human might be more likely to ignore most alarms, whether false or true. Contrary to the findings of Gupta et al. and Cotte et al., other driving simulation studies have found that false alarms of collision avoidance systems do cause drivers to slow down unnecessarily, while high miss rates do not have a significant impact on driving behavior (Maltz & Shinar, 2004).

In a study that did equate the outcome cost of false alarms and misses, Lehto, Papastavrou, Ranney, and Simmons (2000) used a decision aid that notified drivers when it was safe to pass a vehicle in front of them. Participants were rewarded \$0.20 for successfully passing and were penalized \$0.10 for unsafe passing attempts and missed passing opportunities. Their results indicated that a high false alarm rate was more detrimental to performance (measured in monetary earnings) than a system with no false alarms and a few misses. However, in this study, the criterion of the system was confounded by the sensitivity of the decision aid, which means the condition with no false alarms had a higher level of reliability than the condition with the high false alarm

rate (~90% and ~70%, respectively). This disparity in the sensitivity of the decision aid across conditions makes any conclusions about differences in behavior as a function of error type difficult to discern. Furthermore, direct measures of reliance and subjective trust were not reported.

In a recent study by Dixon and Wickens (2003), error type was manipulated in an unmanned aerial vehicle (UAV) simulation. In this multi-task environment, participants were required to detect system failures (visible in a gauges display) with the assistance of an automated alarm while performing a tracking task and a target search task. Their experiment included conditions in which the automation was 100% reliable, 67% reliable (all misses), 67% reliable (all false alarms), and a control group with no automation. In this study, the 100% reliable automation improved performance across all tasks relative to the no automation condition. For the 67% reliable conditions, detection of system failures was worse in the false alarm condition than in the miss condition, suggesting that participants experiencing all false alarms began to ignore most of the alarms therefore missing a considerable number of system failures. Detection of system failures was also worse in the false alarm condition than in the control (no automation) condition.

Contrary to the findings of Gupta et al. (2001), Dixon and Wickens (2003) found that subjective reliability ratings were higher in the false alarm condition than in the miss condition, although in both conditions the actual reliability of the automation was underestimated. Dixon and Wickens also found that in the miss condition, performance in one of the concurrent tasks (identifying targets of opportunity) was significantly lower than in the false alarm condition. This effect of error type on concurrent task performance suggests that participants in the miss condition were paying more attention

to the gauges and neglected some of the concurrent tasks. It also suggests that, because the participants in the miss condition had to monitor the gauges more often, the level of workload in this condition might have been higher. Increased workload is a possible explanation for why the miss condition was conducive of lower reliability ratings.

In an extension of the 2003 study, Dixon and Wickens (2004) used a similar task. They included a mostly misses condition (3 misses, 1 false alarm), a mostly false alarms condition (3 false alarms, 1 misses) and a balanced condition (1 miss, 1 false alarm). One of the more compelling findings in this study was that the average reaction time to respond to alarms was significantly lower in the mostly miss condition than in the mostly false alarms condition. This difference suggests that participants in the mostly misses condition reacted to alarms without verifying whether they were going to make the right choice or not, while participants in the mostly false alarms condition verified the alarm before correcting the failure. Again, these results suggest a difference in the way humans monitor an automated aid as a function of prevalence of error type.

The subjective trust ratings in the Dixon and Wickens (2004) study did not differ as a function of error type and were fairly close to the actual reliability of the automation. However, in this study participants were informed a priori about the range of reliability levels and about the criterion setting of the automation. This information, in part, may have influenced participants' behavior and subjective assessments. Furthermore, the authors did not report whether the different patterns of behavior as a result of error type were evident from the onset of the experiment or if they developed as a result of experience.

Does Error Type Affect Behavior?

The evidence from the limited number of studies that have compared false alarms and misses does suggest that each error type leads to different patterns of human behavior. While both types of errors reduce trust and reliance on the automation, they appear to do so differently. To date, the efforts of Dixon and Wickens (2003; 2004) have produced the clearest evidence for this phenomenon. Their results suggest that each type of error affects the way humans allocate their attention in multi-task environments where verification information is available. However, the measures used in their studies do not directly assess the monitoring behavior associated with the automation-supported task. Rather, assumptions about where attention is allocated are made by examining the performance across the various tasks. A similar criticism can be made of the driving simulator experiments, where inferences about the allocation of attention can be made by examining speed control, but no direct evidence of different monitoring behaviors as a function of error type is offered (e.g., Cotte et al., 2001; Gupta et al., 2001; Maltz & Shinar, 2004). The evidence discussed so far also suggests that the effects of error type on global assessments of trust and perceived reliability are influenced by the amount of workload that each generates (Dixon & Wickens, 2003), the salience of the error (Gupta et al., 2001; Johnson, 2004), and the cost associated with the outcome of error type (Bliss, 2003).

Time on Task

Another open question associated with the effects of error type is how different patterns of behavior emerge through time and experience. To date, there is a lack of data that provide information about how behavior changes as a result of each event, whether it is a false alarm, a miss, a correct detection or a correct rejection. Ideally, humans would weight all events—correct or incorrect—equally, which would facilitate predictions about human behavior associated with the use of automated decision support systems. However, it is likely that the weight assigned to each event changes as a function of experience with the system and the frequency of errors within a specific time period. In most of the studies previously discussed in which error type was manipulated, it was unclear when automation failures occurred. The distribution of errors across time could potentially influence the effects of error type on behavior. To begin to understand how humans accumulate evidence about the automation's reliability across time and how behavior is subsequently affected, it is critical to analyze this issue at a micro-level, where local changes in behavior can be observed.

Does it Matter When Errors Occur?

A considerable number of studies have investigated the effects of reliability on human behavior (e.g., Bliss et al., 1995; Maltz & Shinar, 2004; Parasuraman, Molloy & Singh, 1993; Rovira, Zinni & Parasuraman, 2002; Sanchez et al., 2004; St. John &

Manes, 2002; Vries, Midden & Bouwhuis, 2003; Wiegmann et al., 2001). In all these studies, reliability was manipulated by changing the overall error rate of the automation. For example, Sanchez et al. looked at the effects of three levels of reliability (100%, 80% and 60%) on reliance and trust. Each of these levels was descriptive of the overall or average reliability across the entire experiment. However, none of the previously mentioned studies presented or even mentioned the distribution of automaton errors across the experimental session.

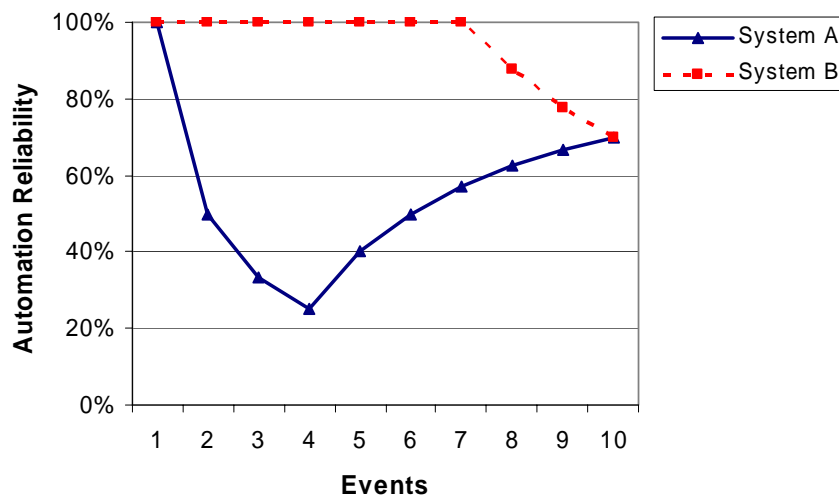


Figure 1. Illustration of two hypothetical systems with the same average reliability (70%) after 10 events but different error configurations through time. Each point in the graph illustrates the average cumulative reliability of the automation up to that point.

Figure 1 illustrates how two systems with the same average reliability (70%) can have two different time configurations of error occurrence. The location of errors within a specific range of time can have different effects on the way humans interact with the automation and on the overall trust that humans report at the end of a session (Wickens & Xu, 2002). Wickens and Xu would argue that humans interacting with System A (Figure 1) would have a different perception of the first automation error than humans interacting

with System B. This difference in perception is a result of different expectations of the automation as a product of experience (Figure 2). Wickens and Xu argued that the first automation failure can result in a more pronounced drop of trust and reliance on the automation than subsequent failures. According to them, the “first failure effect” (p. 8) largely depends on the instructions participants receive prior to their interaction with the system.

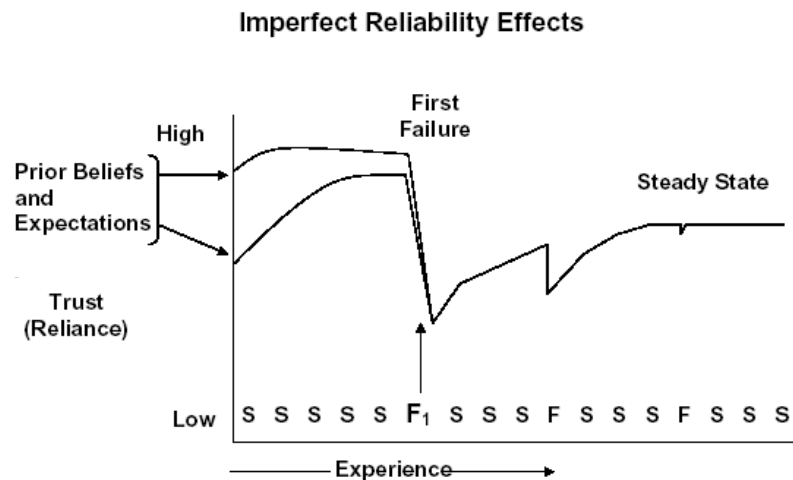


Figure 2. The effects of experience (x-axis) on trust and reliance (y-axis). Each “S” along the x-axis represents a successful event by the automation. Each “F” represents a failure by the automation. Adapted from Wickens and Xu (2002).

To date, limited evidence supports the existence of the “first failure effect” and the impact that it can have on the way humans perceive and interact with the automation (for evidence of the first failure effect see Molloy & Parasuraman, 1996; for evidence against it see Wickens, Helleberg & Xu, 2002). However, the possible existence of the first failure effect, suggests that the location of errors in time is an important component in the relationship between automation reliability and human behavior. Different configurations in the distribution of automation errors across time can have impact on the expectations and perceptions of the automation.

Age-Related Factors in Human-Automation Interaction

A limited amount of research within the automation domain has investigated the effects of age-related factors on automation use. However, age-related performance declines in divided attention, task-switching, and processing speed tasks (for a review see McDowd & Shaw, 2000) can shed light on the age-related differences in human-automation interaction that have been observed to date. The evidence thus far suggests that increased workload that results from age-related declines could be responsible for age-related differences in the use of automation. If workload is defined as “a portion of the operator’s limited capacity actually required to perform a particular task” (O’Donnell & Eggemeier, 1986, p. 42), then conceivably some of the cognitive declines related to age might affect the level of workload in a particular task. The purpose of automation in many cases is to reduce the level of workload for the human; therefore, age-related differences in the use of automation could be a product of increased workload resulting from age-related declines.

Because of the lack of automation studies that have manipulated workload while including an age grouping variable, the effects of workload on the use of automation by older adults can be examined by looking at studies that have manipulated automation reliability and age. Unreliable automation usually leads to increased workload (Johnson, 2004; Sanchez et al., 2004). This increase in workload as a function of reliability is caused by reduced levels of reliance on the automation, which means that humans have to constantly verify the support of automated aids with other information. There is evidence that when reliability is manipulated in a dual-task environment, the performance of older

adults in the non-automated task significantly decreases with lower levels of automation reliability (60% versus 80%, 100%; Sanchez et al., 2004). Sanchez et al. also found that older adults were more sensitive to decreases in automation reliability. This increased sensitivity to changes in reliability as a function of age was evident through older adults' subjective measures of trust and perceived reliability. Conversely, the trust ratings and reliability estimates of younger adults did not differ between the 80% and 60% reliability conditions. The increased sensitivity to reliability drops by older adults might be a product of the increased difficulty in performing two simultaneous tasks when the automation's error rate is higher.

In another study that might shed some light on whether workload affects the interaction of older adults and automation, Johnson (2004) manipulated the type of automation error in a dual-task environment and found that the perceived reliability estimates of older adults were significantly lower in conditions where false alarms were more prominent than misses. In this study, false alarms brought about higher task demands than misses, as participants had to perform an extra step to verify the support provided by the automation. Conversely, in a study using a traffic advisory information system (77% reliable), Kantowitz et al. (1997, experiment 1) found that misses affected older adults' trust ratings more than false alarms. They also found no age differences in subjective trust ratings. However, the task demands, defined by number of simultaneous tasks and time pressure, of the Kantowitz et al. study were lower than the Sanchez et al. (2001) and the Johnson (2004) studies.

Overall, there is some evidence in support of the idea that workload is a key determinant in the interaction of older adults with automation. However, in addition to

workload, other factors appear to contribute to the likelihood of use of the automation by older adults. In a study of visual detection in a luggage screening task, McCarley, Wiegmann, Wickens, and Kramer (2003) found that younger adults benefited from an automated aid by increasing their sensitivity relative to the non-automated condition. However, older adults' detection sensitivity did not increase with the presence of the automated aid. Interestingly, the perceived reliability estimates did not differ as a function of age. This study showed that even when both younger and older adults' perceived reliability estimates were similar there were still differences in how the two groups used the automation. Some have suggested that the acceptance and likelihood of using new technologies decreases with age (Kantowitz et al., 1993). However, age-related factors that might influence automation use have not been well researched.

The field of automation provides a valuable medium for the study of age-related cognitive declines, especially for constructs such as attention and workload. Overall, there is evidence for age-related differences in behavior toward automation. However, it is critical to begin to understand why the patterns of behavior are different as a function of age. One factor that appears to contribute to the age-related differences in the use of automation is increased workload, which is a product of age-related declines. However, there could be other factors that affect age-related differences in automation use, such as pre-conceived biases against the use of new technologies.

Human-Automation Interaction: Micro-Level Analysis of Behavior

A limited number of studies conducted to examine human-automation interaction, have analyzed the effects of reliability on human behavior at a micro level. A micro-level analysis of behavior involves an examination of changes in behavior across time. If humans are constantly adjusting their behavior in response to the characteristics of the automation, a micro-level analysis can provide critical information about the factors that humans are most receptive to and the factors that affect behavior.

Lee and Moray (1992; 1994) conducted experiments to understand the relationship between trust, self-confidence and reliance of an action-implementation automated system (an automatic feedstock pump). With an action-implementation automated system the human can reallocate an entire task/function to the automation. Lee and Moray (1992) examined how trust changed across participants from trial to trial as a function of system failures. A micro-level analysis revealed that while subjective trust was negatively affected by system failures, it recovered promptly in the presence of a properly working system. Furthermore, their analysis indicated that even failures of the same magnitude had different effects on subjective trust, depending on when the failures occurred and on the reliability of the automation during previous trials. For example, during a long sequence of automation failures (the automation was failing on every trial), their results showed that subjective trust actually began to increase (it is likely that participants became accustomed to a faulty system). This finding provides evidence that the way humans perceive automation failures changes as a function of the events that precede each error. In their study, when a failure was preceded by other failures it had

little effect on a person's trust of the automation. Conversely, when a failure was preceded by a period of reliable operation, it significantly decreased trust.

In an extension of their micro-level analysis, Lee and Moray (1994) plotted subjective trust, self-confidence, and use of the automatic controller per participant on a trial by trial basis. Participants had the choice between allocating one of the tasks to the automatic controller and performing the task manually. Self-confidence was assessed through a self-rating of how well the participant believed he/she could perform the task manually.

The results of the Lee and Moray (1994) study showed that during trials in which the difference between trust in the automation and self-confidence ratings was positive, most participants relied on the automation. This analysis also revealed the points in time in which the difference between trust and self-confidence went from positive to negative and vice-versa as a function of errors. Their analysis showed that reliance on the automation changed as a function of both trust and self-confidence.

In summary, the micro-level analyses by Lee and Moray (1992; 1994) provided valuable insight into the decisions by humans to use automation by choosing to perform the task manually or allocate it to the action-implementation automation. This type of micro-level analysis approach might also prove to be valuable for understanding human behavior in environments with automated decision support systems. Specifically, this approach can facilitate the distinctions in behavior as a function of false alarms and misses by the automation.

An additional analysis of an experiment conducted by Johnson (2004) revealed the potential benefits for a micro-level approach in the context of decision support aids.

Johnson's study specifically examined the effects of false alarms and misses of an automated decision aid on trust and reliance. In a multi-task environment, participants received support of an automated aid to help them detect system failures. They had the option to either rely on the automated aid or not rely on it (verifying its support by viewing a pair of gauges). The prevalence of error type was manipulated in three conditions: majority misses (9 misses, 3 false alarms), majority false alarms (3 misses, 9 false alarms), and equal (6 misses, 6 false alarms).

The results of the Johnson (2004) study indicated that across the entire experimental session there were no significant differences in reliance as a function of the error type manipulation for the younger adults. However, when the data were plotted across time, a visual inspection suggests that during periods when the decision support aid was providing an alert, the false alarm group had a higher non-reliance rate than the miss group. Conversely, during periods when the decision support aid was inactive, the miss group had a higher non-reliance rate (see Appendix A for an illustration of the micro-level analysis). However, Johnson's study was not specifically designed to accommodate this type of analysis. For example, within the majority false alarm condition, there is early "contamination" by a miss, which is the second error by the automation. Nevertheless, the patterns observed in these data suggest that a micro-level examination of behavior as a function of type of error can lead to a better understanding of the effects of error type on reliance. Gaining this understanding can provide insight into the way humans perceive errors and how this perception affects behavioral interactions with an automated aid.

Objectives of the Current Investigation

The objective of this investigation was to gain an in-depth understanding of specific variables that affect human interaction with an automated decision support aid. Two types of factors that affect human-automation interaction were investigated, those specific to the automation (Error Type and Distribution of Errors) and those specific to the human (Age and Domain Experience). The aim was to comprehend how these variables influence the way humans accumulate and integrate information about the reliability of an automated decision aid and how the use of this information affects their behavior. This objective was accomplished by analyzing behavior as a function of different variables that affect the weight humans place on the support provided by an automated decision support aid.

Another objective of this research was to evaluate age-related effects on the use of automation. The effect of age on automation use is an area that has received little attention thus far. Automated environments can provide a valuable medium for the study of aging and cognition. Issues such as cautiousness, information accumulation and information integration are important aspects of reliance on automation as well as cognitive aging. Clearly, understanding age-related differences in moderators and mediators of automation usage and trust are critical to theories and models of human-automation interaction. While there is some evidence that the nature of human-automation interaction differs as a function of age (Kantowitz et al., 1997; McCarley et al., 2003; Sanchez et al., 2004), the factors that contribute to these age-related differences are still unclear.

In addition to investigating age-related effects, possible effects of domain experience were evaluated. Two different populations were tested, Farmers and Non-Farmers. The Farmer group had experience operating agricultural vehicles. The aim of evaluating population differences was to evaluate factors such as experience with a specific domain on the manipulated variables in the present study. The automation was comprised of an automated collision avoidance system embedded within a multi-task environment. The experimental environment was a simulated task familiar to Farmers who have harvesting experience.

CHAPTER 2

METHOD (EXPERIMENT 1)

Overview

Two experiments, which included three different populations, were conducted. The first experiment was conducted in the Iowa/Illinois region and it involved the participation of younger adults with experience operating agricultural vehicles. The objective of this experiment was to understand the effects of different error type on trust in and reliance on automation. Participants in Experiment 2 were younger and older adults without experience operating agricultural vehicles. This study involved the investigation of the effects of different distribution of errors on reliance and trust in the automation as well as the effects of error type. In addition, age-related differences in the use of automation were investigated. Both experiments used identical experimental simulator and procedures.

Participants

Twenty younger adults (aged 19-29, $M = 23.2$, $SD = 2.9$, all males) with experience operating agricultural vehicles, which included tractors and harvesting combines, participated in this study. A harvesting combine is a farm machine used to harvest different types of crops. Combines are highly automated currently; hence, the

participants have experience with automation but not the particular automation used in this study. Participants were recruited from the Iowa/Illinois regions and were given a John Deere hat for their participation. General health information was collected before the study, and none of the participants reported conditions that would affect visual acuity. Basic ability data are provided in Appendix B as a function of the Error Type Manipulation. The groups were very similar across all measures with the exception of the memory span (Reverse Digit Span; Wechsler, 1981).

Design

The manipulation in this study was the automation Error Type: False Alarms and Misses. Each condition included 10 such errors (either 10 false alarms or 10 misses). At the end of the session an opposite or exception error occurred (e.g., a false alarm in the Misses Condition). The 10 similar errors were quasi-randomly distributed within the first 57 minutes. The exception error occurred at minute 58. The distribution of the errors can be found in Appendix C (column labeled *Throughout*).

Materials

Equipment

The experiment utilized an IBM-compatible laptop computer (3.2 GHz, 512 MB RAM) with a 15.4" display. The input device was an external, standard QWERTY keyboard, which was attached via USB port. Except for the keys needed to interact with the simulator, all keys were removed. The simulator was coded using Macromedia Flash MX.

Task

The experimental simulator was comprised of two main tasks, a collision avoidance task and a tracking task. The collision avoidance task was supported by an “imperfect” automated collision avoidance system and the tracking task was performed manually. The total experiment lasted approximately two hours, half of which was spent on the experimental phase.

The simulator was divided into three separate windows, the outside view, the collision avoidance window, and the tracking task window (Figure 3). There were a total of six keys mapped onto specific functions in the system. The *up*, *down*, *left*, and *right* arrow keys were used to perform the tracking task. The *spacebar* was used to view the outside window and the *enter* key was used to avoid obstacles. The system only allowed one key to function at a time. For example, if the spacebar was held down, no other key would activate its respective function.

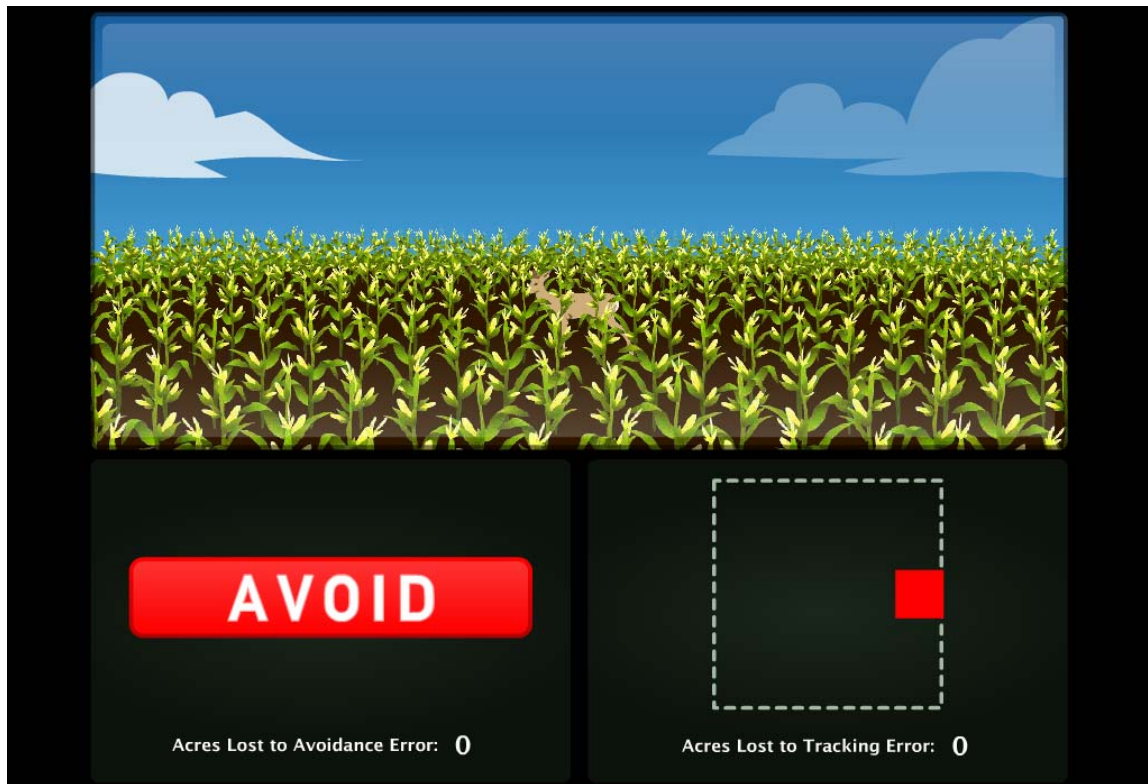


Figure 3. Screenshot of the simulator. Top window is the outside view, bottom left window is the collision avoidance system, and the bottom right window contains the tracking task

Collision Avoidance Task

The top window represented the outside view from the perspective of an operator sitting in the cab of an agricultural vehicle. Obstacles (deer) were visible in the outside view window. The outside view remained hidden unless the spacebar was pressed and it remained visible until it was released. The bottom left window contained the automated collision avoidance system. Within this window, there was an automated collision avoidance indicator that indicated the presence of a deer by turning from white to red and

generating an “AVOID” warning. For correct detections, the indicator turned red as soon as the deer appeared in the top window and remained red for 15 seconds. During automation false alarms, the indicator also turned red for 15 seconds, but there was no deer in the top window. For automation misses, a deer appeared in the top window for 15 seconds, but the indicator did not turn red. During periods of correct rejection by the automation, the indicator remained white and no objects appeared in the top window.



Figure 4. Immediate feedback about the outcome of actions in the collision avoidance task. A “collision” could result from either not pressing the enter key within 15 seconds of the appearance of a deer or from pressing enter while the automation was generating a false alarm. An “unnecessary maneuver” resulted from pressing enter in the absence of a deer and an alarm. The “obstacle avoided” feedback appeared after pressing enter in the presence of a deer.

When a deer appeared in the top window, participants were responsible for pressing the enter key. Participants were told that once enter was pressed, the vehicle would automatically steer around the deer, although there was no way to actually witness the action of steering around the deer because pressing enter meant that the spacebar had been released, which blocked the outside view. No more than one deer ever appeared at once. Deer did not always appear in the same location in the top window, although they always appeared within the middle 50% of the horizontal axis of the outside view window.

The objective of the collision avoidance task was to complete as much of the task without losing any acres due to an “avoidance error.” Any avoidance error resulted in the loss of one acre. Participants received immediate feedback about all actions/inactions associated with the collision avoidance task (Figure 4).

In addition to immediate feedback, participants also received cumulative feedback about their performance via the “acres lost to avoidance error” counter at the bottom of the lower left window (Figure 3). Any time an error was made, an acre was added in this counter. Avoidance errors occurred as a result of the following actions/inactions:

1. Failing to press enter within 15 seconds of the appearance of a deer. This scenario could result from either ignoring a true alarm or failing to detect a deer during a miss by the collision avoidance system.¹ The immediate feedback as a result of failing to press enter within 15 seconds of the appearance of a deer was “COLLISION” (See Figure 4).

¹ A collision could also result from the participant not detecting a deer even if they viewed the outside window. This scenario only appeared to have occurred twice (two participants, once each) throughout the entire investigation (Experiments 1 and 2).

2. Pressing enter during a false alarm by the automation. The immediate feedback for this action was “COLLISION” (See Figure 4). Participants were told that false alarms were generated when the collision avoidance system detected an obstacle that was not in its direct path but the automation perceived it to be in its direct path. Therefore, pressing enter during a false alarm resulted in a collision because the vehicle unnecessarily steered away from its path into the obstacle that caused the false alarm. Consequently, failing to press enter during a miss by the automation and pressing enter during a false alarm both resulted in a collision. Equating the consequences of both Error Types was done in an effort to remove any biases associated with the occurrence of each Error Type.
3. Pressing enter during a correct rejection by the automation. If the vehicle was directed off-path unnecessarily (no deer present) and no alarm from the collision avoidance system, a message saying “UNNECESARY MANUEVER, 1 ACRE LOST” appeared (See Figure 4). This penalty was implemented to prevent participants from randomly pressing enter in the absence of alarms.

If participants pressed *enter* within 15 seconds of when a deer appeared, a message saying “OBSTACLE AVOIDED” was displayed (See Figure 4). All three feedback messages appeared over the top window and remained visible for 15 seconds, irrespective of whether the spacebar was pressed or not.

Reliability of the Collision Avoidance System

Within the 60 minutes of the experimental session there were a total of 240 events. Events lasted 15 seconds each and were comprised of correct detections, correct rejections, false alarms, and misses by the automation. The reason for choosing 15 seconds as the length of an event was because each alarm (true and false) lasted 15 seconds and also when deer appeared, they remained visible for 15 seconds.

The overall reliability of the automation, which was calculated by dividing 11 automation errors over 240 total events, was 95.4%. Of the 240 events, 179 were correct rejections, 50 were correct detections, 10 were either false alarms or misses, and 1 was either a false alarm or a miss. In the False Alarms Condition, the false alarm rate of the automation was 5.3% and the miss rate was 1.97%. In the Misses Condition, the false alarm rate of the automation was 0.56% and the miss rate was 16.7%. This mismatch in false alarm and miss rate exists because of the mismatch in the number of correct rejections (179) and correct detections (50) events. However, the number of errors in each Error Type condition was the same, which allowed comparisons between the two conditions.

Tracking Task

The tracking task was located in the bottom right window (See Figure 1). The objective of this task was to prevent the filled square from reaching the edge of the box. Any time the filled square overlapped with any of the dashed lines, it changed from green

to red. The square was controlled with the *up*, *down*, *left*, or *right* keys. The square randomly deviated from the center in four directions (up, down, left, right), but never in more than one direction at a time. The time it took for the square to go from the center to the edge ranged from 1200 ms to 1400 ms. Once the vehicle began to deviate from the center, its speed was constant. Upon movement of the square in a specific direction (e.g., up), the key pointing to the opposite direction (e.g., down) needed to be pressed once to return it to the center position. Once the square returned to the center position a random amount of time ranging from 1 to 2 seconds elapsed before it began moving again. Participants were told that for every 15 seconds that the square overlapped with the dashed lines they would lose an acre due to a “tracking error”. The 15 seconds that it took to lose an acre were accrued cumulatively. For example, if the square was allowed to reach the edge five times and each time it remained red for three seconds, this resulted in the loss of an acre. Cumulative performance feedback for this task was provided by presenting the number of acres that had been lost because of “tracking error.”

Procedure

Participants were asked to complete a demographics form and perform four abilities tests including Paper Folding (Ekstrom, French, Harman & Dermen, 1976) Shipley Vocabulary (Shipley, 1986), Digit Symbol Substitution (Wechsler, 1981), and Reverse Digit Span (Wechsler, 1981). Next, participants received an explanation of the task and were provided with training. The explanation of the task and the training for both experimental conditions (False Alarms and Misses) was identical.

During training, participants were first exposed to the tracking task for three minutes. The objective during the three minutes of training was to keep the square within the dashed lines. Training on the tracking task continued after three minutes until criterion was reached. Criterion for this task consisted of being able to keep the vehicle inside the dashed lines during the last minute of the training session.²

Next, participants received a thorough explanation of the collision avoidance task. Participants were told the collision avoidance system was “very reliable, but not perfect.” Screen shots of the four different states of the system (correct detections, correct rejections, false alarms, and misses) were presented and the consequences of pressing or not pressing enter for each scenario were reviewed in detail. Participants were allowed to ask any questions that would help them better understand the task. However, the standard response to questions that could potentially bias participants’ perception of the automation such as “how often will the automation be wrong?” or “which type of error is it more likely to generate?” was “I’m not sure of that, all I know is the automation is very reliable, but it’s not perfect.”

Participants received training on the collision avoidance task for five minutes (without the tracking task). During the training session, the automation correctly detected eight objects, missed one, and generated one false alarm. Following training for the collision avoidance task, another five minute training session was conducted with both tasks (tracking and collision avoidance). Participants were given a short break upon completion of the training session.

The experimental phase lasted 60 minutes. Participants were asked to complete the entire session without taking a break. However, they were told that if they really

² All participants in the investigation reached criterion within the first three minutes

needed to take a break they could. All participants completed the experimental phase without requesting a break. Upon completion of the experimental phase, a subjective trust questionnaire was administered (Appendix D). Lastly, participants were debriefed and thanked for their participation.

Dependent Variables

Collision Avoidance Task

All of the actions (key presses) associated with this task were recorded and time-stamped. Performance measures included the following:

1. Number of times the outside was viewed (spacebar presses): the spacebar presses measure accounted for every press of the spacebar, which included instances in which participants chose to “double check” the outside window more than once within a few seconds. For example, in the presence of a single alarm, a participant might have pressed the spacebar three or four times within the span of 3 seconds. This behavior was likely a product of not being entirely sure that a deer had been detected. Therefore, second, third, and fourth presses of the spacebar, within the span of a few seconds were not as pure indicators of non-reliance as the first press.
2. Number of times the outside was viewed (first spacebar presses): this measure consisted of the first spacebar press made in every 15 second event. This measure of

- reliance was the primary one used for the subsequent analyses. The maximum number of first spacebar presses per participant was 240.
3. Number of spacebar presses during alarms: this measure included the first spacebar press made in every 15 second event during alarms.
 4. Number of spacebar presses during non-alarms: this measure included the first spacebar presses during periods when the collision avoidance system was idle (non-alarms). Along with spacebar presses during alarms, this measure made up the first spacebar presses measure (#2).
 5. Total time the spacebar was pressed.
 6. Total number of errors in the collision avoidance task, which included:
 - False positives as a result of an automation false alarm (collision).
 - False positives as a result of pressing enter in the absence of an alarm and a deer (unnecessary maneuver).
 - False negatives as a result of the absence of an alarm during a miss by the automation (collision).
 - False negatives as a result of ignoring a true alarm (collision).
 7. False positive rate: the number of times that a participant pressed enter in the absence of a deer divided by the total number of bins when deer were absent
 8. False negative rate: the number of times a participant failed to press enter in the presence of a deer divided by the number of deer that were present.
 9. Subjective trust (assessed at the end of the experimental session; Appendix D).
 10. Perceived reliability (assessed at the end of the experimental session; Appendix D).

Tracking Task

Performance measures for the tracking task included the following:

1. Total number of acres lost due to tracking error: this measure consisted of the total amount of time that the square was red (at the edge of the box) divided by 15. This measure was consistent with the instructions provided to participants regarding ways in which points would be lost.
2. Red events: number of times the square reached the dashed lines.

Expected Results

The results of previous studies have shown that the prevalence of false alarms, relative to misses, appears to lead to greater decrements in trust and reliance (Johnson, 2004). Some have suggested that the greater impact of false alarms on trust is a product of the saliency associated with this type of error. Others have suggested that misses actually have a greater impact on trust and reliance because they often have more serious consequences associated with them. Given that the cost of failing to detect both types of automation errors was equal in the current experiment, it was expected that the False Alarms Condition would lead to lower levels of subjective trust and perceived reliability of the automation.

The results of several studies (Cotte et al., 2000; Dixon & Wickens, 2003; 2004; Gupta et al., 2001; Lehto et al., 2000) provide some evidence that human behavior is affected differently as a function of the prevalence of a specific type of error by the

automation. Therefore, it was expected that during periods when the collision avoidance system is idle (non-alarms), participants who are in Misses Condition would view the outside window more frequently than those in the False Alarms Condition. Conversely, participants in the False Alarms Condition were expected to view the outside more frequently during periods when the alarm is active (red). This prediction was based on the assumption that humans will adjust their behavior according to the characteristics of the automation in an effort to reduce the amount of time and effort they have to use in verifying the automation while trying to achieve a high level of performance in both tasks.

Because the simulation used for the experiment is a dual-task environment, it was expected that performance on the tracking task would be inversely related to the level of non-reliance on the automation. This prediction was based on previous evidence, which suggests that unreliable automation can have a negative impact on the performance of concurrent tasks (Dixon & Wickens, 2003; Sanchez et al., 2004). Low automation reliability usually leads to low levels of reliance on the automation. Not relying on the automation means that more cognitive and physical resources are allocated to the verification of other information rather than to the performance of concurrent manual tasks.

Method (Experiment 2)

Participants

Two groups of participants completed this study. Sixty younger adults (aged 18-24, $M = 19.5$, $SD = 1.6$) and sixty-one older adults (aged 65-75, $M = 69.7$, $SD = 3.5$). One older adult was excluded from the analysis because this participant dozed off during the experiment. The analyses were performed with the remaining sixty older adults (aged 65-75, $M = 69.6$, $SD = 3.5$). The younger adults were recruited from introductory psychology courses at the Georgia Institute of Technology and were given course credit for their voluntary participation. The older adults were paid \$25 for two hours of participation and were recruited from the Atlanta metropolitan area. Older adults were screened over the phone for good health and for any disease or condition that would affect their vision. All participants' general health information was also collected before the study.

General participant characteristics by Age, Error Type and Distribution of Errors can be found in Tables E1, E2, and E3 (Appendix E). Younger Adults performed significantly better than Older Adults in the Digit Symbol Substitution and Reverse Digit Span tests while Older Adults outperformed Younger Adults in the Shipley Vocabulary test.

Design

Experiment 2 was a 2 (Age: Younger Adults and Older Adults) x 2 (Error Type: False Alarms and Misses) x 3 (Distribution of Errors: Throughout, First Half, and Second Half) fully factorial design. All the variables were manipulated between-subjects; therefore, there were a total of 10 participants per experimental group.

The main experimental design difference between Experiment 1 and Experiment 2 was the addition of the Distribution of Errors manipulation. As in Experiment 1, every group was exposed to 11 automation errors (10 similar errors and one exception error). The exception error occurred after all the similar errors had occurred. Within the Throughout Condition, errors were distributed in the same order as in Experiment 1 (the 10 similar errors were distributed quasi-randomly within the first 57 minutes and the exception error occurred at minute 58). Within the First Half Condition, the 10 similar errors were quasi-randomly distributed within the first 27 minutes of the task and the exception error occurred at minute 28. Within the Second Half Condition, the 10 similar errors were quasi-randomly distributed between minute 30 and minute 57, and the exception error occurred at minute 58. The distribution of errors during the first 30 minutes of the First Half Condition was identical to the distribution of errors during the last 30 minutes of the Second Half Condition (See Appendix B for the distribution of the errors for all three conditions).

Materials

Equipment

The experiment utilized IBM-compatible computers (3.2 GHz, 4.1 GB RAM) connected to 19 inch cathode-ray tube displays. Refresh rate was set at 85 Hz. The input device was a standard QWERTY keyboard. The experimental simulator was identical to the one used in Experiment 1.

Task

The experimental simulator and the task were identical to Experiment 1.

Procedure

The experimental procedure was identical to the procedure followed in Experiment 1.

Dependent Variables

The same measurements were collected as in Experiment 1.

Expected Results

It was expected that a higher concentration of errors in the second half of the experiment would have a stronger negative impact on trust ratings and perceived reliability estimates of the automation than a concentration of errors in the first half of the experiment. Therefore, participants in the First Half Condition should have higher trust ratings and estimates of the automation's reliability than participants in the Second Half Condition. This pattern of results will not only provide evidence that the location of errors in time is an important component of reliability, but it will also suggest that the closer errors occur to the time when humans are asked to provide an estimate of the reliability, the lower estimates are likely to be.

Reliance patterns as a function of Error Type were expected to be similar to the patterns observed in Experiment 1. Participants in the False Alarms Condition were expected to view the outside window more frequently during alarms than participants in the Misses Condition. Conversely, during non-alarms it was expected that participants in the Misses Condition would exhibit less reliance on the automation than participants in the False Alarms Condition. Because participants were expected to develop different patterns of reliance as a function of Error Type, it was expected that the detection rate of the exception error (e.g., the miss in the mostly False Alarms Condition) would be less than the detection rate of the rest of the errors.

The results of previous research suggest that the negative impact of unreliable events is greater for older adults than for younger adults (Sanchez et al., 2004).

However, given that the reliability of the automation used in this investigation was rather high, it was expected that older adults will benefit more from automated support and their estimates of reliability will likely be higher than the estimates of younger adults. Age-related differences in tracking and obstacle detection performance favoring the younger adults were also expected.

The negative effects of an error were expected to decrease as time passed and participants were exposed to more errors. To test this hypothesis, reliance on the automation before and after errors was compared. It was expected, for example, that the increase in monitoring of the outside view from before the first error to after it will be greater than the increase in monitoring of the outside view from before the sixth error to after it. If humans perceive errors as “common” occurrences in the system, they should adjust their behavior accordingly.

A comparison will be made between participants in Experiment 1 and the Younger Adults in the Throughout Condition from Experiment 2. This comparison will shed light on the possible effects of domain experience on human-automation interaction. It is expected that errors will have a greater effect on the trust and reliance of the participants in Experiment 1 (Farmers). Because of their experience with agricultural vehicles and highly reliable automation, the Farmer group may be less willing to accept any failure by the automation than individuals without that type of experience.

CHAPTER 3

RESULTS

Overview of Analyses

The results of this investigation were analyzed at both the micro- and macro-levels. The macro-level approach consisted of analyses of variance (ANOVA) to make comparisons across all experimental conditions. A correlation analysis was also performed to identify any meaningful relationships between the variables of interest. The complete set of results for the ANOVAs and correlations can be found in Appendices F and G respectively.

The micro-level analyses were geared toward examining and understanding how human behavior, specifically reliance on the automation, changed as a function of time and the variables of interest. In this analysis, the measures “first spacebar presses during alarms” and “first spacebar presses during non-alarms” were plotted across the entire experimental session, which encompassed 240 bins of 15 seconds each.

The results from Experiments 1 and 2 were organized in the following order:

1. The results of Experiment 2 were presented first; beginning with a macro-level analysis of the effects of Age, Error Type, and Distribution of Errors on reliance, trust and perceived reliability.

2. Next, a micro-level analysis of the reliance data for the Younger Adults provided a closer look at the development of reliance across time as a function of Error Type and Distribution of Errors.
3. A micro-level analysis of the reliance data for the Older Adults facilitated a closer examination of age-related differences in the development of reliance as a function of Error Type and Distribution of Errors.
4. The effects of Agricultural (domain) Experience on reliance and trust were presented at both the macro and micro levels (Experiment 1 versus Younger Adults in Experiment 2).
5. Lastly, the effects of Age, Error Type, and Distribution of Errors on the following measures were presented:
 - a. Detection of the exception error
 - b. Effects of first and subsequent failures on reliance
 - c. Tracking task performance
 - i. Number of acres lost to tracking error
 - ii. Number of red events
 - d. Obstacle detection performance
 - i. Number of acres lost to collision errors
 - ii. False negative and positive rates

Summary of Results

Table 1 contains a summarized version of the results from Experiments 1 and 2. The effects of the experimental manipulations on the main dependent measures are listed.

Table 1. Summary of results from Experiments 1 and 2

Dependent Measures	Experimental Manipulations			
	<i>Age</i>	<i>Error Type</i>	<i>Distribution of Errors</i>	<i>Agricultural Experience</i>
<i>Reliance</i>	a) OA relied less on automated alarms b) OA took longer to adjust their reliance behavior as a function of Error Type	a) During alarms, FA decreased reliance & Misses increased reliance* b) during non-alarms, FA increased reliance & Misses decreased reliance*	a) Throughout condition led to less overall reliance b) higher concentration of errors did not lead to faster development of reliance patterns as a function of Error Type	Farmers relied less on automated alarms
<i>Trust</i>	OA reported higher trust ratings*	FA led to higher trust ratings Misses	Recency of errors led to lower trust ratings*	No effect
<i>Perceived reliability</i>	OA had higher estimates of reliability*	No effect	Recency of errors led to lower estimates of reliability*	No effect
<i>Tracking</i>	YA had higher performance*	FA led to higher performance	No effect	No effect
<i>Obstacle detection</i>	YA had higher performance*	FA led to higher performance	No effect	No effect
Note. * represents expected result YA = Younger Adults, OA = Older Adults, FA = False Alarms				

Reliance and Trust in Automation: A Macro-Level Analysis (Experiment 2)

First Spacebar Presses

The main measure of non-reliance in this investigation was the first spacebar press per event. Therefore, because there were a total of 240 events, the maximum number of first spacebar presses per participant was 240. The results of the total number of spacebar presses can be found in Appendix F. It is worth noting that there was a strong relationship between total spacebar presses and first spacebar presses, $r(120) = .94, p < .01$.

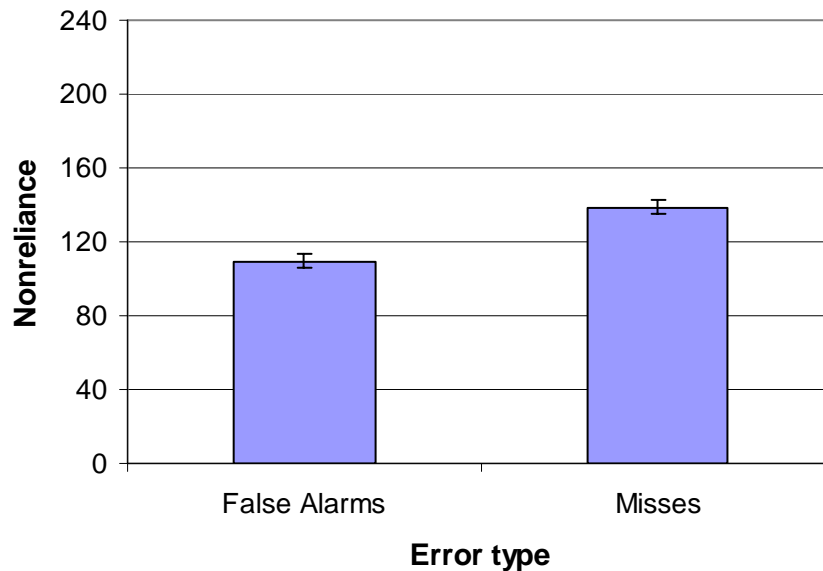


Figure 5. Non-reliance (first spacebar presses) by Error Type for all participants in Experiment 2. The maximum number of spacebar presses per participant was 240. Error bars represent the standard error of the mean.

For first spacebar presses, there was a main effect of Error Type where participants in the Misses Condition relied less on the automation than those in the False

Alarms Condition, $F(1, 108) = 9.70, \eta^2 = .08, p < .05$ (Figure 5). Given the nature of the system used in this study, this finding was not surprising. In the Misses Condition, participants had to constantly check the outside view in an attempt to “catch” misses by the automation. This state of the system, when the collision avoidance system was inactive, consisted of approximately 75% of the total time in the experiment. However, in the False Alarms Condition, participants only had to double check the output of the automation when it generated an alarm, which consisted of approximately 25% of the total experiment time.

There was also a main effect of Distribution of Errors on first spacebar presses, $F(2, 108) = 3.74, \eta^2 = .07, p < .05$. A Tukey Post Hoc analysis revealed that participants in the Throughout Condition relied less on the automation than those in the First Half Condition. It is worth noting that there was a significant interaction between Error Type and Distribution of Errors, $F(2, 108) = 10.98, \eta^2 = .17, p < .05$ on first spacebar presses.

Spacebar Presses during Alarms and Non-Alarms

In an effort to determine if the experimental manipulations had an effect on reliance during different states of the automation, the *first spacebar presses* data were divided into periods when the collision avoidance alarm was active (alarms) and periods when it was inactive (non-alarms). In contrast to the results obtained for the first spacebar press, during periods when the collision avoidance system was active (alarms), participants in the False Alarms Condition relied less on the automation than those in the Misses Condition, $F(1, 108) = 52.4, \eta^2 = .33, p < .05$ (Figure 6). However, during periods

when the collision avoidance system was inactive (non-alarms), participants in the Misses Condition relied less on the automation than participants in the False Alarms condition, $F(1, 108) = 22.5, \eta^2 = .17, p < .05$ (Figure 7).

Interestingly, there was a main effect of Age for spacebar presses during alarms, where Older Adults relied less on the automation than Younger Adults, $F(1, 108) = 10.6, \eta^2 = .09, p < .05$. However, there was not a main effect of Age for spacebar presses during non-alarm periods. This finding suggests that older adults are less likely than younger adults to rely on automation when it is actively providing support in the form of an alarm. However, older adults are just as willing to rely on the automation during periods when the automation is idle (non-alarms).

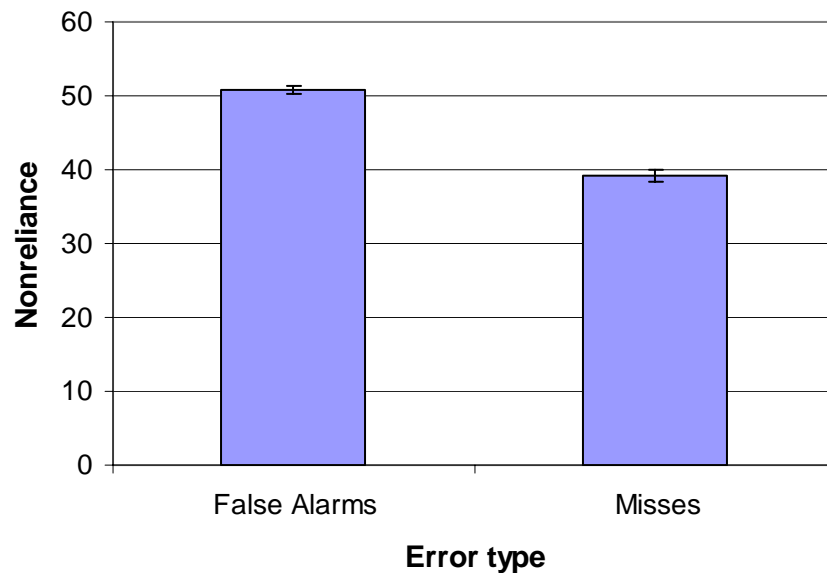


Figure 6. Non-reliance (first spacebar presses during alarms) by Error Type for all participants in Experiment 2. The maximum number of spacebar presses during alarms was 60. Error bars represent the standard error of the mean.

The contrast in reliance behavior as a function of Error Type during different system states brings to light the complex relationship between automation reliability and reliance on the automation. While the prevalence of each type of error can increase reliance on the automation during a specific system state, it has the opposite effect on the other state. For example, participants who were used to an automated system with a looser detection criterion (False Alarms Condition) relied less on the automation during alarms than those in the Misses Condition, but were more reliant on it during non-alarm periods. This effect was explored in more detail in the following section in which patterns of reliance were examined and analyzed across time.



Figure 7. Non-reliance (first spacebar presses during non-alarms) by Error Type. The maximum number of spacebar presses during non-alarms was 180. Error bars represent the standard error of the mean.

Total Time for Spacebar Presses

The pattern of results for the total time that the spacebar was pressed was similar to the pattern of results for first spacebar presses. There was a significant interaction between Error Type and Distribution of Errors, $F(2, 108) = 11.5$, $\eta^2 = .18$, $p < .05$ (Figure 8). There were also main effects of Age, $F(1, 108) = 11.0$, $\eta^2 = .09$, $p < .05$; Error Type, $F(1, 108) = 15.0$, $\eta^2 = .12$, $p < .05$; and Distribution of Errors, $F(2, 108) = 4.4$, $\eta^2 = .08$, $p < .05$. The average amount of time that spacebars were pressed across all participants was 870 ms ($SD = 340$ ms). The short amount of time per spacebar press was not surprising because participants had to constantly attend to the tracking task.

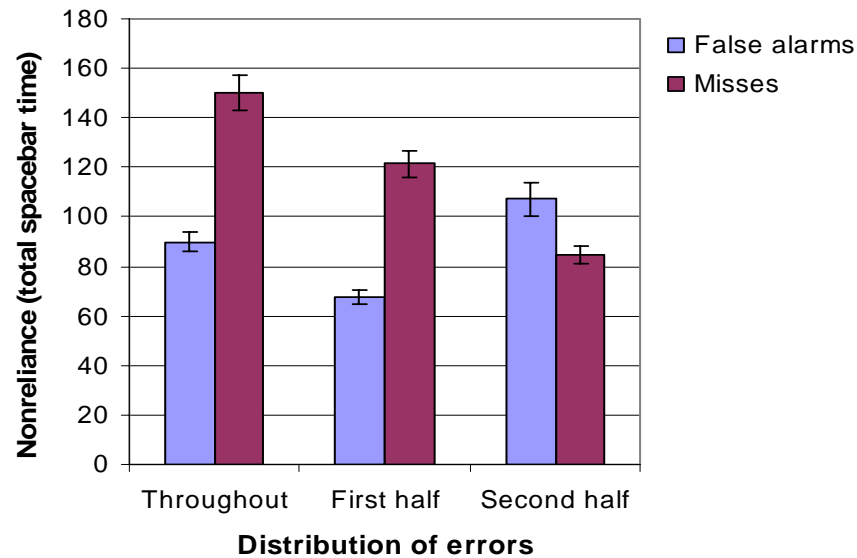


Figure 8. Total amount of time the spacebar was pressed by Error Type and Distribution of Errors. Error bars represent the standard error of the mean.

Subjective Trust

As expected, trust ratings were negatively correlated to the measures of non-reliance (See Appendix G for correlations table). These correlations provide more evidence suggesting a strong relationship between trust in and reliance on the automation. The analysis on the trust data showed that False Alarms led to higher trust ratings than Misses, $F(1, 108) = 15.2, \eta^2 = .12, p < .05$. This finding was consistent with previous studies that have found that misses lead to lower levels of trust relative to false alarms (Dixon & Wickens, 2003), but inconsistent with the results of other studies that have found the opposite (Gupta et al., 2001; Johnson, 2004).

There was a main effect of Distribution of Errors, $F(2, 108) = 22.6, \eta^2 = .30, p < .05$. A Post Hoc Tukey analysis showed that both the Throughout and First Half Conditions led to higher trust ratings than the Second Half Condition (both p 's < 0.05). Given that in all three conditions the total number of errors was the same, this finding provides some evidence for the degrading effects of recency of errors on trust. This finding was explored in greater detail with the micro-level analyses following this section.

There was a significant interaction between Age and Distribution of Errors, $F(2, 108) = 3.4, \eta^2 = .06, p < .05$. Older Adults reported higher trust ratings than Younger Adults within the Throughout and First Half Conditions but not within the Second Half Condition (Figure 9). This interaction suggests that older adults were more likely to base their trust ratings of the automation on their most recent interaction with it. Furthermore, there was a main effect of Age, $F(1, 108) = 18.0, \eta^2 = .14, p < .05$, where Older Adults

reported higher trust ratings than Younger Adults. The main effect of Age provides support to the idea that if the level of reliability of the automation is high enough to make it useful, then the benefit of the automation to older adults, relative to younger adults, is reflected in their subjective assessment of it (Sanchez et al., 2004).

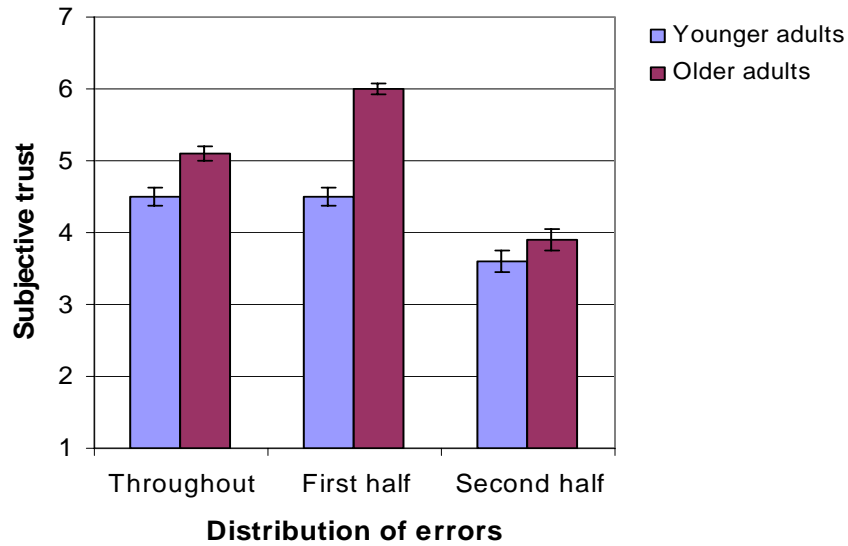


Figure 9. Subjective trust ratings by Age and Distribution of Errors. Error bars represent the standard error of the mean.

Perceived Reliability

The pattern of results for the perceived reliability measure was consistent with the pattern observed for the trust ratings. Although there was a main effect of Age, where Older Adults reported higher reliability estimates than Younger Adults, $F(1, 108) = 4.6$, $\eta^2 = .04$, $p < .05$, there was also a significant interaction between Age and Distribution of Errors, $F(2, 108) = 4.2$, $\eta^2 = .07$, $p < .05$. Older Adults reported higher reliability

estimates than Younger Adults within the Throughout and First Half Conditions, but not within the Second Half Condition (Figure 10). Interestingly, within the Younger Adult group there were no differences in perceived reliability as a function of the Distribution of Errors. However, within the Older Adult group, reliability estimates were significantly lower in the Second Half Condition than in the Throughout and First Half Conditions.³ These results suggest that when evaluating the reliability of the automation, older adults are more heavily influenced by their most recent experience with it. There was also a main effect of Distribution of Errors, $F(2, 108) = 9.7, \eta^2 = .15, p < .05$. A Post Hoc Tukey analysis showed that both the Throughout and First Half Conditions led to higher reliability estimates than the Second Half Condition (both p 's < 0.05). These results provide more evidence for a recency effect in estimating the reliability of the automation.

Another explanation for the effect of the Distribution of Errors on reliability estimates is that participants in the Second Half Condition had built up their expectations of the automation during the first 30 minutes and the sudden appearance of errors during the second half led to lower estimates of reliability. There is some evidence that humans have high expectations about the reliability of automated systems (Crocoll & Coury, 1990; Dijkstra, Liebrand & Timminga, 1998; Dzindolet, Pierce, Beck, Dawe & Anderson, 2001). Therefore, a sudden change from perfect reliability to imperfect could have a considerable impact in humans' perception of it. A mismatch between expected and actual reliability could also be a reason for the underestimation of reliability by most participants (See Figure 10). This underestimation of reliability was consistent with the

³ Even though the pattern of results of the subjective trust ratings and perceived reliability measures were similar, and there was a strong correlation between the two measures, $r(120) = .54, p < .01$, the pattern of results were not exactly the same. Specifically, the trust ratings of Younger Adults differed as a function of Distribution of Errors, but this was not the case for estimates of reliability. This difference suggests that trust ratings and perceived reliability measure different components of the construct of trust.

results of previous studies (e.g., Dixon & Wickens, 2003; Johnson, 2004; Moray et al., 2000; Sanchez et al., 2004).

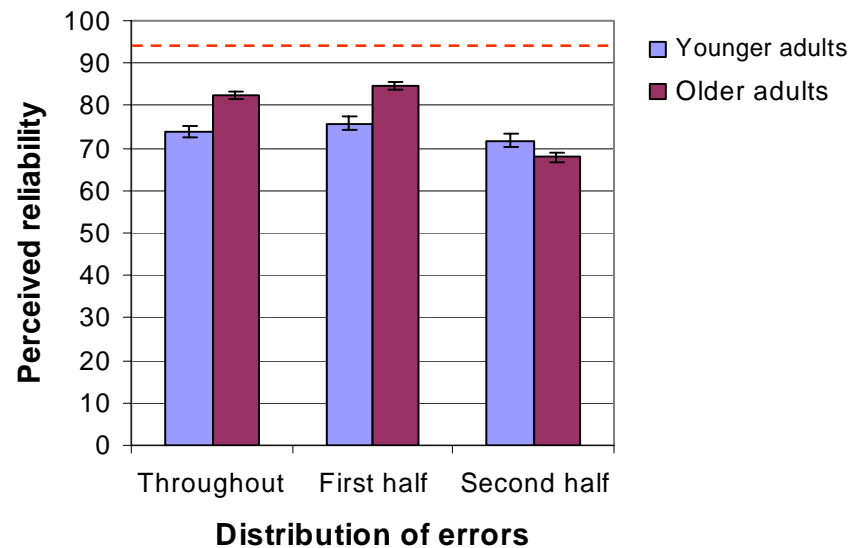


Figure 10. Perceived reliability ratings by age and Distribution of Errors. The dashed line indicates the actual reliability of the automation, which was 95.4%. Error bars represent the standard error of the mean.

The Effects of Error and Distribution of Errors on Reliance: A Micro-Level

Analysis

The results from the macro-level analysis showed that differences in reliance as a function of Error Type were dependent on the state of the system. The focus of this section was to explore in greater detail how behavior changes over time and as a function of Error Type and Distribution of Errors. The reliance data of the Younger Adults in

Experiment 2 are presented and discussed. The reliance data of the Older Adults are presented in the following section.

Figures 11 - 22 illustrate the changes in the percentage of participants who relied on the automation over the entire experiment as a function of Error Type and the state of the automation (alarm and non-alarm periods). A best fit regression line was plotted for each Error Type Condition during the first 30 minutes (bins 0 – 120) and during the last 30 minutes (bins 120 – 240) of the experiment. The y-axes in Figures 11 – 22 represent the percentage of participants who pressed the spacebar across each of the 240 time bins (x-axes). In the figures that illustrate the number of participants who pressed the spacebar during alarms there are a total of 50 bins in which data points are plotted (50 points for each False Alarms and Misses).⁴ In the figures that illustrate the number of participants who pressed the spacebar during non-alarms there are a total of 180 bins in which data points are plotted (False Alarms and Misses).

Because there were 10 participants in each experimental group, a value of 100% on the y-axes means that 10 of 10 participants pressed the spacebar during a specific time bin. The Non-Parametric Kruskal-Wallis test was used to analyze the effects of Error Type on the percentage of participants who relied on the automation during alarms and non-alarms. These analyses were performed for both the first and the last 30 minutes of the experiment.

⁴ Only correct detections (true alarms) are plotted. False alarms are omitted because the corresponding errors were misses, and the objective of this analysis was to compare behavior as a function of Error Type in the presence and absence of alarms.

Throughout Condition

Figure 11 illustrates the percentage of younger adults who pressed the spacebar during alarms and Figure 12 during non-alarms. Both figures illustrate data from the condition in which all of the automation errors were distributed throughout the experiment. As expected, reliance on the automation differed as a function of Error Type and the state of the automation. During alarms (Figure 11), participants in both the False Alarms and Misses Conditions had similar levels of non-reliance for the first 30 minutes of the experiment, $\chi^2 = 2.17$, $\eta^2 = .04$, $p > .05$. However, during the last 30 minutes, distinct patterns of behavior emerged. Participants in the Misses Condition began to rely more on the automation than participants in the False Alarms Condition, $\chi^2 = 30.16$, $\eta^2 = .62$, $p < .05$. This difference suggests that after approximately 30 minutes, most participants in the Misses Condition began to adjust their behavior in accordance to the criterion of the automation. Participants in the Misses Condition realized that it was not necessary to frequently check the validity of the alarms by pressing the spacebar and viewing the outside window because the errors by the automation up to that point had consisted of only misses and automated alarms up to that point had been perfectly reliable. However, most participants in the False Alarms Condition continued to frequently press the spacebar during alarms. These results suggest that the criterion of an automated system can have an important impact on the psychological meaning of an alarm.

The percentage of younger adults who pressed the spacebar during non-alarms is illustrated in Figure 12. The pattern of behavior as a function of Error Type observed

during non-alarms was the opposite of the behavior observed during alarms.

Interestingly, the different patterns of behavior as a function of Error Type emerged during the first 30 minutes of the experiment, $\chi^2 = 19.25$, $\eta^2 = .11$, $p < .05$, and remained stable during the last 30 minutes, $\chi^2 = 1108.08$, $\eta^2 = .60$, $p < .05$. The majority of participants in the Misses Condition continued to press the spacebar throughout the experiment while most participants in the False Alarms Condition had adjusted their behavior toward the end of the first 30 minutes to better match the criterion of the automation.

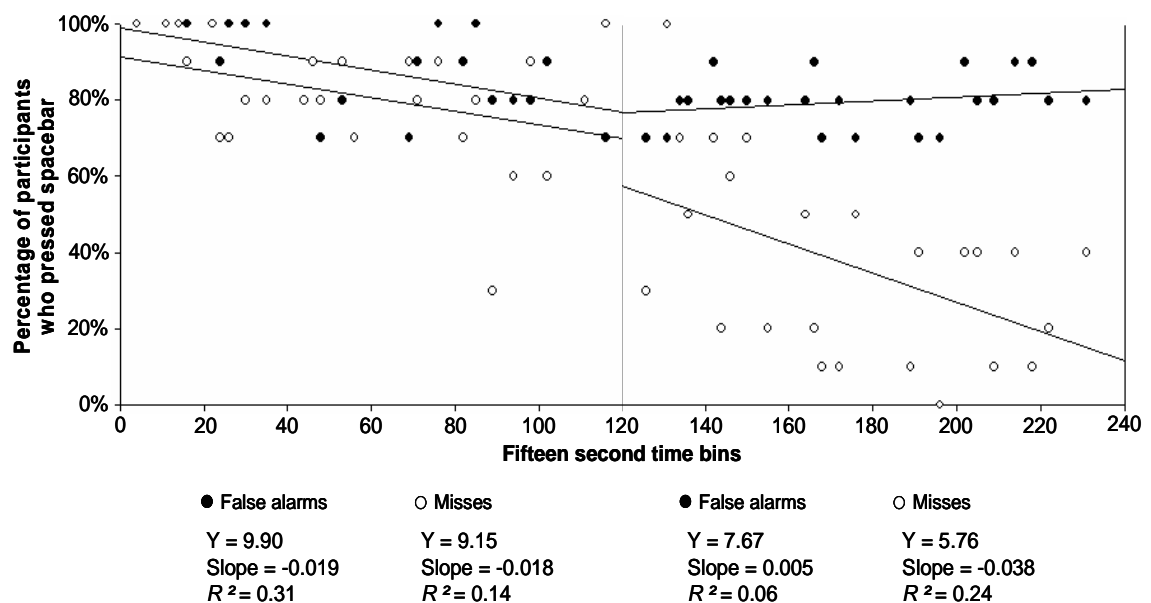


Figure 11. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Younger Adults who pressed the spacebar **during alarms** in the Throughout Condition. [(Y-intercept and slope) x 10].

The earlier emergence of different patterns of reliance during non-alarm periods relative to alarm periods suggests a clear difference in the manner in which humans interact with a system as a function of the state of the automation. Even though the automation is providing support while it is idle (during non-alarms) by letting the human know that there are no obstacles, an alarm is a more salient way for an automated system to provide support. The salience of alarms relative to non-alarms is a feasible explanation for the delayed shift in behavior toward reliance on alarms by participants in the Misses Condition (See Figure 11). Because non-alarms are less salient, participants in the False Alarm Condition were more inclined to begin to rely on the automation earlier, but only during non-alarms (See Figure 12).

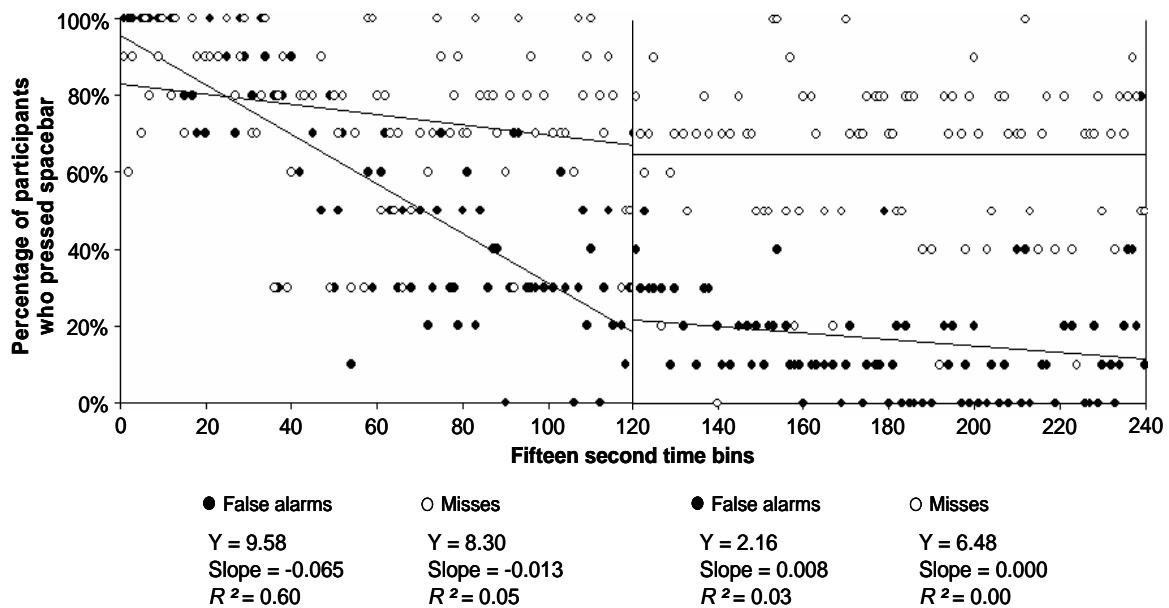


Figure 12. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Younger Adults who pressed the spacebar **during non-alarms** in the Throughout Condition. [(Y-intercept and slope) x 10].

Another interesting outcome of this analysis was that in three of the four sets of regression lines in Figures 11 and 12, the R^2 values for the False Alarms regression lines were higher than the R^2 values for the Misses regression lines. The fact that most of the False Alarms lines have a better fit suggests that the reliance behavior across participants in that condition was less variable than the behavior of the participants in the Misses Condition. The implication of this trend is that the emergence of reliance behavior by humans who interact with a system in which false alarms are prevalent is more predictable relative to a system in which misses are more common.

The pattern of behavior illustrated in Figures 11 and 12 provides an important insight into the relationship between automation reliability and human behavior. To date, the goal of various research efforts has been to understand how different levels of reliability affect trust-related behaviors, such as reliance (e.g., Moray et al., 2000; Parasuraman et al., 1993; Sanchez et al., 2004; Wiegmann et al., 2001). The results of the current investigation show that reliance on automation is not simply shaped by the automation's reliability (the percentage of errors it makes), but also by its characteristics, specifically the prevalence of each error type. These results imply that in a system in which most errors consist of false alarms, one could argue that reliance on the automation is likely to be high, as long as the automation is idle. Conversely, when the automation is providing support information in the form of an alarm, reliance on the automation is likely to be low. This contrast in behavior, which is dependent on the state of the system, highlights the complexity of the construct of trust in automation and has important theoretical and practical implications.

From a theoretical standpoint, it is clear that changes in human behavior as a result of automation errors are not random. Instead, changes in behavior appear to be systematic responses to the characteristics of the automation in an effort to more effectively collaborate with it. The prevalence of each type of error by the automation also has important effects on the allocation of attention by the human.

From a practical standpoint, the current results suggest that it is critical to set realistic expectations when defining the role of a human in any system that requires collaboration with automation. For example, the prevalence of each type of error can have different effects on the workload of the human. In a system in which false alarms are prevalent, workload will likely be higher during alarm periods because the human has to constantly check the validity of the alarms. Therefore, the human should not be expected to attend to many other stimuli during alarms.

First Half Condition

Figure 13 illustrates the percentage of Younger Adults who relied on the automation during alarm periods and Figure 14 during non-alarm periods. Both Figures illustrate data from the condition in which all of the automation errors were distributed within the first half of the experiment. Interestingly, the pattern of data illustrated in Figure 13 closely resembles the pattern in Figure 11 (Throughout Condition). During alarms, most participants in both the False Alarms and Misses Conditions did not rely on the automation during the first 30 minutes of the experiment, $\chi^2 = .70$, $\eta^2 = .01$, $p > .05$. However, during the last 30 minutes, more participants in the Misses Condition began to

rely on the automated alarms than in the False Alarms Condition, $\chi^2 = 14.76$, $\eta^2 = .30$, $p < .05$. Because of the greater concentration of errors by the automation during the first half of the experiment, it was expected that the distinct patterns of behavior as a function of Error Type would have emerged earlier than in the condition in which the errors were distributed throughout the experiment. Instead, the different patterns did not emerge until the second half of the experiment, when the automation was perfect.

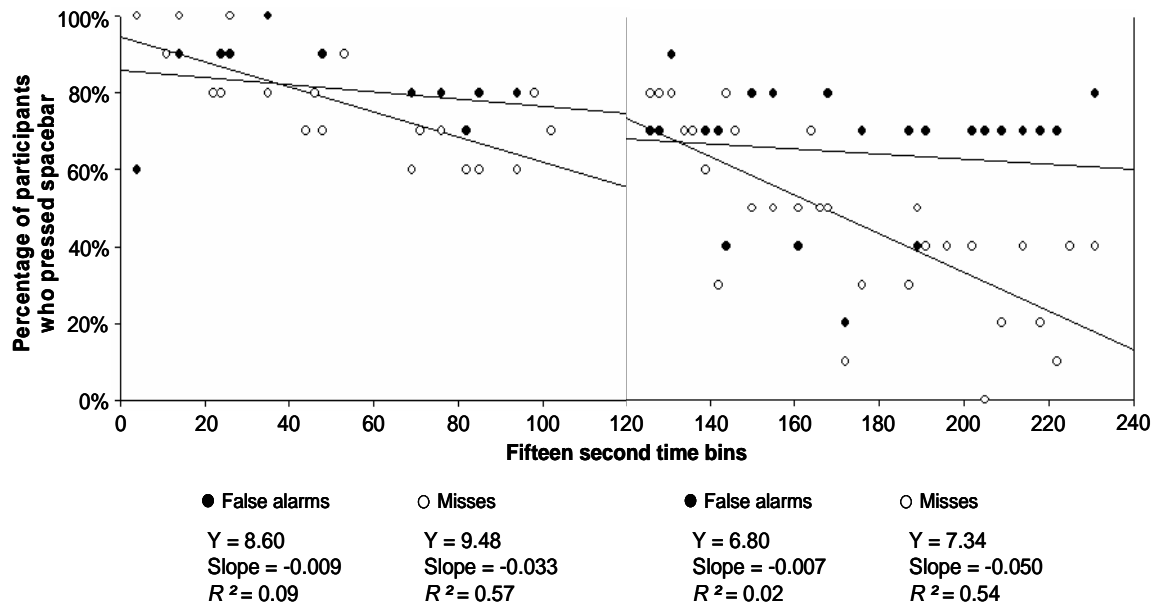


Figure 13. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Younger Adults who pressed the spacebar **during alarms** in the First Half Condition. [(Y-intercept and slope) x 10].

The results illustrated in Figure 13 suggest that a greater concentration of automation errors, at least to the degree in which it was changed with the Distribution of Errors manipulation, does not significantly impact the effects that error types have on the

reliance behavior during alarms. The emergence of different patterns of behavior during the last 30 minutes of the experiment as a function of Error Type does suggest that once humans alter their behavior to better match the characteristics of the automation there is a lasting effect, even in the presence of perfect automation.

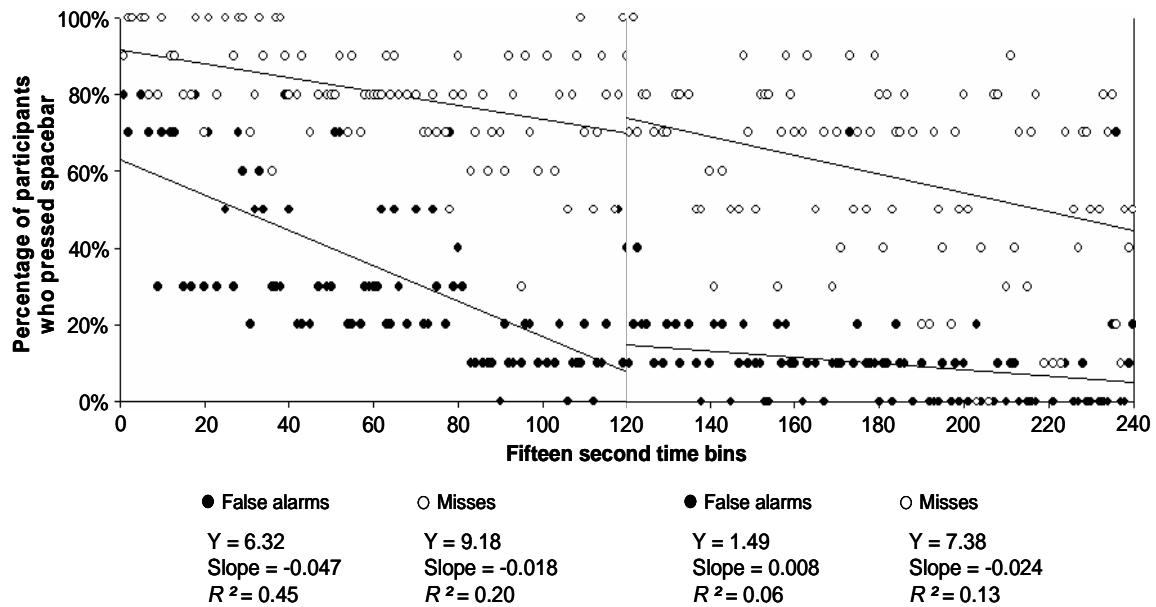


Figure 14. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Younger Adults who pressed the spacebar **during non-alarms** in the First Half Condition. [(Y-intercept and slope) x 10].

Surprisingly, the pattern of behavior of the participants in the False Alarms Condition during the last 30 minutes of the experiment (Figure 13) does not suggest a strong trend toward reliance on the automation. The results of previous studies show that humans are likely to begin to trust and rely on automation shortly after an error occurs (Lee & Moray, 1992; Riley 1996). However, the current results suggest this tendency toward reliance is not evident when humans have learned through experience that the

automation has a tendency to generate false alarms and are faced with the decision to rely or not rely on automated alarms.

Figure 14 illustrates the reliance data for non-alarm periods. The pattern of data observed in the first 30 minutes of Figure 14 resembles the pattern observed in the first 30 minutes of Figure 12, which was the Throughout Condition. During the first 30 minutes, more participants in the False Alarms Condition began to rely on the automation than in the Misses Condition, $\chi^2 = 98.87$, $\eta^2 = .56$, $p < .05$. The similarity between the trends observed in figures 12 and 14 suggests that the concentration of errors did not affect the rate at which reliance changed as a function of Error Type during non-alarms. However, during the last 30 minutes, when the automation was perfect, the negative slope of the best fit line for the Misses Condition indicates a trend toward the recovery of trust in the automation, manifested through reliance (Figure 14). This trend toward reliance by participants in the Misses Condition provides some evidence that when transitioning from unreliable to perfect automation, the recovery of trust is more likely to be observed first in the reliance behavior during non-alarms and by those who are interacting with a system laden with misses. Even with the trend toward reliance by participants in the Misses group during the last 30 minutes, the different patterns of reliance remained different as a function of Error Type, $\chi^2 = 109.71$, $\eta^2 = .61$, $p < .05$.

Second Half Condition

Figure 15 illustrates the percentage of younger adults who relied on the automation during alarm periods and Figure 16 during non-alarm periods. Both Figures

illustrate data from the condition in which all of the automation errors were distributed within the last 30 minutes of the experiment (Second Half Condition). The first 30 minutes graphed in Figures 15 and 16 provide an illustration of how reliance changes across time with perfect automation. As expected, the negative slopes of all four lines in Figures 15 and 16 indicate an increase in the percentage of participants who relied on the automation across time. It is worth noting that while Error Type did not have a significant effect on the percentage of participants who relied on alarms during the first 30 minutes, more participants in the False Alarms Condition pressed the spacebar during the first 30 minutes during non-alarms, $\chi^2 = 20.54$, $\eta^2 = .12$, $p < .05$. This difference as a function of Error Type was not expected during the first 30 minutes, when the automation was perfect for all groups.

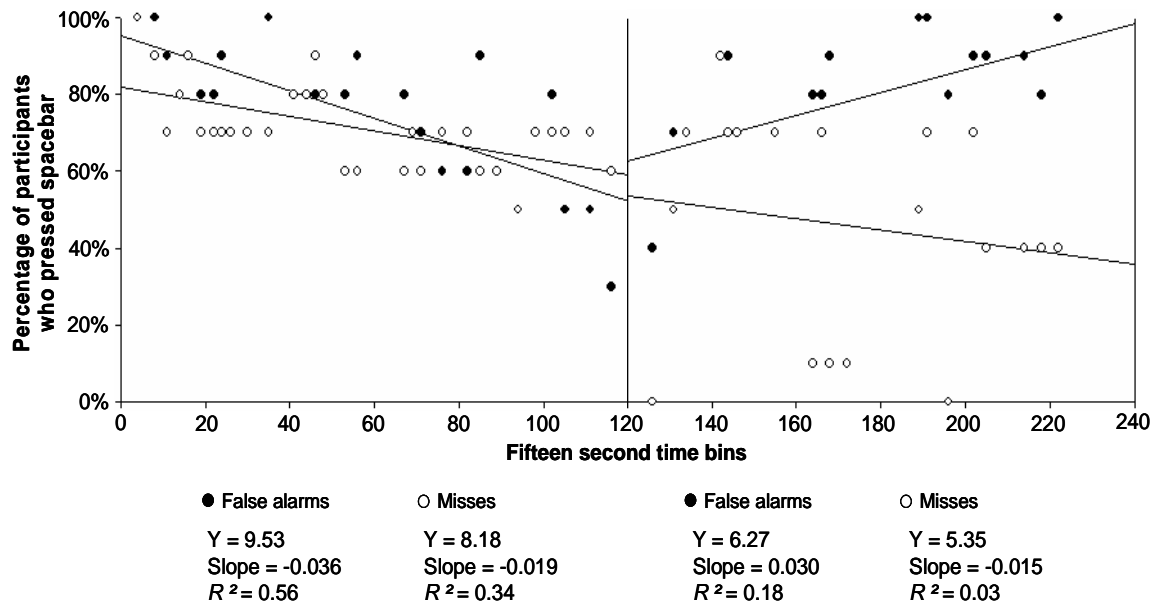


Figure 15. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Younger Adults who pressed the spacebar **during alarms** in the Second Half Condition. [(Y-intercept and slope) x 10].

Interestingly, during the first 30 minutes there was a noticeable difference in behavior as a function of the state of the automation. The trend toward reliance on the automation was greater during non-alarms than during alarms. The evidence for this trend is in the steeper slopes of the regression lines during non-alarms (-0.043 for False Alarms and -0.048 for Misses) relative to the slopes during alarms (-0.036 for False Alarms and -0.019 for Misses) as well as the greater Y-intercept values during alarms (9.53 for False Alarms and 8.18 for Misses) relative to the Y-intercept values during non-alarms (8.24 for False Alarms and 7.01 for Misses). The differences in the magnitude of the slopes and Y-intercept values suggests that even in the presence of perfectly reliable automation, most participants were not as willing to rely on alarms as they were to rely on non-alarms. Presumably, alarms were still salient enough to prompt most participants to double-check their validity, even when they had proven reliable for several minutes. Another interesting aspect of the data from Figures 15 and 16 is that the R^2 values of the lines from the first 30 minutes are higher than most of R^2 values from previous conditions in which automation errors were involved. This observed difference in R^2 values suggests that the reliance behavior of a group is more predictable as a function of time when the automation is perfectly reliable.

In general, the pattern of behavior observed in the last 30 minutes of Figures 15 and 16 was consistent with the pattern of behavior from previous conditions. During alarms (Figure 15), most participants in the False Alarms Condition did not rely on the automated alarms, while most participants in the Misses Condition began to rely on them toward the end of the experiment, $\chi^2 = 14.3$, $\eta^2 = .29$, $p < .05$. During non-alarms (Figure 16), most participants in the False Alarms Condition began to rely on the automation,

while most participants in the Misses Condition began to check the outside view with increased frequency, $\chi^2 = 24.5$, $\eta^2 = .14$, $p < .05$.

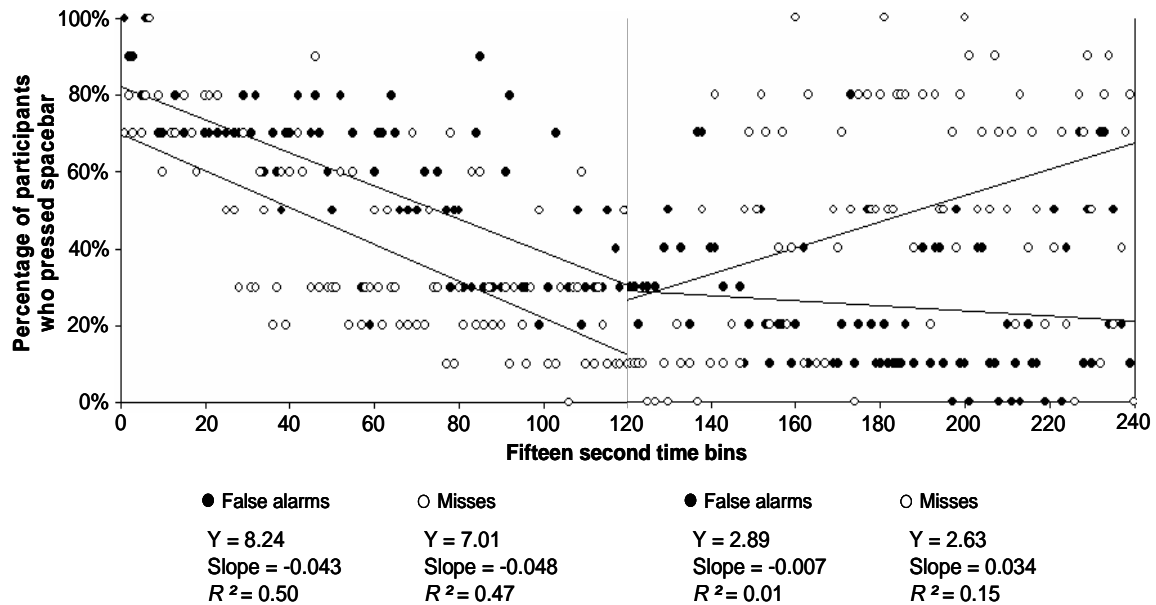


Figure 16. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Younger Adults who pressed the spacebar **during non-alarms** in the Second Half Condition. [(Y-intercept and slope) x 10].

A notable difference between the Second Half Condition and the Throughout and First Half Conditions was that during alarms, differences in reliance as a function of Error Type became apparent at different stages. In the Throughout and First Half Conditions, participants had been exposed to errors for approximately 30 minutes before different reliance patterns emerged. However, in the Second Half Condition (Figure 15), this difference in reliance behavior as a function of Error Type appears to emerge only after approximately 12 minutes (~bin 170) of being exposed to errors by the automation.

This finding suggests that there may be other factors, such as experience with the system, which affect the ability of humans to perceive the characteristics of the automation and adjust their behavior accordingly. In the Throughout and First Half Conditions, participants were exposed to automation errors from the beginning of the experiment while they were still trying to develop a mental representation of the automation. However, in the Second Half Condition they were allowed to interact with the system for 30 minutes before any errors were introduced. As errors appeared, greater familiarity with the system might have contributed to the quicker adjustment in behavior during alarms.

Age-related Effects on Reliance: A Micro-Level Analysis

Figures 17 – 22 illustrate the reliance data from the Older Adults in the same format that the reliance data from the Younger Adults was presented in the previous section. The objective of this section was to explore at a micro-level, how Error Type and the Distribution of Errors affect the reliance behavior of Older Adults across time. This analysis showed that there were some similarities in the patterns that older adults appear to develop when exposed to the same manipulations. However, one notable difference between the two Age groups was the differences in reliance on the automation during alarms.

Throughout Condition

Figures 17 and 18 illustrate the percentage of Older Adults in the Throughout Condition who pressed the spacebar during alarms and during non-alarms respectively. During alarms, neither the participants in the Misses nor those in the False Alarms Condition relied on the automated alarms consistently (Figure 17). During the first 30 minutes there were no differences in the percentage of participants who relied on the alarms as a function of Error Type, $\chi^2 = .26$, $\eta^2 = .01$, $p > .05$. However, during the last 30 minutes, more participants in the Misses Condition pressed the spacebar than in the False Alarms Condition, $\chi^2 = 11.45$, $\eta^2 = .23$, $p < .05$. While this difference was contrary to what was expected, the percentage of participants in the False Alarms Condition who pressed the spacebar during alarms in the last 30 minutes was consistently at or above 80% (See Figure 17). Therefore, the data suggest that most participants in both Error Type Conditions did not rely on alarms throughout the entire experiment.

Interestingly, the patterns of reliance as a function of Error Type were considerably different within the Younger Adult group, where participants in the Misses Condition began to rely on the automated alarms after approximately 30 minutes. These results are consistent with the results of the macro-level analysis, which showed that Older Adults relied significantly less on alarms than Younger Adults. One explanation for the different patterns of reliance during alarms as a function of age is older adults' aversion to engaging in risk taking behavior by relying on a system that they know could be wrong. However, in a high workload, dual-task environment, such as the one used in this experiment, not relying on the automated alarms increases workload and can

negatively affect the performance in concurrent tasks (e.g., the tracking task). Therefore, relying on the automated alarms would have been to their advantage.

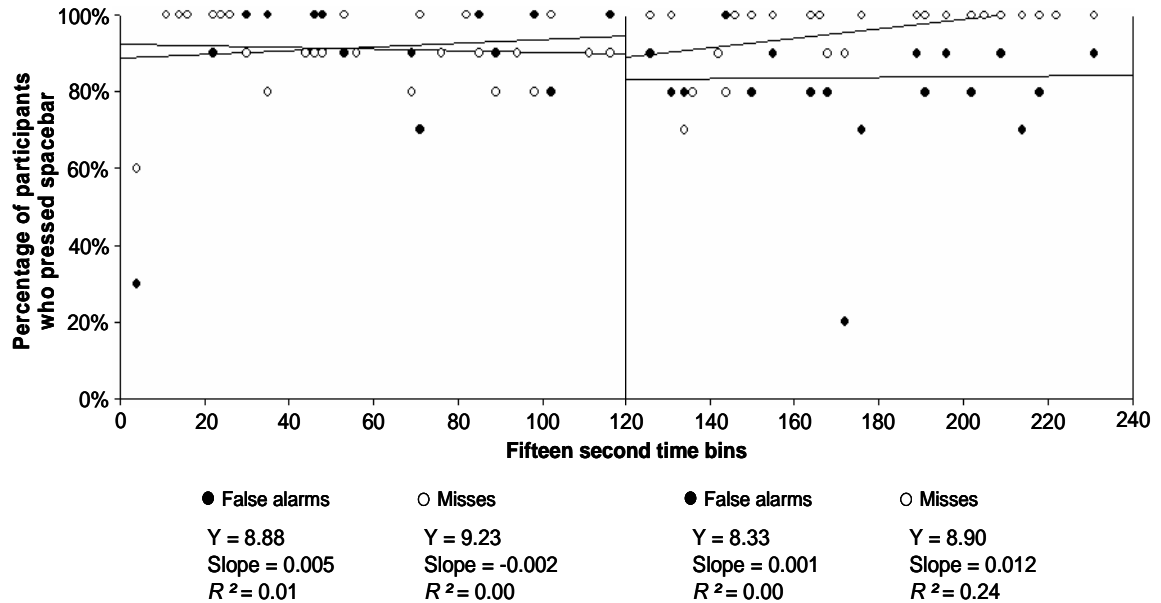


Figure 17. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Older Adults who pressed the spacebar **during alarms** in the Throughout Condition. [(Y-intercept and slope) x 10].

During non-alarms the Older Adults in the False Alarms condition relied more on the automation than those in the Misses Condition within the first 30 minutes, $\chi^2 = 92.1$, $\eta^2 = .53$, $p < .05$, and within the last 30 minutes, $\chi^2 = 106.21$, $\eta^2 = .59$, $p < .05$ (Figure 18). This pattern was similar to the one observed for the Younger Adult group, which suggests that at least for non-alarms, Older Adults did adjust their behavior according to the detection criterion of the automation. Also, Older Adults demonstrated willingness to rely on an automated system that had proven, at times, to be unreliable. Therefore, a

possible explanation for the age-related differences in the pattern of reliance during alarms is that the saliency of alarms had a more pronounced effect on older adults, making it more challenging for older adults to discern the characteristics of the automation in the presence of alarms.

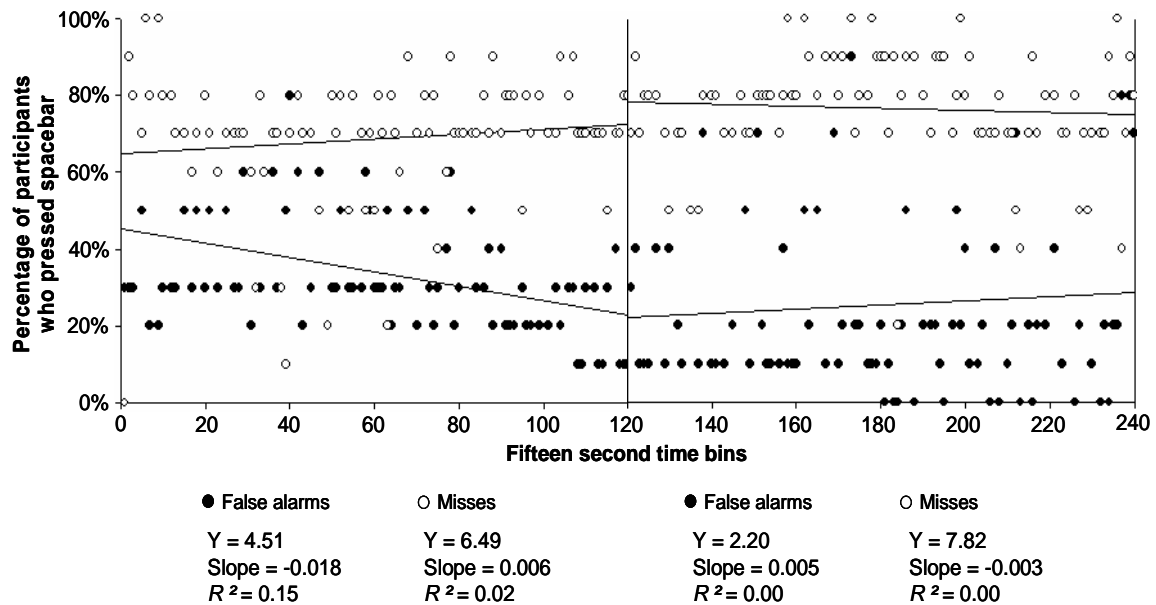


Figure 18. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Older Adults who pressed the spacebar **during non-alarms** in the Throughout Condition. [(Y-intercept and slope) x 10].

First Half Condition

Figures 19 and 20 illustrate the percentage of Older Adults in the First Half Condition who pressed the spacebar during alarms and during non-alarms, respectively. Similar to the results from the Throughout Condition, the pattern of reliance during

alarms did not differ as a function of Error Type (Figure 19). Therefore, it appears that a higher concentration of errors during the first 30 minutes did not have an impact on the development of distinct patterns of behavior as a function of Error Type. The higher concentration of errors during the first 30 minutes did not impact the behavior of Younger Adults either. However, unlike the behavior of Younger Adults, during the last 30 minutes of the experiment, when the automation was perfect, there was not a clear trend toward reliance by the Older Adults as was expected.

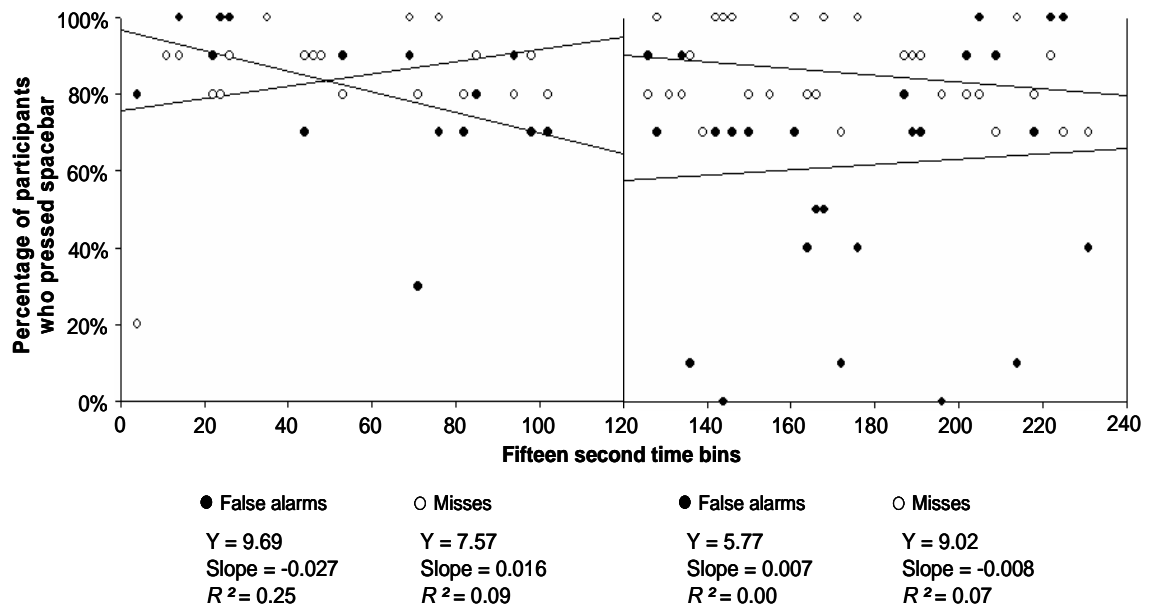


Figure 19. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Older Adults who pressed the spacebar **during alarms** in the First Half Condition. [(Y-intercept and slope) x 10].

This lack of a trend toward reliance during the last 30 minutes of the experiment was not consistent with the higher reliability estimates reported by Older Adults in the

First Half Condition. The lack of consistency between the reliance behavior of older adults and their estimates of reliability suggests that in some cases objective behaviors and subjective estimates of reliability are loosely coupled. It is also possible that subjective estimates of reliability are predictive of future reliance patterns. Therefore, with a longer exposure to the task Older Adults' behavior might have moved toward reliance.

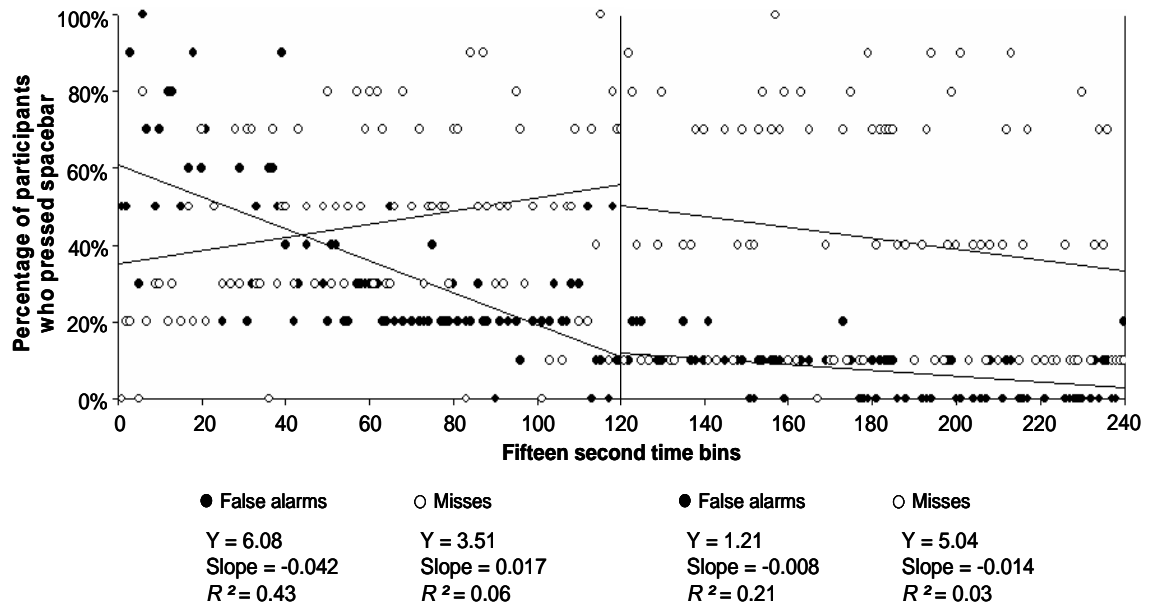


Figure 20. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Older Adults who pressed the spacebar during **non-alarms** in the First Half Condition. [(Y-intercept and slope) x 10].

During non-alarms, most Older Adults in the False Alarms Condition began to rely on the automation toward the end of the first 30 minutes, $\chi^2 = 11.05$, $\eta^2 = .06$, $p < .05$ and continued to rely on it during the last 30 minutes, $\chi^2 = 76.2$, $\eta^2 = .43$, $p < .05$ (Figure

20). Conversely, Older Adults in the Misses Condition were not as likely to rely on the automation throughout the experiment, although during the last 30 minutes there does appear to be a trend toward reliance. The pattern of behavior observed during the last 30 minutes of the non-alarm periods resembled the pattern of behavior of the Younger Adults.

Second Half Condition

Figures 21 and 22 illustrate the percentage of Older Adults in the Second Half Condition who pressed the spacebar during alarms and during non-alarms, respectively. The pattern of behavior illustrated in the first 30 minutes of Figures 21 and 22 indicates that with perfect automation, reliance increased as a function of time. Similar to the behavior of younger adults, the trend toward reliance during the first 30 minutes was greater during non-alarms than during alarms. This finding provides more support for the idea that reliance on automation needs to be examined as a function of the state of the system. Alarms, because of their greater salience, are less conducive to absolute reliance than non-alarms.

During alarms, the pattern of behavior within the last 30 minutes of the experiment was similar to the pattern of behavior of the Younger Adults, although within the False Alarms Condition, Older Adults relied less on the automation than Younger Adults. Most Older Adults in the False Alarms Condition did not rely on the alarms, while participants in the Misses Condition continued to rely on the automation, $\chi^2 = 26.87, \eta^2 = .55, p < .05$. Interestingly, this condition was the only one in which the Older

Adults exhibited different patterns of behavior during alarms as a function of Error Type (Figure 21). This behavior suggests that older adults, when given the opportunity to interact with a system under optimal conditions, appear to build enough trust in the automation and are able to discern the characteristics of the automation and adjust their behavior accordingly.

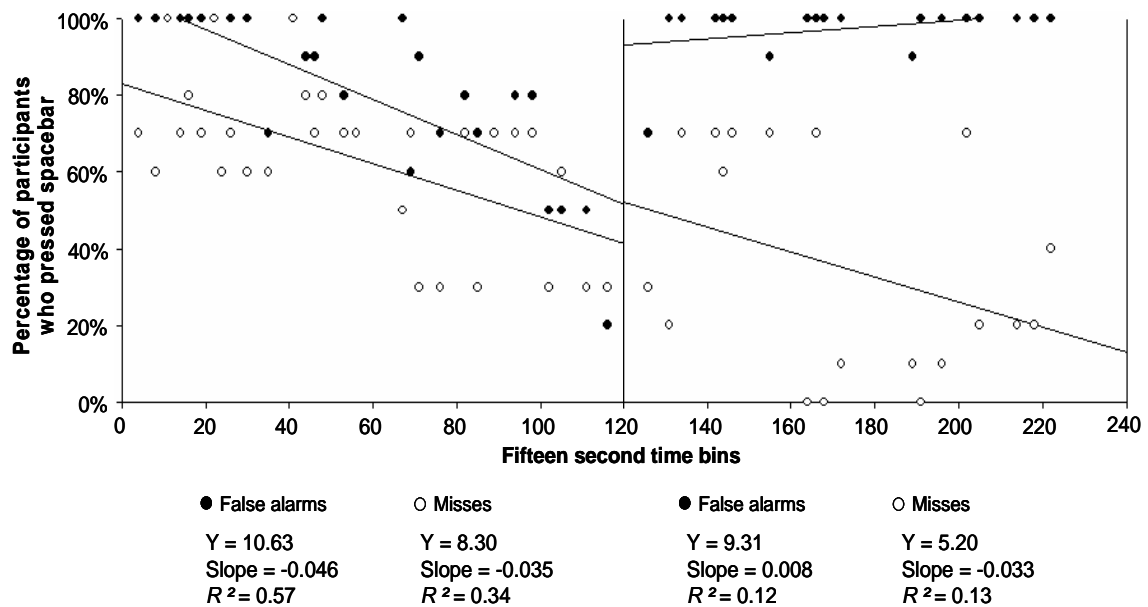


Figure 21. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Older Adults who pressed the spacebar **during alarms** in the Second Half Condition. [(Y-intercept and slope) x 10].

Another interesting aspect related to the behavior of the Older Adults during the first 30 minutes of the Second Half Condition is that the R^2 values of the best fit lines were higher than for most of the conditions in which there were automation errors. This trend was also observed with the Younger Adults. This finding suggests that in the

presence of perfect automation, the reliance behavior across older adults is more predictable than when the automation is not perfectly reliable.

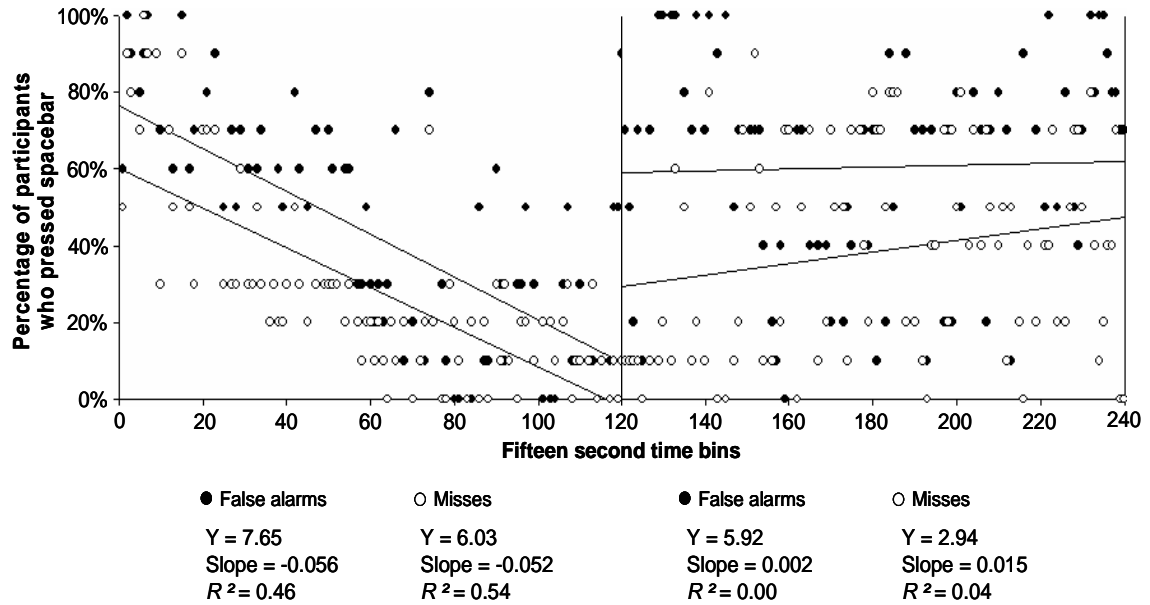


Figure 22. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Older Adults who pressed the spacebar during **non-alarms** in the Second Half Condition. [(Y-intercept and slope) x 10].

The Effects of Agricultural Experience on Reliance and Trust

For this analysis, the behavior and performance of the Younger Adult group in Experiment 2, who did not have experience operating agricultural vehicles (Non-Farmers), was compared to the behavior and performance of the participants from Experiment 1 (Farmers). Therefore, the macro-level analysis was treated as a 2 (Agricultural Experience: Farmers, Non-Farmers) x 2 (Error Type: False Alarms, Misses)

between-subjects experiment. Other than the results reported in this section, there were no other main effects of agricultural experience on any of the dependent variables.⁵

Participant Characteristics

The Non-Farmer groups performed significantly better than the Farmer group in the Digit Symbol Substitution test, $F(1, 34) = 16.8$, $\eta^2 = .31$, $p < .05$, the Shipley Vocabulary test, $F(1, 34) = 4.7$, $\eta^2 = .11$, $p < .05$, and the Paper Folding Test. As expected, Farmers had more years of experience with agricultural vehicles, $F(1, 34) = 104.6$, $\eta^2 = .73$, $p < .05$. Farmers were also significantly older than Non-Farmers, $F(1, 34) = 24.1$, $\eta^2 = .40$, $p < .05$.

Spacebar Presses

There were main effects of Error Type on first spacebar presses, spacebar presses during alarms, and spacebar presses during non-alarms. These results were consistent with the results from Experiment 2. Overall, participants in the Misses Condition relied less on the automation than participants in the False Alarms Condition, $F(1, 34) = 7.3$, $\eta^2 = .17$, $p < .05$. Similarly, participants in the Misses Condition relied less on the automation during non-alarms than participants in the False Alarms Condition, $F(1, 34) = 16.4$, $\eta^2 = .31$, $p < .05$. During alarms, participants in the False Alarms Condition relied

⁵ For the ANOVA table with all of the comparisons by Age and Error Type see Appendix H

less on the automation than those in the Misses Condition, $F(1, 34) = 58.7, \eta^2 = .62, p < .05$.

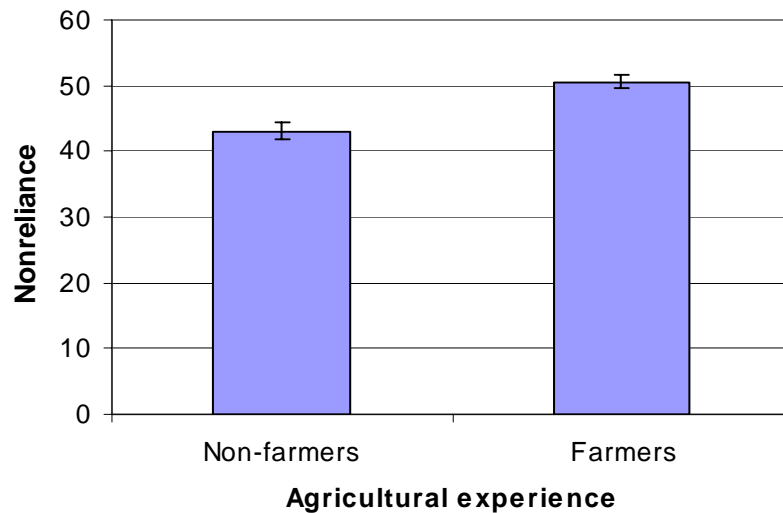


Figure 23. Non-reliance (first spacebar presses) by agricultural experience for all Farmers and Non-Farmers. Error bars represent the standard error of the mean.

While there were no differences as a function of agricultural experience on first spacebar presses, there was a main effect of agricultural experience on spacebar presses during alarms, where Farmers relied less on the automation than Non-Farmers, $F(1, 34) = 11.9, \eta^2 = .25, p < .05$ (Figure 23). This finding suggests that Farmers were less willing to rely on alarms from an automated system that had proven to be unreliable. Interestingly, there were no main effects of agricultural experience on subjective trust and perceived reliability. The lack of a difference suggests that subjective trust ratings and estimates of reliability, while they are usually related to overall reliance, they are not sensitive enough to predict reliance during specific states of the automation. A more in-

depth breakdown of the reliance data is presented in the following section, which examines reliance patterns of the Farmer group at the micro-level.

Micro-Level Analysis

Figures 24 and 25 illustrate the percentage of Farmers in the Throughout Condition who pressed the spacebar during alarms and during non-alarms, respectively. The main effect of agricultural experience on reliance during alarms found in the macro-level analysis appears to be a product of Farmers in the Misses Condition not adjusting their behavior to the criterion of the automation. The analysis of the Non-Farmers' reliance data during alarms, indicated that during the last 30 minutes of the experiment most participants in the Misses Condition began to rely on the alarms. However, this trend was not as obvious for the Farmer group.

Even though the percentage of Farmers in the False Alarms Condition who did not rely on alarms is significantly higher than the percentage in the Misses Condition during the first 30 minutes, $\chi^2 = 17.14$, $\eta^2 = .35$, $p < .05$, and during the last 30 minutes, $\chi^2 = 23.12$, $\eta^2 = .47$, $p < .05$, the data from Figure 24 show that most Farmers (70% or more) in the Misses Condition continued to double check the validity of the alarms throughout the experiment.

One explanation for the different patterns of reliance on alarms as a function of agricultural experience is that Farmers are accustomed to interacting with automated systems that are generally robust, which means that the automated systems seldom make errors, unless they are caused by a permanent malfunction. Therefore, any error by the

automation might have been perceived by the Farmers as a sign that the collision avoidance system was permanently damaged and therefore they were less willing to rely on the automated alarms. Another factor that could have contributed to this difference in behavior as a function of Agricultural Experience is that the perception of the cost of an error might have been inherently higher for Farmers. The collision of an agricultural vehicle with any object usually has a very high cost associated with it. Not only is the equipment expensive, but it usually means that the vehicle is disabled to some capacity during harvest or planting season, when time is of the essence.

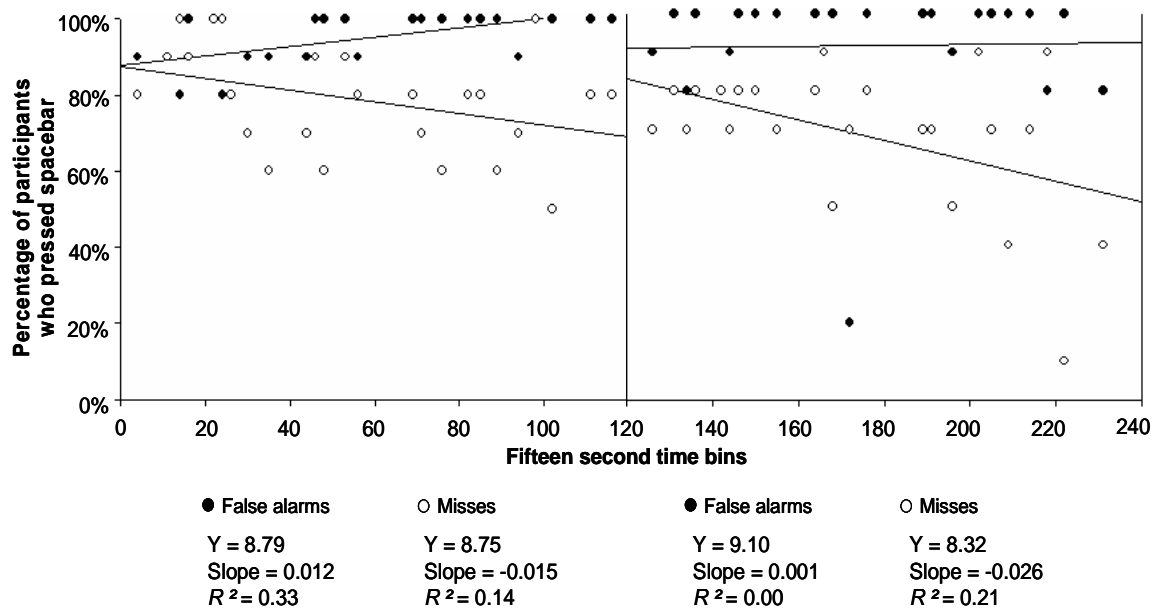


Figure 24. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Farmers who pressed the spacebar **during alarms**. [(Y-intercept and slope) x 10].

The reliance data from the non-alarm periods (Figure 25) show that different patterns of behavior did emerge as a function of Error Type, especially during the last 30 minutes, $\chi^2 = 66.6$, $\eta^2 = .37$, $p < .05$. This behavior suggests that Farmers are willing to rely on a faulty automated system during non-alarms. The results from Experiment 1 are consistent with the behavior observed in Experiment 2, where most participants were more reluctant to rely on automated alarms than they were to rely on automation during non-alarms.

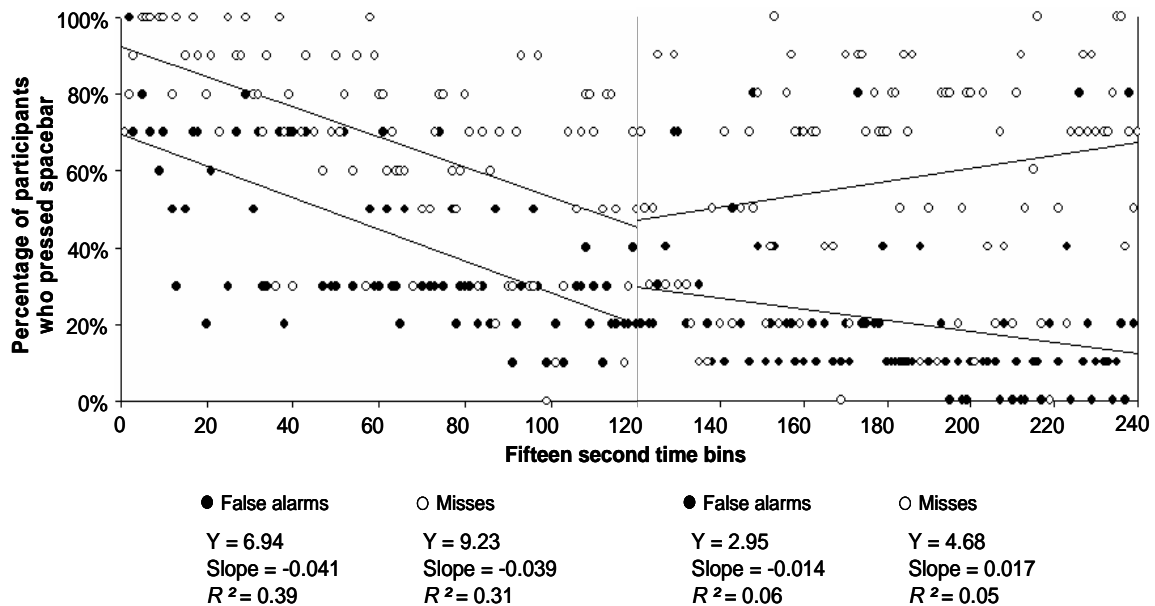


Figure 25. Percentage of participants who pressed the spacebar (y-axis) across each 15 second time bin (x-axis) as a function of Error Type. Each point represents the percentage of Farmers who pressed the spacebar **during non-alarms**. [(Y-intercept and slope) x 10].

Interestingly, the slopes of both lines during the first 30 minutes in Figure 25 (during non-alarms) suggest a trend toward reliance by participants in both the False

Alarms and Misses Conditions. This pattern toward reliance was not observed for the Non-Farmer group. The trend toward reliance of non-alarms by the Farmer group could be a product of prior experience. Most automated systems in agricultural vehicles have a loose criterion setting, which means that if the system does generate an error it is more likely to be a false alarm (A. Greer, personal communication, February, 2005).

Therefore, farmers are not used to unreliable events being in the form of automation misses, which is a possible reason why there is a trend toward reliance during non-alarms within the first 30 minutes of the experiment. However, with some exposure to the system, the different patterns of reliance became clearer (see the last 30 minutes in Figure 25). As expected, most participants in the False Alarms Condition continued the trend toward reliance on the automation, while those in the Misses Condition began to verify its validity with increased frequency.

The tendency of Farmers to not rely on the automation as much as Non-Farmers was not consistent with the results of Riley (1996). The results of his experiment showed that pilots, who were the experimental group with more automation experience, were more likely to rely on faulty automation than non-pilots (college undergraduates). Riley's explanation for this finding was that the costs associated with errors in the experimental task were considerably less than the costs of an error in the flight deck. Furthermore, the automation supported task was distinguishing letters from numbers. Perhaps this task did not invoke the tendencies and biases of pilots associated with the use of automation. Given the apparent lack of ecological validity of the experimental task, it is not surprising that pilots' experience with automation in the flight deck did not transfer into their interaction with the automation in the experiment.

Detection of the Exception Error

The last automation error in each experimental condition was the opposite of the rest of the errors. The main purpose for including the exception error was to contribute to the understanding of the effects that the prevalence of each Error Type have on the probability of “catching” an automation error. If reliance on the automation is affected by the prevalence of an Error Type in the way in which the data of this investigation have shown, then in the event that an uncommon automation error occurs, the probability of a collision should be higher than if the non-reliance behavior of humans was simply random. The combined results from Experiments 1 and 2 showed that 123 out of 140 participants failed to detect the exception error by the automation, which means only 12.1% of the exception errors were detected and did not result in a collision. This percentage was considerably lower than the percentage of the rest of the automation errors that were detected (54%).

The considerably lower detection rate of the exception error, relative to the rest of the errors, provides more evidence that humans adjust their behavior based on the criterion of the automation. This finding has an important practical implication. Once human behavior is changed due to the prevalence of a particular error, there is a lower probability that the human will detect an uncommon automation error. For example, in a system in which false alarms are prevalent, the human will likely have adjusted his/her behavior such that reliance on the automation during non-alarm periods is high. Therefore, in the event of an automation miss, it is not likely that the human will serve an effective backup role to the automation.

Effects of First and Subsequent Failures on Reliance

To determine the effects of each error on reliance, the number of participants who pressed the spacebar in the fifteen second bin after every automation error was subtracted from the number of participants who pressed the spacebar in the bin before the error. Figures I1 – I6, in Appendix I, illustrate changes in non-reliance for each error. Overall, the results of this analysis did not provide clear evidence for drastic changes toward non-reliance after each error. It appears that the trend toward non-reliance as result of exposure to errors was more gradual and therefore was not apparent through an analysis of such narrow scope (Figures I1 – I6). Riley (1996) also found that “automation use after each failure was not less than before the failure” (p. 26).

There is however, partial evidence for the possible existence of the first error effect. Wickens and Xu (2002) argued that when the first automation error is preceded by an extensive period of perfect automation, the expectations of the human have a tendency to increase. Therefore, the appearance of the first error can have a considerable impact on reliance and trust. In an analysis of the Second Half Condition, where the first error appeared after the automation had been perfect for 30 minutes, there was a notable change toward non-reliance after the first error (Table I1, Appendix I). However, the first error effect was not as apparent in the Throughout and First Half Conditions. An explanation for the lack of the first error effect in the Throughout and First Half Conditions was the low expectations of the automation prior to beginning the experiment, which were evident by the high percentage of participants who pressed the spacebar during the first minute of the experiment (See Figures I1 – I2).

Tracking Task Performance

One of the benefits of receiving support from an automated system is that it can reduce the workload for a specific task while freeing up cognitive and physical resources. In a multi-task environment, these resources can be allocated to performing concurrent tasks (Wickens & Xu, 2002). Therefore, human performance in tasks without automation support is often an indicator of reliance on the automation. If the human trusts and relies on the automation, performance on concurrent tasks should be higher than if the human has to constantly use time and resources to verify the validity of the support provided by the automation.

The simulation used in the current investigation required participants to perform two tasks. The first was the collision avoidance task which was supported by an automated system. The second task was a tracking task, which was performed manually. The simulator was configured so that participants were forced to abandon one task to perform an action on the other task. For example, if someone wanted to view the outside window, the buttons necessary to perform the tracking task would cease to work while the spacebar was being pressed.

The main measure of performance for the tracking task was the “number of acres lost due to tracking error.” This measure was calculated by adding the total amount of time that the square was left on the edge of the box and dividing it by 15. In study 2, participants in the Misses Condition committed significantly more errors than participants in the False Alarms Condition, $F(1, 108) = 7.6, \eta^2 = .07, p < .05$. Furthermore, Older Adults performed significantly worse than Younger Adults, $F(1, 108) = 63.9, \eta^2 = .37, p$

< .05. This age-related difference in tracking performance might have been due in part to the increased non-reliance on the automation during alarms by Older Adults.

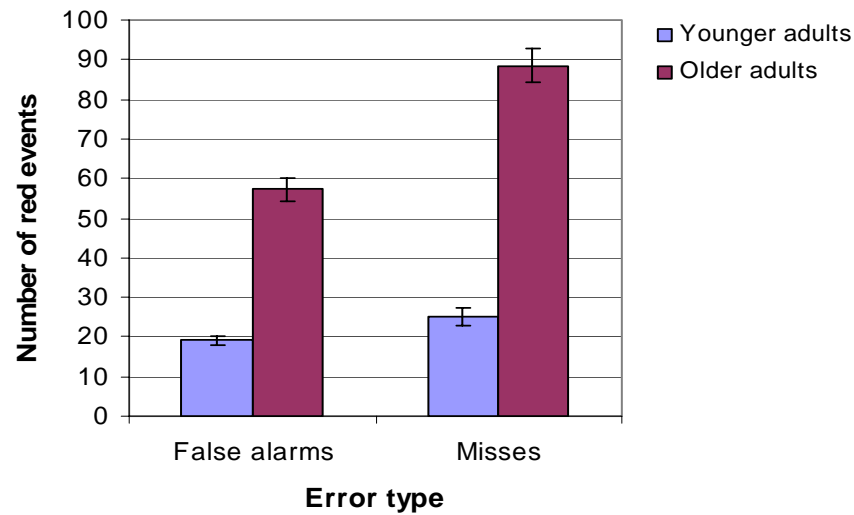


Figure 26. Tracking task performance (red events) by Age and Error Type. Error bars represent the standard error of the mean.

Another performance measure for the tracking task was the number of times that participants allowed the square to reach the edge; these instances were labeled “red events.” The advantage of this measure relative to the number of acres lost due to tracking error was that it accounted for all instances in which the square was allowed to reach the edge. In Experiment 2, there was a main effect of Error Type on number of red events where participants in the Misses Condition allowed significantly more red events than participants in the False Alarms Condition, $F(1, 108) = 10.6$, $\eta^2 = .09$, $p < .05$. There was also a main effect of Age, where Older Adults allowed more red events than Younger Adults, $F(1, 108) = 78.1$, $\eta^2 = .42$, $p < .05$. However, there was a significant

interaction between Age and Error Type, $F(1, 108) = 4.9, \eta^2 = .04, p < .05$ (See Figure 26). For the Older Adults, the prevalence of Misses had a significantly negative impact on the tracking task performance relative to False Alarms. This difference as a function of Error Type was not present for the Younger Adults. A correlation analysis also showed that there was a positive relationship between the number of red events and the number of acres lost to tracking error, $r(120) = .95, p < .01$.

The results of the tracking performance data were consistent with the pattern of results obtained for participants' reliance on the automation, which showed that participants in the Misses Condition relied less on the automation than participants in the False Alarms Condition. This behavior likely contributed to their poor performance in the tracking task. These current results support the idea that not relying on the automation can increase workload, which can contribute to performance decrements in other tasks. However, the main question of interest in the analysis of the tracking task performance was does Error Type affect concurrent task performance? The answer to this question depends on the type of system that is being evaluated. In a system that generates alarms with a high frequency, the prevalence of automation false alarms, relative to misses, would likely result in worse performance for concurrent tasks. Conversely, in a system in which alarms are rare, the prevalence of automation misses would be more detrimental to performance in concurrent tasks. This does not mean that the decision to set the criterion of the automation should be based purely on trying to increase performance in concurrent tasks. However, having some knowledge of the effects of Error Type on concurrent task performance can help predict the types of situations when a human's attentional resources might be limited.

Obstacle Detection Performance

The performance measures used to evaluate obstacle detection by participants were total number of errors, false positive and false negative rates. The total number of errors consisted of any mistake that was made in the collision avoidance task. The maximum number of errors possible during the experiment was 240. The maximum number of errors caused by the automation was 11.

Number of Collision Errors

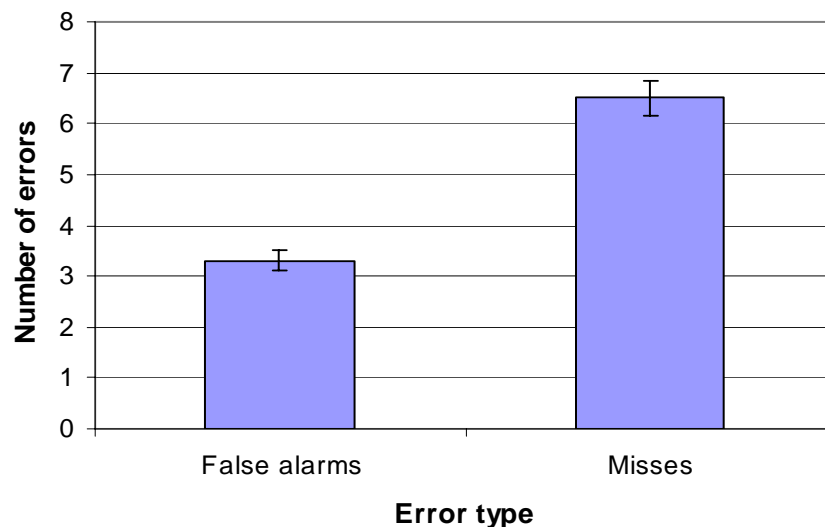


Figure 27. Total number of collision errors by Error Type. Error bars represent the standard error of the mean.

Older Adults committed more errors than Younger Adults, $F(1, 108) = 28.2, \eta^2 = .21, p < .05$, and participants in the Misses Condition made more errors than those in the False Alarms Condition, $F(1, 108) = 52.4, \eta^2 = .17, p < .05$ (Figure 27). Given the nature

of the task, it was not surprising that participants in the Misses Condition made more errors than participants in the False Alarms Condition. Participants in the Misses Condition had to constantly check the outside window during non-alarms if they wanted to avoid colliding with a deer, while participants in the False Alarms Condition only had to check the outside window during alarms. However, if participants had not adjusted their reliance behavior as a function of Error Type, it is likely that the average number of errors would have been much higher for both Error Type groups.

False Negative and Positive Rates

False positive and negative rates represent the human's ability to correctly "avoid" deer, irrespective of what the collision avoidance alarm was reporting. The reason for examining rates rather than total number of false positives and negatives independently is that the ratio of events with no deer and deer differed as a function of the Error Type manipulation (189:51 for the false alarm condition and 180:60 for the Misses Condition).

For false negative rate, there was a significant interaction between Age and Error Type, $F(1, 108) = 13.1, \eta^2 = .11, p < .05$. Older Adults had a higher false negative rate than Younger Adults within the Misses Condition but not within the False Alarms Condition. There was a main effect of Age, where Older adults had a higher false negative rate than Younger Adults, $F(1, 108) = 16.9, \eta^2 = .14, p < .05$. Furthermore, the Misses Condition led to a higher false negative rate than the False Alarms Condition, $F(1, 108) = 52.4, \eta^2 = .33, p < .05$.

Overall, the false positive rate was higher for Older Adults than for Younger Adults, $F(1, 108) = 27.5, \eta^2 = .20, p < .05$, and the False Alarms Condition also led to a higher false positive rate than the Misses Condition, $F(1, 108) = 14.7, \eta^2 = .12, p < .05$. It was not surprising that a prevalence of automation misses led to a higher rate of false negatives and more automation false alarms led to a higher rate of false positives. However, had humans not adjusted their reliance behavior as a function of the criterion setting of the collision avoidance system, it is expected that the difference in false positive and false negative rates as a function of Error Type would have been greater.

CHAPTER 4

DISCUSSION

The main objective of the current investigation was to gain an understanding of how specific variables affect human behavior in automated environments. The specific variables that were targeted in this investigation were ones that have either been overlooked in previous research efforts or have been the source of contradictory findings. A unique aspect of this investigation was the aim to understand how humans gather evidence about the characteristics of the automation and how this information affects behavior over time. An analysis of changes in behavior over time was critical to understanding the development of different patterns of reliance on the automation as a function of the experimental manipulations.

Does Error Type Affect Reliance on Automation?

Yes. One of the most influential findings from this investigation was the different patterns of reliance that emerged as a function of Error Type. Clearly, the detection criterion of an automated system has an impact on the reliance behavior of humans. When the unreliability of a system consists primarily of automation misses, humans adjust their reliance behavior accordingly to constantly monitor alternate sources of information while the automation is idle (during non-alarms). However, a prevalence of automation misses, while it leads to non-reliance on the automation during idle states, is

also conducive to reliance on automated alarms. Conversely, when a system is laden with false alarms, non-reliance behavior becomes prevalent in the presence of automated alarms. In this situation, the appearance of an alarm becomes a proxy to check alternate sources of information with the goal of verifying the validity of that alarm.

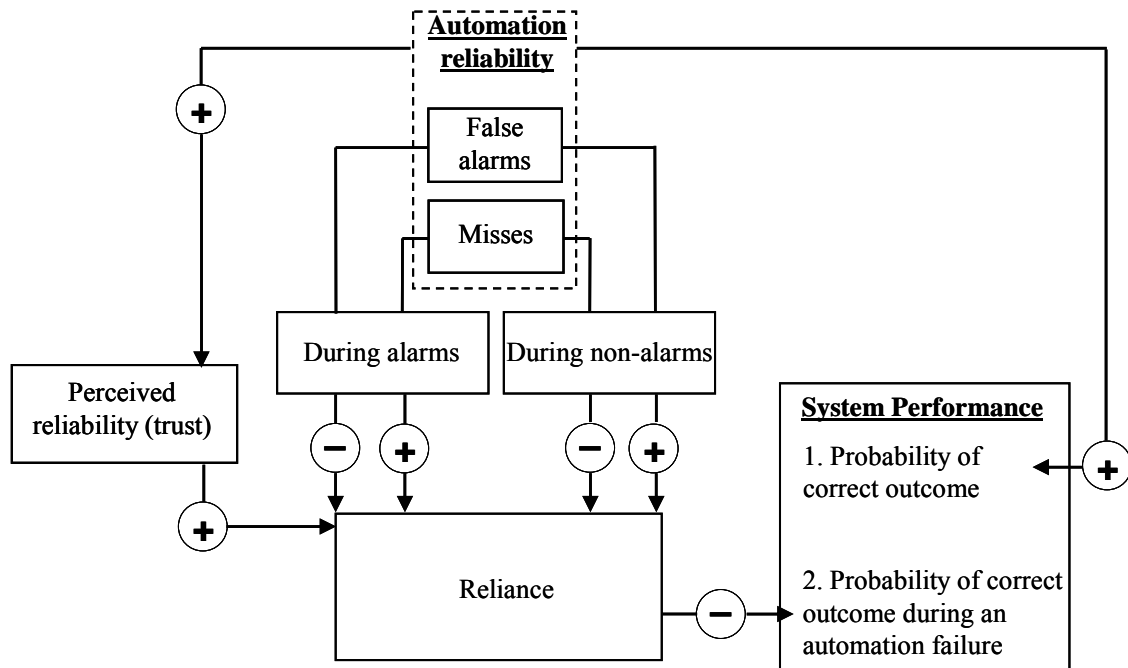


Figure 28. The relationship between variables specific to the automation, reliance, and system performance. The “+” sign indicates a positive relationship between two variables and the “-” sign indicates a negative relationship. The entire model can be found in Appendix J.

Figure 28 illustrates a conceptual model of the relationships between reliability, perceived reliability, reliance, and system performance. The “+” sign indicates a positive relationship between two variables and the “-” sign a negative one. For example, the relationship between automation reliability and perceived reliability is positive, which means that as automation reliability increases, perceived reliability also increases. The

different patterns of reliance that result as a function of different error types by the automation add a level of complexity to the relationship between automation reliability and reliance. Previous models of trust in automation (e.g., Dzindolet et al., 2001; Riley, 1996; Sanchez, 2005) have suggested that there are positive relationships between automation reliability, perceived reliability, and reliance. The positive relationship between automation reliability and reliance is usually based on the definition of reliability that treats it as a “consistency across measurements”, which is how reliability is generally defined in a statistical sense (Jackson & Messick, 1978). However, the findings of this investigation suggest that depending on the state of the system, false alarms and misses can each lead to both positive and negative relationships between reliability and reliance. Therefore, when describing the effects that automation reliability has on reliance, reliability can no longer be defined strictly as the rate of errors that are made by the automation (Wickens & Xu, 2002), but also the type of error that is prevalent need to be considered. An integrated version of the relationships illustrated in Figure 28 and the Sanchez (2005) model can be found in Appendix J.

Do Humans Adjust Optimally to Automation Error Type?

Perfect performance in the experimental task used in this investigation is comprised of zero errors (no acres lost) in both the collision avoidance task and the tracking task. No participant was able reach this level of performance. Hypothetically, if a participant had chosen to always rely on the automation, they would have ended with

11 errors in the collision avoidance task.⁶ Conversely, if someone had chosen to “never” rely on the automation by constantly pressing the spacebar, they *might* have achieved optimal performance on the collision avoidance task, but their performance on the tracking task would have suffered. The strategy to achieve perfect performance consisted of an “optimal balance”⁷ in the amount of time and effort spent pressing the spacebar and performing the tracking task.

The patterns of behavior that emerged as a function of Error Type suggest that humans adjust their behavior in an attempt to collaborate more effectively with the automation. In essence, humans calibrate their behavior according to the capabilities and limitations of the automation in an attempt to increase overall system performance. The identification of this self-adjustment in reliance makes an important contribution to the field of human-automation interaction, specifically to the idea of appropriate trust/reliance.

Appropriate trust and *appropriate reliance* are terms used to describe a match between the perceived capabilities of the automation by the human and the actual capabilities of the automation (Lee & Moray, 1994; Lee & See, 2004). For example, if a collision avoidance system has a high false alarm rate during rainy conditions, appropriate reliance would consist of frequently checking the validity of automated alarms when it is raining, while relying on the automated alarms when it is not raining. The results of this investigation indicate that humans adjust their behavior as a function of the types of errors by the automation, which suggests that humans can and do gather

⁶ No participant adopted this strategy

⁷ There is not an exact amount of time and effort that defines an “optimal balance,” rather “optimal balance” likely consists of a range of time and effort spent on each task. The exact point in that range depends on individual differences in the abilities required to perform the task.

information that helps them increase the instances of appropriate reliance on the automation. This trend toward appropriate reliance suggests that humans are constantly trying to adjust their behavior to achieve equilibrium between overall system performance and workload.

The relationships illustrated in Figure 28 suggest that while high automation reliability leads to an increase in the probability of correct outcomes, it also indirectly leads to a lower probability of a correct outcome during a failure by the automation (See Skitka et al., 1999 for an example). The reason for this paradoxical relationship between automation reliability and system performance is that interacting with automation of high reliability for extended periods of time leads to over reliance (Parasuraman et al., 1993). Once the human begins to over rely on the automation, the probability of a system error⁸ increases in the event that the automation makes an error. The effects of different error types have a similar consequence on reliance such that the prevalence of each error type can lead to over reliance on the automation depending on the state of the system. For example, in a system in which automation misses are prevalent, humans will likely adjust their reliance behavior to almost always rely on alarms. This behavior will have a negative impact on the probability of avoiding a system error during an automation false alarm. In a system where automation false alarms are prevalent, the human's reliance behavior will reduce the probability of detecting an automation miss. Therefore, while the adjustment of behavior as a function of the criterion of the automation suggests a trend toward appropriate reliance, such a shift in behavior does not occur without some shortcomings to overall system performance.

⁸ A system error is an instance in which the wrong output is generated by the system, irrespective of which of the system components (human or automation) caused it. For a list of technical terms and definitions see Appendix K

The Meaning of Alarms

Another important finding of this investigation was the reluctance of humans to rely on the automation during alarms relative to non-alarm periods. For example, in the first 30 minutes of the experiment, most Younger Adults (First Half and Throughout Conditions) were more reluctant to rely on the automation during alarms than they were to rely on it during non-alarms. Even in the Second Half Condition, where the automation was perfect for 30 minutes, the trend toward reliance on the automation was greater during non-alarm periods than during alarms. This behavior suggests that the different ways in which automation can provide support, via alarms and non-alarms, affect the way humans gather evidence about the reliability of the automation and the manner in which they adjust their behavior accordingly. Presumably, because of their salience, alarms prompt an increased tendency toward non-reliance relative to non-alarms.

Does it Matter When Errors Occur?

Yes. The Distribution of Errors along a specific range of time affects estimates of the reliability of the automation and subjective trust ratings. The results of this investigation suggest that a more recent exposure to errors has a greater impact on perceived reliability and trust than an exposure to errors during the initial interaction with an automated system. This finding is consistent with the results of previous studies, which have shown that trust does recover when the automation behaves reliably for an

extended period of time following an error or string of errors (Lee & Moray, 1992; 1994).

The findings of this investigation suggest that there is a negative relationship between perceived reliability and the recency of errors⁹ by the automation. This means that if humans are given sufficient time with error-free automation, their perception of it is likely to “recover.” The negative relationship between recency of errors and perceived reliability is illustrated in the conceptual model of human-automation interaction, which can be found in Appendix J.

The different Distribution of Errors also had an impact on overall reliance. Overall, participants in the First and Second Half conditions relied more on the automation than participants in the Throughout Condition. This finding suggests that when automation frequently and randomly generates errors, humans’ reliance is more likely to remain lower than if the automation behaves reliably for an extended period of time.

Interestingly, a higher concentration of errors in the first half of the experiment did not lead to an earlier development of different patterns of reliance as a function of Error Type as was expected. In both the First Half and Throughout Conditions, the different patterns of behavior emerged in the last 30 minutes of the experiment. However, in the Second Half Condition in which participants interacted with perfect automation for the first 30 minutes, the different patterns of behavior as a function of Error Type emerged shortly after the appearance of errors.

The earlier emergence of different reliance patterns during the last 30 minutes of the Second Half Condition highlights the effect that familiarity with a system has on

⁹ Recency of errors: time between errors and when the participant is asked to provide an estimate of the automation’s reliability or a rating of trust.

human interaction with automated systems. Wickens and Xu (2002) suggested that experience with a system is an important factor in the effects that different automation specific variables, specifically errors by the automation, have on human behavior. The results of this investigation suggest that when humans are allowed to build trust in an automated system, the development of different reliance patterns as a function of error type, especially during alarms, occurs faster than if humans are exposed to errors early in their interaction with the automation.

The effects of Distribution of Errors on reliance and trust observed in this investigation also shed light on the complex relationship between automation reliability and reliance. Specifically, these results suggest that when investigating the relationship between automation reliability and human reliance, it is critical to consider when automation errors occur. The impact of each error on objective behaviors, such as reliance, and subjective perceptions, such as trust, depends on the amount of experience and familiarity the human has with the automation. The implication of these findings to future research efforts in the field of human-automation interaction is that the term “overall reliability” is not sufficient to describe the effect that the error rate of the automation has on human behavior. Perhaps when referring to the error rate of the automation, a term such as “consistency of the automation” might be better suited than “reliability of the automation.” In addition to defining the consistency of the automation, it is also critical to describe and consider how errors are distributed throughout the experimental session.

Does Age Matter in Human-Automation Interaction?

Yes. Older Adults relied less on automated alarms than Younger Adults.

However, it appears that this age-related difference is not due to an aversion to using automation. The reluctance to rely on the automation during alarms, relative to non-alarms suggests that older adults have a more conservative approach when dealing with a system that has proven to be faulty at times. This conservative approach is only evident with reliance on alarms. However, the unwillingness of older adults to rely on alarms is considerably mitigated when they are given sufficient time and experience to build their trust in the automation. Therefore, the different patterns of reliance as a function of age are more likely to be a product of age-related differences in the development of appropriate trust in the automation. Because it takes older adults longer to discern the detection criterion of the automation, it also takes them longer to adjust their reliance behavior. This finding is consistent with the results of Sit and Fisk (1999), which showed that older adults take longer to modify their biases, especially with tasks that require “higher-order, strategic processing” (p. 26).

Interestingly, the reliability estimates and trust ratings of Older Adults were higher than those of Younger Adults. This finding is likely to be a product of the age-related benefits associated with having automated support in a multi-task environment with high attentional demands. When automated support has a high overall reliability, as it did in this investigation, older adults experience greater benefits from its presence. Without the automated support it is likely that age-related differences in performance would be greater. Previous research has found that as the reliability of the automation

degrades, older adults report lower trust ratings and reliability estimates than younger adults (Sanchez et al., 2004). However, other studies in which task demands (number of tasks and time pressure) were low, have not reported age-related differences in trust and perceived reliability (McCarley et al., 2003). This contrast in findings suggests that the workload demands of a system have an important effect on the perceived utility of the automation as a function of age.

In summary, it appears that an aversion toward technology by older adults is not the main factor driving age-related differences in human-automation interaction. The evidence from this investigation suggests that while it does take older adults longer (more trials) to adjust their behavior in accordance to the characteristics of the automation, they eventually do adjust. Once older adults adjust their behavior, the age-related differences in reliance on the automation are considerably mitigated. Also, the results of previous studies, as well as the current one, indicate that when the automation has a high level of reliability, age-related differences in subjective trust and perceived reliability are uncommon.

Does Domain Experience Matter in Human-Automation Interaction?

The most notable domain experience-related difference found in this investigation was the unwillingness of most Farmers to rely on the automated alarms. One explanation for this domain experience-related difference in reliance is that there may be differences in the preconceived meaning of an automation error by the Farmers group. Farmers might view any automation error as a permanent malfunction. This difference in the way

errors are perceived can lead to a more cautious approach in the reliance on alarms. This finding highlights the importance of prior expectations of the automation and reliance on it. It appears that there is a negative relationship between prior expectations of the automation and perceived reliability. As expectations of the automation are higher, any error will have a stronger negative impact on perceived reliability (See Appendix J).

The Construct of Trust

In one of the first publications that focused on the integration of the trust construct and the interaction of humans and automated systems, Muir (1987) argued that “the concept of trust is a critical one in the design of decision support systems” (p. 527). Since Muir’s publication, the role of trust in human-automation systems has received a considerable amount of attention, evident by the number of studies have been conducted to asses and understand the role of trust in human-automation interaction. The results of these studies have shed some light on the effects that various system variables such as reliability have on subjective attitudes toward automated systems (for reviews see Cohen, Parasuraman & Freeman, 1999; Lee & See, 2004; Muir, 1994; for a scale of trust see Jian, Bisantz & Drury, 2000).

In a recent and comprehensive review of the literature relevant to trust in automation, Lee and See (2004) defined trust as “the attitude that an agent will help achieve and individual’s goals in a situation characterized by uncertainty and vulnerability” (p. 54). There are two key components in Lee and See’s definition. The first is that trust is defined as an attitude, not an objective behavior. This is an important

distinction, as “trust” is often times used interchangeably to describe objective behaviors such as reliance. While there is evidence of a strong relationship between trust and the use of automation (Moray et al., 2000; Sanchez et al., 2004; Skitka et al., 1999; Vries et al., 2003), there is also evidence to suggest that it is not the only factor that predicts use (Dzindolet et al., 2003; Lee & Moray, 1992; Riley, 1996).

The second part of Lee and See’s definition mentions situations of uncertainty and vulnerability. Because automated systems are seldom 100% accurate, understanding the level of trust on a system serves as a valuable indicator of the likelihood of use. Therefore, from an applied standpoint, gauging operator trust is helpful in making predictions about behavior toward automation in situations of uncertainty. For example, in complex systems where the human does not have access to or an understanding of the information needed to make a decision, trust will likely predict behaviors such as reliance. From a research standpoint, trust serves as an indicator of the perceived reliability of an automated system. This allows researchers to parse out other variables that might influence objective behaviors. For example, if trust ratings are low while the use of the automation is high, it is likely that other variables such as workload might be influencing objective behavior. However, the low levels of trust indicate that the human is perceptive to factors such as low reliability, which affect the perception of the automation.

The results of the current investigation indicated that there was a strong relationship between subjective trust and reliance. Furthermore, there was also a strong relationship between reliance and perceived reliability, which is another way to measure perceptions of the human about the automation. Interestingly, the relationship between

trust and perceived reliability was not significant, and the patterns of trust and perceived reliability differed as a function of Distribution of Errors. This disconnection between trust and perceived reliability, suggests that the construct of subjective trust has a number of different components. To date, we do not have a good understanding of what these components might be and how much variance can be accounted for.

Practical Applications

Design of Automated Systems

The main issue with human-automation interaction across a variety of domains is the paradoxical problem that stems from the presence of automation in systems, which is that highly reliable automation, while desirable in terms of improving overall system performance, negatively affects human performance (evident by poor monitoring). As long as the human remains an integral component of the human-automation system, the formula for successful human-automation interaction is the congruency between the human's system representation and the design parameters of the automation. The findings of this investigation showed that humans can and do gather information about the automation that helps them adjust their behavior to best collaborate with it. In general, the acquired patterns of behavior as a function of the characteristics of the automation suggest that humans strive for optimal system performance.

Knowing the effects of specific variables on human-automation interaction can help designers of automated systems make some predictions about human behavior and system performance as a function of the characteristics of the automation. For example, based on the detection criterion of the automation, human operators are likely to develop distinct patterns of reliance. These distinct patterns can be used to make some predictions about the allocation of attention and workload during different states of the automation. Of course, the current findings regarding the effects of error type should not be used as the sole basis by which the detection criterion of an automated system is set. This decision must be made with knowledge of other important parameters specific to any system, such as the cost of errors.

Another finding from this investigation with important practical implications was the low detection rate of the exception error. This result suggests that once an operator develops a specific pattern of non-reliance, the probability of catching an uncommon automation error will be low. This can be particularly important if the criterion of the automation is set to protect against the type of error with the highest cost. For example, an automation miss by a system that detects ice on the wings of a plane can lead to the plane crashing at takeoff. Therefore, the criterion of this system is likely to be set so that it generates an alarm in the presence of the slightest indication of ice. This criterion setting is also likely to generate false alarms with a much higher frequency than it misses. Over time, the human responsible for monitoring the output from the ice warning system might become over reliant on the automation, especially in the absence of alarms. Therefore, the probability of detecting ice on the wing in the event of miss by the automation is likely to decrease.

Future Investigations of Human-Automation Interaction

The patterns of reliance observed in this investigation not only have an impact on our understanding of human-automation interaction, but also inform future investigations in this field about methodological issues that should be considered. The existence of different patterns of behavior as a function of error type suggest that, in addition to considering variables such as the overall reliability of the automation, it is critical to control for and anticipate the effects that variables such as error type and the distribution of errors can have on behavior. Furthermore, the analyses of behaviors such as reliance, must take into consideration that humans develop distinct patterns of reliance that are dependent on the state of the automation. Therefore, it is not sufficient to examine behavior at the macro-level if the objective of this field is to gain a thorough understanding of the way humans gather evidence about the automation and how they adjust their behavior accordingly.

To truly understand the effects of reliability on trust and reliance, it is critical to clearly and effectively calculate automation reliability. The reported reliability of the automated collision avoidance system used in this investigation was 95.4%. This value was calculated by dividing the number of events in which the automation provided accurate support (229) by the total number of events in which the automation provided support (240). However, this is not the way in which automation reliability is always calculated in this field of research. For example, some studies do not include periods of correct rejection when calculating the reliability of the automation (e.g., Crocoll & Curry, 1990; Dixon & Wickens, 2003; 2004; Sanchez et al., 2004). One reason for the exclusion

of correct rejection periods in calculating the reliability of automated systems, is that the lack of understanding of the effect of periods of correct rejection on trust in and reliance on the automation. For example, assuming that a correct alarm leads to an increase in trust in the automation, it is unclear what the equivalent length of correct rejection time is to have the same effect on trust. Therefore, future empirical investigations that shed light into the way humans weigh different events by the automation are critical to further our understanding of the automation reliability – reliance relationship.

Where We Have Been, Where We Are Now, and Where We Should Go

Rather than providing an extensive list of follow-up experiments to the current investigation, it may prove more productive to briefly review the historical research trends from the field of human-automation interaction and examine the current state of the field. This approach might make the contributions that have been made to both the theoretical and applied areas more clear, and shed light on the future paths of research that are worth pursuing. The objective of this field is to understand human behavior in automated environments. From a scientific standpoint, understanding human behavior in automated environments can inform psychological theories of attention (Molloy & Parasuraman, 1996; Parasuraman, Mouloua & Molloy, 1996; Parasuraman et al., 1993), and workload (Bliss & Dunn, 2000; Wickens & Dixon, 2005) among others. From an applied standpoint, understanding human capabilities and limitations impacts the design of automated systems across a variety of domains.

Where We Have Been

Some of the earlier work related to human-automation interaction focused on the optimal allocation of functions between humans and machines (e.g., Fitts, 1951) and the distribution of responsibilities within a single task between human and machine agents (e.g., Price, 1985; Sheridan & Verplank, 1978). This line of research generated much discussion about the criteria that should be used for function allocation. Naturally, these discussions stimulated research to understand human capabilities and limitations with the goal of making more informed decisions about function allocation.

Parasuraman (1987) and Muir (1987) were some of the first to consider the effects of automation on human behavior. They discussed issues such as over and under reliance on automation and the consequences associated with such acquired behaviors. During this time there was also a growing concern about the increase of accidents as a possible result of automation-induced behaviors (National Research Council, 1997).

Since the Parasuraman (1987) and Muir (1987) articles, a considerable amount of research has been conducted in an effort to understand human behavior in automated environments. In one of the first studies that examined human behavior, Lee and Moray (1992) investigated the effects of failures by an automated system on trust and reliance. Research efforts since Lee and Moray's paper have followed in their footsteps, trying to understand how variables such as automation reliability, workload, and the type of automation among others, affect trust in automation and reliance on it.

Unfortunately, the last 20 years of the research in this field lacks a truly systematic approach. With few exceptions, a majority of the studies have been isolated

efforts that failed to extract the key findings from previous investigations to formulate and test new hypotheses. A clear example of the lack of a systematic approach is the number of studies that have attempted to understand the relationship between automation reliability and reliance while ignoring critical components of reliability such as error type. As it turns out, the results of the current investigation show that the effects of reliability on reliance are moderated by the type of error that is prevalent in the system. The findings of the current investigation, suggest that it may be worth revisiting some of the conclusions that have been made in studies that investigated the effects of automation reliability on reliance.

Where We Are Now

Stepping back on the optimistic side, even with the lack of a systematic approach, the findings from this field have had an impact on a number of psychological theories (e.g., workload, sustained attention) as well as the design of systems across a variety of domains (e.g., surface transportation, aviation, medical systems). During the last decade, a handful of literature reviews have been instrumental in redirecting the efforts of this field (e.g., Meyer, 2004; Parasuraman & Riley, 1997; Pritchett, 2001; Sheridan, 2002). Most recently, Lee and See (2004) discussed the importance of working toward an understanding of *appropriate trust* and *reliance* (Lee & See, 2004), which means using or relying on the automation only in situations that it is designed to handle. The underlying importance of the concept of appropriate reliance is that it redefines the automation reliability – reliance relationship as a conditional, dynamic process, rather than an

absolute (rely or not rely), static one. A number of research efforts are beginning to investigate human-automation interaction while attempting to discern the variables critical to appropriate reliance.

Another important trend in the current research efforts of the field is the attempt to understand how different Levels of Automation (LOA) affect human behavior. Since Parasuraman et al. (2000) proposed a framework of automation, there has been an increase in awareness that this is a variable that should not be overlooked. The results of recent studies suggest that reliance and trust change as a function of LOA (Clamman, Wright & Kaber, 2002; Endsley & Kaber, 1999; Moray et al., 2000). However, the interaction effects between LOA and variables such as reliability are still not well understood.

So where are we now? We do have a general understanding of the effects that a number of variables have on the behavior of humans in automated environments. This knowledge has been captured and summarized by some of the reviews of the literature, and by a few comprehensive conceptual models of human-automation interaction that have conceptualized some of the cause-effect relationships we are beginning to understand (e.g., Riley, 1996; Dzindolet et al., 2001; Sanchez, 2005). However, the contradictory results of investigations that claim to target similar issues still go unexplained. Furthermore, overgeneralizations based on the results and discussions of a few studies are made too often, which results in a failure to address and understand the key issues and variables.

Where We Should Go

Research in this field will continue to be motivated in part by the emergence of new technologies that allow machines to perform higher level functions such as information integration to support decision making. The constant development of new technologies will continue to drive new research ideas that are aimed at understanding the relationship between the characteristics of specific technologies and human behavior. However, research efforts that are simply reactionary initiatives to emerging technologies will “only brighten an inconsequential corner rather than contributing to the science” (Adams, 1987; p. 41). Therefore, the following are some suggestions about the paths that should be traveled by future research efforts and the factors that those efforts should consider:

- We must begin to ask WHY the manipulations of specific variables lead to specific changes in behavior. For example, simply noting that changes in the reliability of an automated system lead to changes in reliance without asking why, merely provides a surface level understanding of the factors that affect use in automation. A deeper interpretation of the results of our experiments will force us to pay more attention to small but important details that appear to moderate and mediate some of the behaviors being observed. For example, one key variable that is often overlooked by researchers when drawing conclusions from previous research efforts and generating new experiments is the amount of information about the reliability of the automation that participants receive (See Sanchez, 2005 for a review). This variable alone

appears to be an important moderator and/or mediator of the effects that factors such as reliability have on trust and reliance.

- Efforts to generate conceptual and statistical models that describe and extend what we know about human-automation interaction need to continue. The development of conceptual and statistical models will allow us to more easily detect the relationships that we do not yet understand and highlight what we do understand.
- We must realize that the term automation covers a wide spectrum of systems, which range from very “simple” sensing technologies to complex decision support aids. The existing taxonomies of automation (Endsley& Kaber, 1999; Parasuraman et al., 2000; Sheridan & Verplank, 1978) provide a solid starting point, but are not sufficient to classify the numerous variables specific to different types of automation that can affect human behavior. Developing a more comprehensive taxonomy of automation will facilitate the interpretation and organization of results from studies that use different types of automation.

APPENDIX A

RELIANCE DATA FROM JOHNSON (2004)

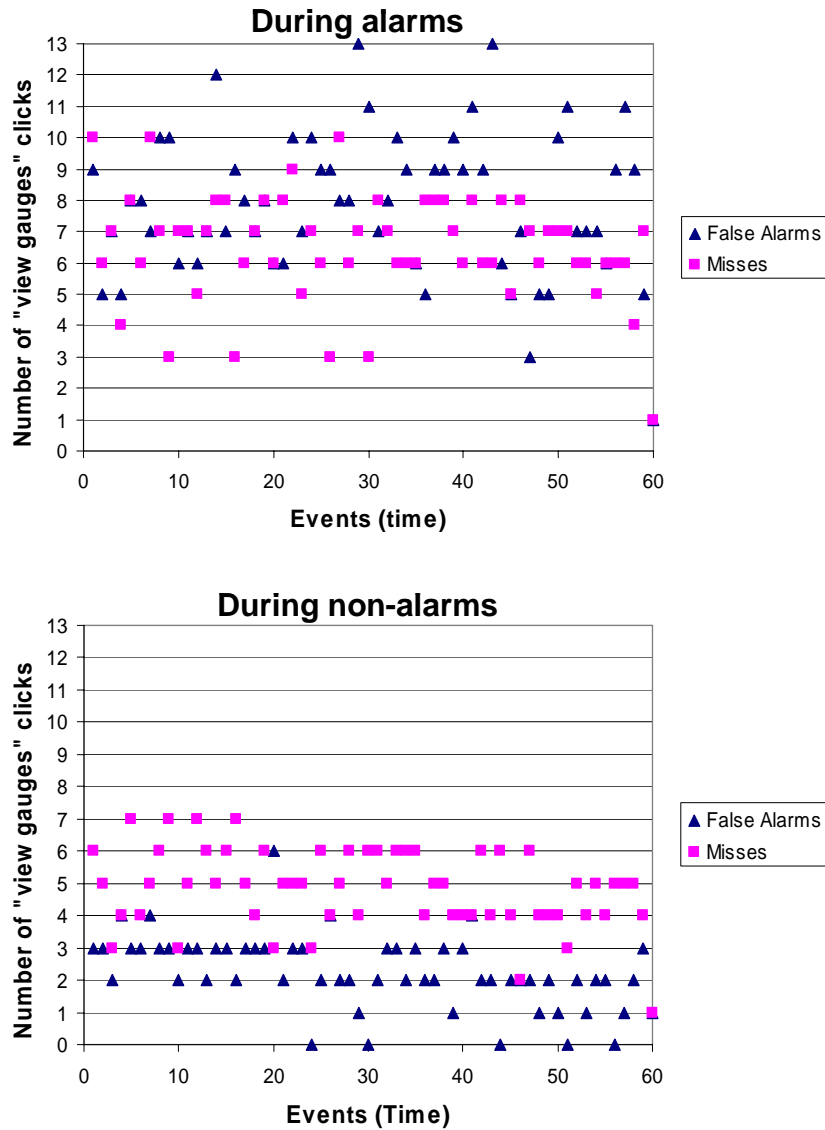


Figure A1. Reliance data from Johnson (2004). The top graph shows the non-reliance behavior during alerts by the decision support aid and the bottom graph shows non-reliance behavior during times when the decision support aid was not alerting. The squares represent the majority false alarm condition and the triangles the majority misses condition.

APPENDIX B

PARTICIPANT CHARACTERISTICS BY ERROR TYPE (EXPERIMENT 1)

Table B1. Participant characteristics by Error Type Condition (Experiment 1)

	Age		Paper Folding Task* ¹		Shipley Vocabulary Test* ²		Digit-symbol Substitution* ³		Reverse Digit Span* ⁴	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
False Alarms	23.5	2.6	9.3	2.1	29.1	3.1	62.8	9.8	5.8	1.5
Misses	22.8	3.3	9.5	2.0	28.5	3.0	62.5	9.6	6.7	1.3

*¹ Number of correct items (maximum 20); *² Number of correct items (maximum 40); *³ Number of completed items (maximum 100); *⁴ Number of digits recalled in the correct order (maximum 14)

APPENDIX C

DISTRIBUTION OF ERRORS

Table C1. Distribution of errors

Time bin	Throughout	First half	Second half
1	correct rejection	correct rejection	correct rejection
2	correct rejection	correct rejection	correct rejection
3	correct rejection	correct rejection	correct rejection
4	correct alarm	correct alarm	correct alarm
5	correct rejection	correct rejection	correct rejection
6	correct rejection	correct rejection	correct rejection
7	correct rejection	correct rejection	correct rejection
8	error	error	correct alarm
9	correct rejection	correct rejection	correct rejection
10	correct rejection	correct rejection	correct rejection
11	correct alarm	correct alarm	correct alarm
12	correct rejection	correct rejection	correct rejection
13	correct rejection	correct rejection	correct rejection
14	correct alarm	correct alarm	correct alarm
15	correct rejection	correct rejection	correct rejection
16	correct alarm	error	correct alarm
17	correct alarm	correct alarm	correct alarm
18	correct rejection	correct rejection	correct rejection
19	error	error	correct alarm
20	correct alarm	correct alarm	correct alarm
21	correct rejection	correct rejection	correct rejection
22	correct alarm	correct alarm	correct alarm
23	correct rejection	correct rejection	correct rejection
24	correct alarm	correct alarm	correct alarm
25	correct rejection	correct rejection	correct rejection
26	correct alarm	correct alarm	correct alarm
27	correct rejection	correct rejection	correct rejection
28	correct rejection	correct rejection	correct rejection
29	correct rejection	correct rejection	correct rejection
30	correct alarm	error	correct alarm
31	correct rejection	correct rejection	correct rejection
32	correct rejection	correct rejection	correct rejection
33	correct rejection	correct rejection	correct rejection
34	correct rejection	correct rejection	correct rejection
35	correct alarm	correct alarm	correct alarm
36	correct rejection	correct rejection	correct rejection
37	correct rejection	correct rejection	correct rejection
38	correct rejection	correct rejection	correct rejection
39	correct rejection	correct rejection	correct rejection

88	correct rejection	correct rejection	correct rejection
89	correct alarm	error	correct alarm
90	correct rejection	correct rejection	correct rejection
91	correct rejection	correct rejection	correct rejection
92	correct rejection	correct rejection	correct rejection
93	correct rejection	correct rejection	correct rejection
94	correct alarm	correct alarm	correct alarm
95	correct rejection	correct rejection	correct rejection
96	correct rejection	correct rejection	correct rejection
97	correct rejection	correct rejection	correct rejection
98	correct alarm	correct alarm	correct alarm
99	correct rejection	correct rejection	correct rejection
100	correct alarm	correct alarm	correct alarm
101	correct rejection	correct rejection	correct rejection
102	correct alarm	correct alarm	correct alarm
103	correct rejection	correct rejection	correct rejection
104	correct rejection	correct rejection	correct rejection
105	error	error	correct alarm
106	correct rejection	correct rejection	correct rejection
107	correct rejection	correct rejection	correct rejection
108	correct rejection	correct rejection	correct rejection
109	correct rejection	correct rejection	correct rejection
110	correct rejection	correct rejection	correct rejection
111	correct alarm	error	correct alarm
112	correct rejection	correct rejection	correct rejection
113	correct rejection	correct rejection	correct rejection
114	correct rejection	correct rejection	correct rejection
115	correct rejection	correct rejection	correct rejection
116	correct alarm	exception error	correct alarm
117	correct rejection	correct rejection	correct rejection
118	correct rejection	correct rejection	correct rejection
119	correct rejection	correct rejection	correct rejection
120	correct rejection	correct rejection	correct rejection
121	correct rejection	correct rejection	correct rejection
122	correct rejection	correct rejection	correct rejection
123	correct rejection	correct rejection	correct rejection
124	correct rejection	correct rejection	correct rejection
125	correct rejection	correct rejection	correct rejection
126	correct alarm	correct alarm	correct alarm
127	correct rejection	correct rejection	correct rejection
128	error	correct alarm	error
129	correct rejection	correct rejection	correct rejection
130	correct rejection	correct rejection	correct rejection
131	correct alarm	correct alarm	correct alarm
132	correct rejection	correct rejection	correct rejection
133	correct rejection	correct rejection	correct rejection
134	correct alarm	correct alarm	correct alarm
135	correct rejection	correct rejection	correct rejection

232	correct rejection	correct rejection	correct rejection
233	correct rejection	correct rejection	correct rejection
234	correct rejection	correct rejection	correct rejection
235	correct rejection	correct rejection	correct rejection
236	correct rejection	correct rejection	correct rejection
237	correct rejection	correct rejection	correct rejection
238	correct rejection	correct rejection	correct rejection
239	correct rejection	correct rejection	correct rejection
240	correct rejection	correct rejection	correct rejection

APPENDIX D

SUBJECTIVE TRUST AND PERCEIVED RELIABILITY QUESTIONNAIRE

Participant Number _____

1. Overall how much do you trust the Collision Avoidance System?

1	2	3	4	5	6	7
Not at all						Completely

2. To what extent can you count on the Collision Avoidance System to do its job?

1	2	3	4	5	6	7
Not at all						Completely

3. Please indicate, using a number, your estimate of the reliability of the Collision Avoidance System

(Example: I think the Collision Avoidance System was XX% reliable)

_____ %

APPENDIX E

PARTICIPANT CHARACTERISTICS BY AGE, ERROR TYPE AND DISTRIBUTION OF ERRORS (EXPERIMENT 2)

Table E1. Participant characteristics by Age and Error Type for the Throughout Condition

	Age		Paper Folding Task* ¹		Shipley Vocabulary Test* ²		Digit-symbol Substitution* ³		Reverse Digit Span* ⁴	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Younger Adults										
False Alarms	19.6	2.2	11.3	2.2	31.7	4.5	74.9	7.4	7.0	1.8
Misses	19.1	1.3	11.1	1.7	30.9	4.2	72.0	6.7	7.1	1.6
Older Adults										
False Alarms	69.5	3.2	7.6	1.7	33.8	3.1	59.0	9.3	5.2	0.9
Misses	69.3	4.5	8.1	1.9	33.8	6.4	52.8	14.9	4.3	1.7

*¹ Number of correct items (maximum 20); *² Number of correct items (maximum 40); *³ Number of completed items (maximum 100); *⁴ Number of digits recalled in the correct order (maximum 14)

Table E2. Participant characteristics by Age and Error Type for the First Half Condition

	Age		Paper Folding Task* ¹		Shipley Vocabulary Test* ²		Digit-symbol Substitution* ³		Reverse Digit Span* ⁴	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Younger Adults										
False Alarms	19.5	1.4	10.1	1.9	29.8	2.8	76.4	8.7	7.3	1.6
Misses	19.4	1.4	11.4	1.4	28.6	3.6	75.8	9.2	7.1	2.1
Older Adults										
False Alarms	70.3	4.0	8.1	2.1	34.7	4.3	53.8	8.0	5.3	0.8
Misses	71.3	2.8	7.3	1.9	34.5	2.4	60.0	11.5	5.0	1.2

*¹ Number of correct items (maximum 20); *² Number of correct items (maximum 40); *³ Number of completed items (maximum 100); *⁴ Number of digits recalled in the correct order (maximum 14)

Table E3. Participant characteristics by Age and Error Type for the Second Half Condition

	Age		Paper Folding Task* ¹		Shipley Vocabulary Test* ²		Digit-symbol Substitution* ³		Reverse Digit Span* ⁴	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Younger Adults										
False Alarms	19.7	1.9	10.2	1.9	29.5	3.4	74.8	7.7	6.7	1.8
Misses	19.8	1.3	10.2	2.1	29.8	5.5	70.6	8.8	6.3	2.0
Older Adult										
False Alarms	67.7	2.5	9.0	1.9	34.0	3.1	49.4	12.7	5.2	1.2
Misses	69.6	3.3	8.8	1.8	34.3	5.5	48.1	14.5	4.9	1.1
* ¹ Number of correct items (maximum 20); * ² Number of correct items (maximum 40); * ³ Number of completed items (maximum 100); * ⁴ Number of digits recalled in the correct order (maximum 14)										

APPENDIX F

ANOVA TABLE (EXPERIMENT 2)

Table F1. ANOVA table (Experiment 2)

Source	Dependent Variable	SS	df	MS	<i>F</i>	<i>p</i>
Age	Spacebars	572.0	1	572.0	.04	.839
	First spacebars	658.0	1	658.0	.25	.615
	Total spacebar time*	20399.4	1	20399.4	11.00	.001
	Spacebars during alarms*	715.4	1	715.4	10.55	.002
	Spacebars during non-alarms	2650.8	1	2650.8	1.21	.274
	Trust*	20.0	1	20.0	17.96	.000
	Perceived reliability*	585.2	1	585.2	4.59	.034
	Collision errors*	392.4	1	392.4	28.16	.000
	False negative rate*	.0	1	.0	16.89	.000
	False positive rate*	.0	1	.0	27.45	.000
	Red events*	77114.7	1	77114.7	78.15	.000
	Tracking errors*	403.3	1	403.3	63.95	.000
	Digit substitution*	12281.6	1	12281.6	115.3	.000
	Shipley vocabulary*	512.5	1	512.5	29.12	.000
	Reverse digit*	112.1	1	112.1	49.96	.000
	Paper folding*	197.6	1	197.6	55.96	.000
	Participants' age*	75300.3	1	75300.3	1x10 ⁴	.000
Error Type	Spacebars	19712.0	1	19712.0	1.43	.234
	First spacebars*	25085.2	1	25085.2	9.70	.002
	Total spacebar time*	27808.6	1	27808.6	14.99	.000
	Spacebars during alarms*	3553.4	1	3553.4	52.40	.000
	Spacebars during non-alarms*	49207.5	1	49207.5	22.48	.000
	Trust*	16.9	1	16.9	15.15	.000
	Perceived reliability	161.0	1	161.0	1.26	.263
	Collision errors*	304.0	1	304.0	21.81	.000
	False negative rate*	.1	1	.1	52.35	.000
	False positive rate*	.0	1	.0	14.72	.000
	Red events*	10490.7	1	10490.7	10.63	.001
	Tracking errors*	48.1	1	48.1	7.63	.007
	Digit substitution	67.5	1	67.5	.63	.428
	Shipley vocabulary	2.1	1	2.1	.12	.728
	Reverse digit	3.3	1	3.3	1.49	.226
	Paper folding	.3	1	.3	.08	.771
	Participants' age	4.0	1	4.0	.55	.459
Distribution of Errors	Spacebars	71368.8	2	35684.4	2.59	.079
	First spacebars*	19349.8	2	9674.9	3.74	.027
	Total spacebar time*	16468.2	2	8234.1	4.44	.014

Age x Error Type	Spacebars during alarms*	1154.4	2	577.2	8.51	.000
	Spacebars during non-alarms	12371.5	2	6185.7	2.83	.064
	Trust*	50.5	2	25.2	22.65	.000
	Perceived reliability*	2474.1	2	1237.1	9.71	.000
	Collision errors	23.3	2	11.6	.83	.437
	False negative rate	.0	2	.0	.90	.409
	False positive rate	.0	2	.0	2.74	.069
	Red events	3665.9	2	1832.9	1.86	.161
	Tracking errors	7.3	2	3.7	.58	.562
	Digit substitution*	697.1	2	348.6	3.27	.042
	Shipley vocabulary	11.3	2	5.6	.32	.727
	Reverse digit	3.4	2	1.7	.75	.477
	Paper folding	2.6	2	1.3	.37	.691
	Participants' age	19.3	2	9.7	1.33	.270
	Spacebars	4966.5	1	4966.5	.36	.549
	First spacebars	2367.4	1	2367.4	.92	.341
	Total spacebar time	1055.6	1	1055.6	.57	.452
	Spacebars during alarms	249.4	1	249.4	3.68	.058
	Spacebars during non-alarms	4662.5	1	4662.5	2.13	.147
	Trust	1.0	1	1.0	.91	.344
	Perceived reliability	99.0	1	99.0	.78	.380
	Collision errors	20.0	1	20.0	1.44	.233
	False negative rate*	.0	1	.0	13.08	.000
	False positive rate*	.0	1	.0	13.41	.000
	Red events*	4813.3	1	4813.3	4.88	.029
	Tracking errors	12.0	1	12.0	1.91	.170
	Digit substitution	34.1	1	34.1	.32	.572
	Shipley vocabulary	2.7	1	2.7	.15	.696
	Reverse digit	.8	1	.8	.37	.544
	Paper folding	2.1	1	2.1	.60	.439
	Participants' age	8.5	1	8.5	1.17	.282
Age x Distribution of Errors	Spacebars	11054.1	2	5527.1	.40	.670
	First spacebars	3585.6	2	1792.8	.69	.502
	Total spacebar time	3797.2	2	1898.6	1.02	.363
	Spacebars during alarms*	704.3	2	352.1	5.19	.007
	Spacebars during non-alarms	3509.5	2	1754.7	.80	.451
	Trust*	7.6	2	3.8	3.42	.036
	Perceived reliability*	1064.1	2	532.1	4.18	.018
	Collision errors	46.7	2	23.3	1.67	.192
	False negative rate	.0	2	.0	.48	.618
	False positive rate*	.0	2	.0	3.37	.038
	Red events	463.9	2	231.9	.24	.791
	Tracking errors	6.2	2	3.1	.49	.612
	Digit substitution	220.8	2	110.4	1.04	.358
	Shipley vocabulary	44.1	2	22.0	1.25	.290
	Reverse digit*	3.8	2	1.9	.85	.430
	Paper folding	24.5	2	12.3	3.47	.035

Error Type x Distribution of Errors	Participants' age	30.1	2	15.0	2.06	.132
	Spacebars*	97367.6	2	48683.8	3.54	.032
	First spacebars*	56769.0	2	28384.5	10.98	.000
	Total spacebar time*	42528.8	2	21264.4	11.46	.000
	Spacebars during alarms*	1556.9	2	778.4	11.48	.000
	Spacebars during non-alarms*	39647.0	2	19823.5	9.06	.000
	Trust	4.9	2	2.4	2.18	.118
	Perceived reliability	341.3	2	170.7	1.34	.266
	Collision errors	70.1	2	35.0	2.51	.086
	False negative rate	.0	2	.0	1.43	.244
	False positive rate	.0	2	.0	2.96	.056
	Red events	5233.6	2	2616.8	2.65	.075
	Tracking errors	22.0	2	11.0	1.75	.179
	Digit substitution	293.6	2	146.8	1.38	.256
	Shipley vocabulary	5.3	2	2.6	.15	.861
	Reverse digit	.1	2	.1	.03	.974
	Paper folding	.7	2	.3	.09	.912
Age x Error Type x Distribution of Errors	Participants' age	9.2	2	4.6	.63	.533
	Spacebars	11626.5	2	5813.3	.42	.656
	First spacebars	8811.8	2	4405.9	1.70	.187
	Total spacebar time	10094.5	2	5047.2	2.72	.070
	Spacebars during alarms*	1150.9	2	575.4	8.49	.000
	Spacebars during non-alarms	4882.6	2	2441.3	1.12	.332
	Trust	2.2	2	1.1	1.00	.373
	Perceived reliability	198.3	2	99.2	.78	.462
	Collision errors	13.1	2	6.5	.47	.627
	False negative rate	.0	2	.0	.32	.728
	False positive rate*	.0	2	.0	4.47	.014
	Red events	810.8	2	405.4	.41	.664
	Tracking errors	.1	2	.1	.01	.991
	Digit substitution	129.7	2	64.9	.61	.546
	Shipley vocabulary	1.4	2	.7	.04	.961
	Reverse digit	1.7	2	.9	.38	.683
	Paper folding	10.2	2	5.1	1.45	.240
	Participants' age	2.8	2	1.4	.19	.825

Note. * indicates significance at $p < .05$

APPENDIX G

CORRELATIONS TABLE (EXPERIMENT 2)

Table G1. Correlations table (Experiment 2)

	First spacebars	Spacebars (alarms)	Spacebars (non- alarms)	Trust	Perceived reliability	Number of errors	Tracking (red events)
Total spacebars	.94**	.41**	.90**	-.27**	-.17	-.43	.17
First spacebars		.42**	.96**	-.24**	-.13	-.35	.15
Spacebars (alarms)			.18	.20*	.06	-.29**	.19*
Spacebars (non- alarms)				-.34**	-.17	-.29**	.14
Trust ratings					.54**	.16	.08
Perceived reliability						.02	-.02
Number of errors							.33**
Note. * indicates significance at $p < .05$; ** indicates significance at $p < .01$							

APPENDIX H

ANOVA TABLE (EXPERIMENT 1)

Table H1. ANOVA table (Experiment 1)

Source	Dependent Variable	SS	df	MS	<i>F</i>	<i>p</i>
Agricultural Experience	Spacebars	90.0	1	90.0	.01	.924
	First spacebars	144.4	1	144.4	.06	.811
	Total spacebar time*	29268.1	1	29268.1	7.14	.011
	Spacebars during alarms*	555.0	1	555.0	11.89	.001
	Spacebars during non-alarms	1357.2	1	1357.2	.63	.433
	Trust	.4	1	.4	.26	.616
	Perceived reliability	2.5	1	2.5	.02	.896
	Collision errors	3.6	1	3.6	.84	.365
	False negative rate	.0	1	.0	1.58	.216
	False positive rate	.0	1	.0	3.06	.089
	Red events	115.6	1	115.6	.37	.548
	Digit substitution*	1166.4	1	1166.4	16.18	.000
	Shipley vocabulary*	62.5	1	62.5	4.50	.041
	Reverse digit	6.4	1	6.4	2.66	.112
	Paper folding*	32.4	1	32.4	8.12	.007
	Participant's age*	144.4	1	144.4	24.13	.000
Error Type	Domain experience*	490.0	1	490.0	132.6	.000
	Spacebars	14745.6	1	14745.6	1.53	.224
	First spacebars*	18147.6	1	18147.6	7.29	.010
	Total spacebar time	15792.7	1	15792.7	3.85	.057
	Spacebars during alarms*	2739.0	1	2739.0	58.67	.000
	Spacebars during non-alarms*	35343.0	1	35343.0	16.39	.000
	Trust	.9	1	.9	.58	.453
	Perceived reliability	108.9	1	108.9	.76	.390
	Collision errors*	32.4	1	32.4	7.57	.009
	False negative rate*	.0	1	.0	9.49	.004
	False positive rate	9E-005	1	9E-005	2.57	.118
	Red events*	2592.1	1	2592.1	8.26	.007
	Digit substitution	25.6	1	25.6	.36	.555
	Shipley vocabulary	4.9	1	4.9	.35	.556
	Reverse digit	2.5	1	2.5	1.04	.315
	Paper folding	.0	1	.0	.00	1.000
Agricultural Experience *	Participant's age	3.6	1	3.6	.60	.443
	Experience*	22.5	1	22.5	6.09	.018
	Spacebars	230.4	1	230.4	.02	.878
	First spacebars	25.6	1	25.6	.01	.920

Error Type	Total spacebar time	423.8	1	423.8	.10	.750
	Spacebars during alarms	75.6	1	75.6	1.62	.211
	Spacebars during non-alarms	216.2	1	216.2	.10	.753
	Trust	1.6	1	1.6	1.02	.318
	Perceived reliability	2.5	1	2.5	.02	.896
	Collision errors*	32.4	1	32.4	7.57	.009
	False negative rate	.0	1	.0	2.77	.105
	False positive rate	.0	1	.0	1.30	.262
	Red events	129.6	1	129.6	.41	.524
	Digit substitution	16.9	1	16.9	.23	.631
	Shipleigh vocabulary	.1	1	.1	.01	.933
	Reverse digit	1.6	1	1.6	.67	.420
	Paper folding	.4	1	.4	.10	.753
	Participant's age	.1	1	.1	.02	.898
	Experience*	22.5	1	22.5	6.09	.018
Note. * indicates significance at $p < .05$						

APPENDIX I

“FIRST ERROR EFFECT”

The y-axis represents the difference in the number of participants who pressed the spacebar before and after each error. A positive number on the y-axis indicates a change toward non-reliance and a negative number toward reliance. The x-axis represents each error in the experiment. The exception error was excluded from this analysis.

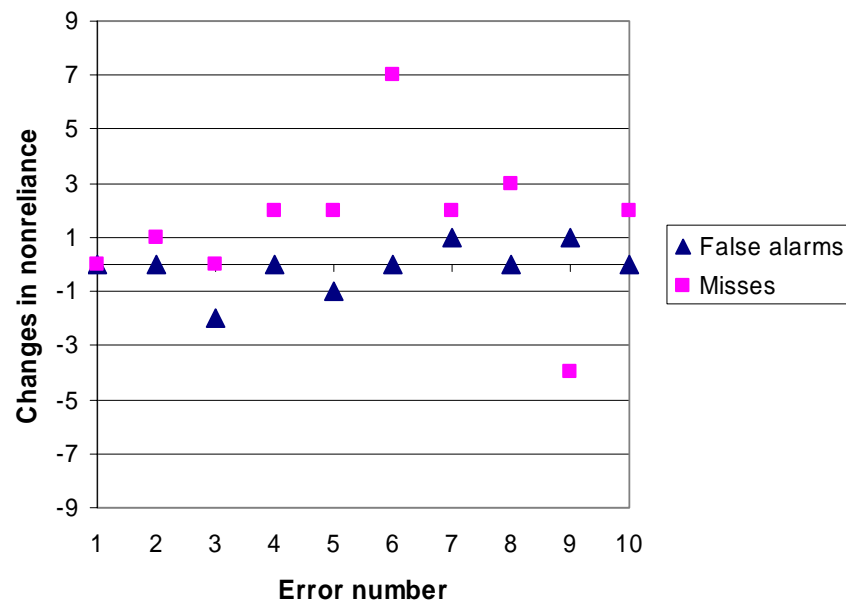


Figure 11. Changes in the number of participants who pressed the spacebar before and after each error during alarms in the Throughout Condition.

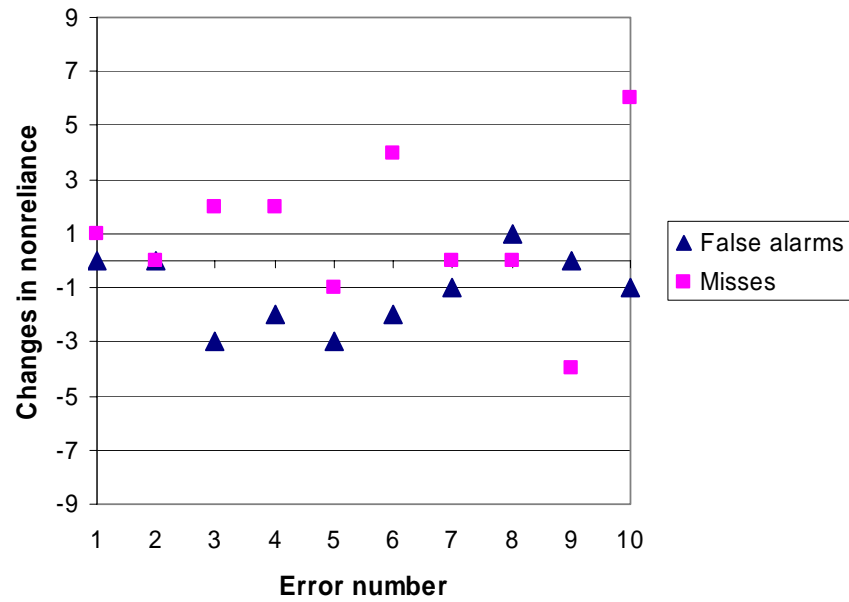


Figure I2. Changes in the number of participants who pressed the spacebar before and after each error during non-alarms in the Throughout Condition.

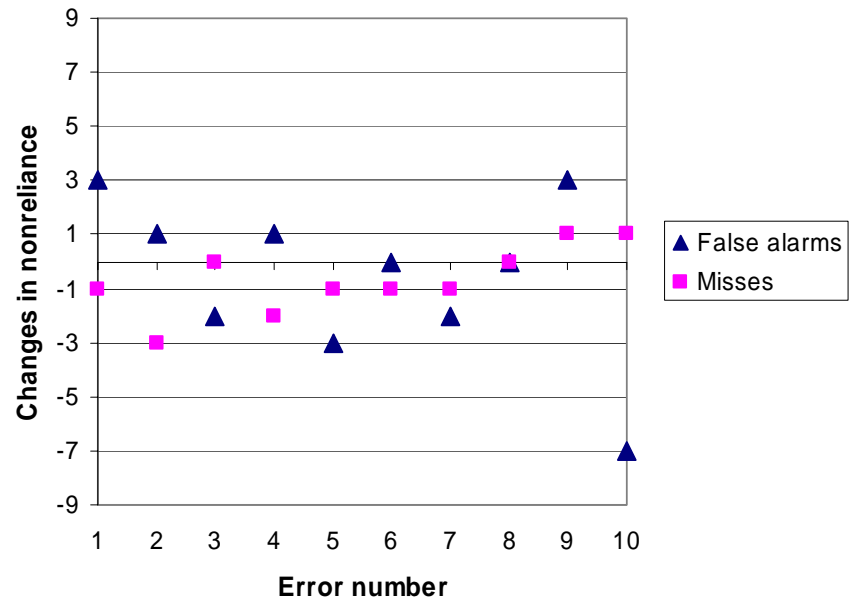


Figure I3. Changes in the number of participants who pressed the spacebar before and after each error during alarms in the First Half Condition.

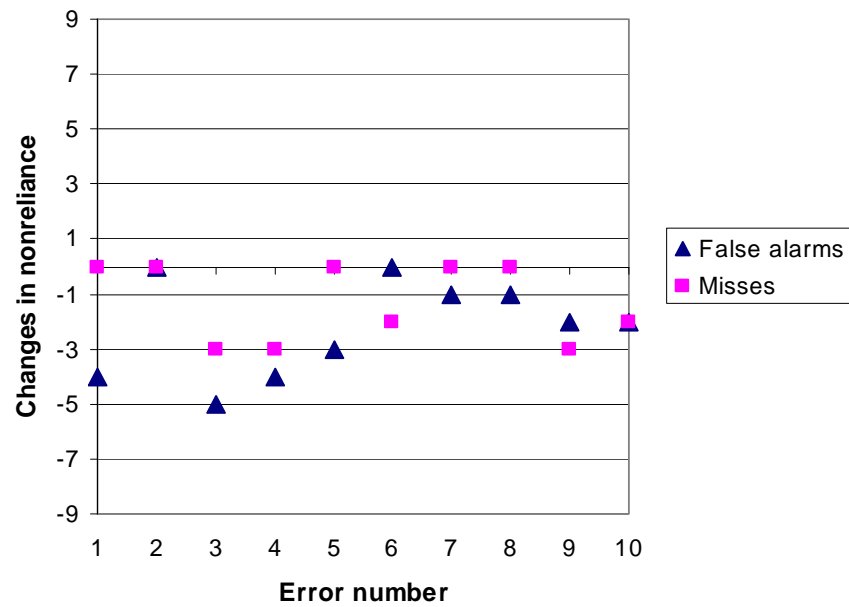


Figure I4. Changes in the number of participants who pressed the spacebar before and after each error during non-alarms in the First Half Condition.

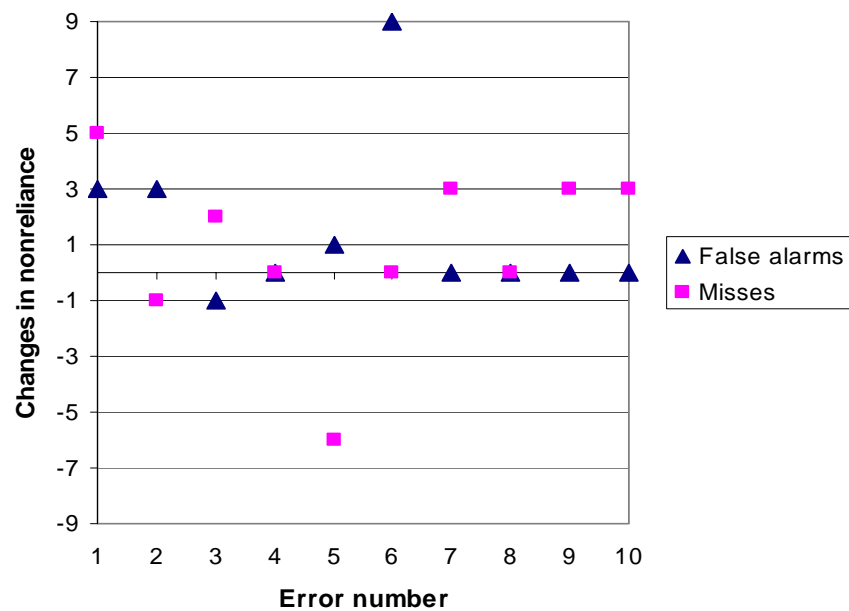


Figure I5. Changes in the number of participants who pressed the spacebar before and after each error during alarms in the Second Half Condition.

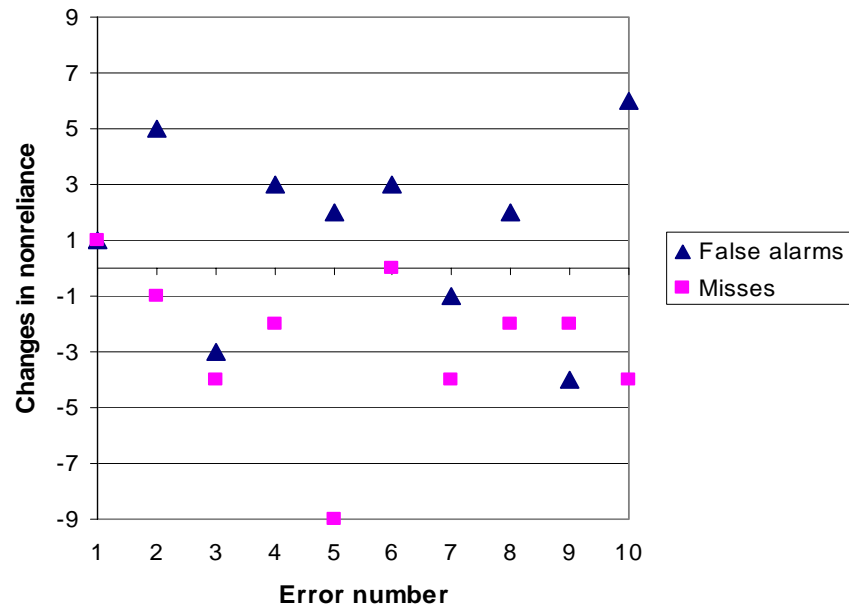


Figure I6. Changes in the number of participants who pressed the spacebar before and after each error during non-alarms in the Second Half Condition.

Table II. Changes in the number of Younger Adults in Experiment 2 who pressed the spacebar before and after the first automation error.

	During Alarms		During Non-alarms	
	False Alarms Condition	Misses Condition	False Alarms Condition	Misses Condition
Throughout	0	0	0	+1
First half	+3	-1	0	-4
Second half	+3	+5	+1	+1

APPENDIX J

CONCEPTUAL MODEL OF HUMAN-AUTOMATION INTERACTION

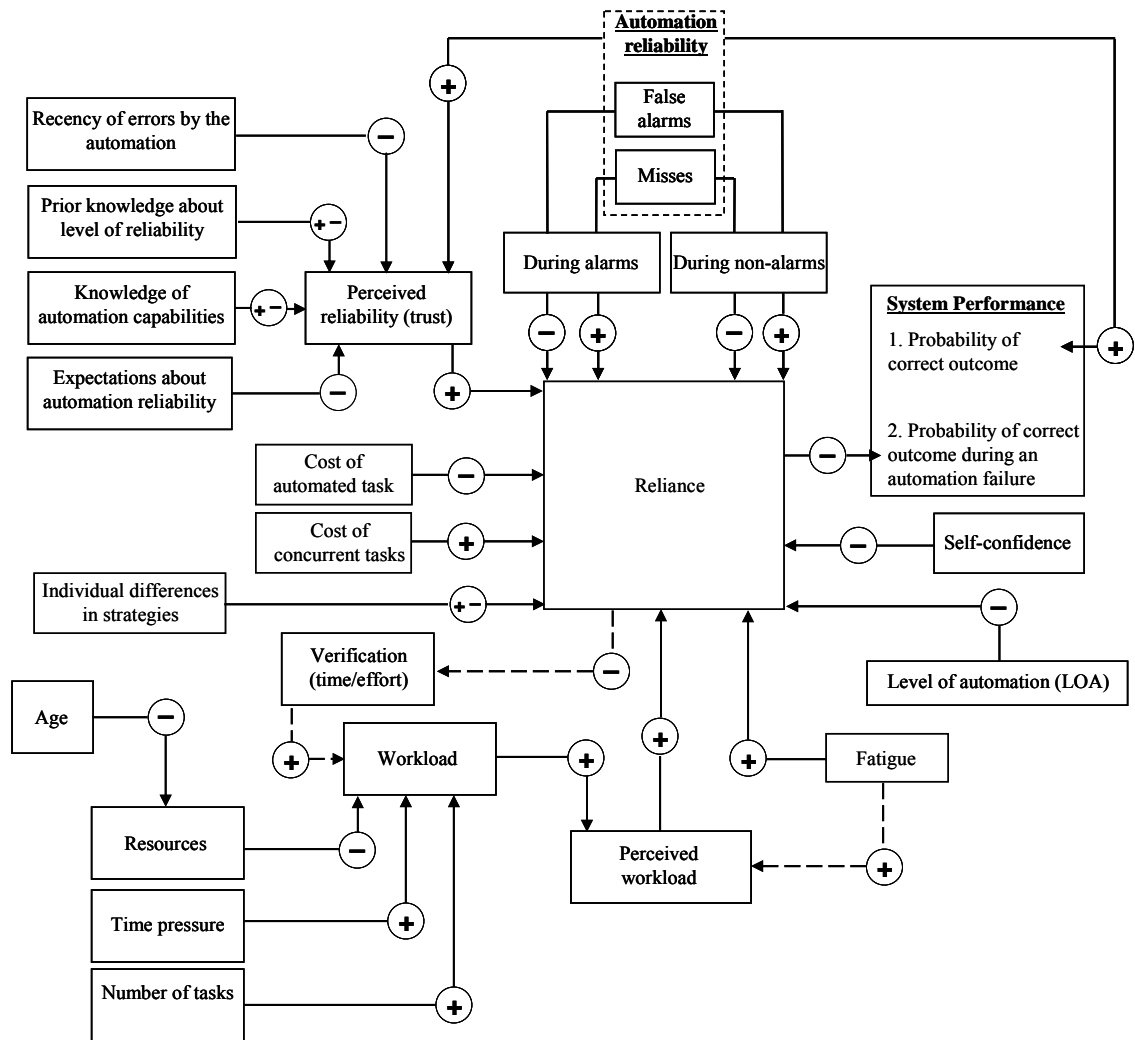


Figure J1. Conceptual model of human-automation interaction

APPENDIX K

TECHNICAL TERMS & DEFINITIONS

- *Actual reliability*: Describes the true or actual reliability of the automation. Reliability is calculated by dividing the number of events in which the automation provides correct support by the total number of events in which the automation provides support. For example, a system that is “60% reliable” would generate 6 correct responses out of 10.
- *Appropriate reliance/trust*: A match between the actual capabilities of the automation and the perceived capabilities by the human. Instances of appropriate reliance refer to situations in which the human correctly chooses to use or not use the automation based on knowledge they have about its capabilities (Lee & See, 2004).
- *Automation error*: An event for which the automation fails to provide accurate support. For example, in automated decision support systems, errors usually consist of false alarms and misses.
- *Automation reliability*: see *actual reliability*
- *Automation*: a technologically-based system used to partially or fully assist the human in tasks involving sensing, detecting, processing information, making decisions and/or executing actions. In this document, the word automation generally used to describe a system that helps sense, process information, and make decisions.
- *Decision support aid*: A type of automated system that is designed to help humans make decisions faster, more accurately, and/or with less effort.
- *Distribution of errors*: Within this investigation, this variable describes the distribution of errors as a function of time.
- *Error type*: Within this investigation, it described the two types of errors that a decision support aid can make (false alarms and misses)
 - *False alarm*: Occurs when the automation generates a warning or alert in the absence of a true signal.
 - *Miss*: Occurs when the automation fails to notify the human of a true signal.
- *Levels of Automation (LOA)*: The level of support that an automated system provides. Parasuraman et al. (2000) suggested that there are four levels of automation: information acquisition, information analysis, decision/action selection, and action implementation.
- *Perceived reliability*: A subjective measure commonly used in human-automation interaction studies. It is an estimate of the automation’s reliability by participants.
- *Reliability*: See *Actual reliability*
- *Reliance*: Within this study it is the behavior of agreeing with the automation without verifying other sources of information.
- *System error*: Within this document, the term system is used to describe the human-automation team or system. Therefore a system error is an instance in which the wrong output is generated by the system, irrespective of which of the system components (human or automation) caused it.
- *Trust in automation*: Generally defined as one’s willingness to use the automation in the face of uncertainty. Trust is often measured subjectively by asking participants to provide a rating of their trust in the automation. Objective proxies for trust in automation are objective measures such as reliance.

REFERENCES

- Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. *Psychological Bulletin*, 101, 41 – 74.
- Bailey, N. R. (2004). The effects of operator trust, complacency potential, and task complexity on monitoring a highly reliable automated system. Unpublished doctoral dissertation, Old Dominion University, Virginia.
- Bliss, J. (2003). An investigation of alarm related accidents and incidents in aviation. *International Journal of Aviation Psychology*, 13, 249 – 268.
- Bliss, J. P., & Dunn, M. C. (2000). Behavioural implications of alarm mistrust as a function of task workload. *Ergonomics*, 43, 1283-1300.
- Bliss, J.P., Dunn, M., & Fuller, B. (1995). Reversal of the cry-wolf effect: An investigation of two methods to increase alarm response rates. *Perceptual and Motor Skills*, 80, 1231-1242.
- Bliss, J. P., & Gilson, R. D. (1998). Emergency signal failure: implications and recommendations, *Ergonomics*, 41, 57-72.
- Bliss, J. P., Gilson, R. D., & Deaton, J. E. (1995). Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics*, 38, 2300-2312.
- Bliss, J. P., & McAbee, P. E. (1995). Alarm responses in a dual-task paradigm as a function of primary task criticality. *Proceedings of the Human Factors Society 39th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 1395 – 1399).
- Breznitz, S. (1984). Cry wolf: The psychology of false alarms, Hillsdale, NJ: LEA.
- Clamann, M. P., Wright, M. C. & Kaber, D. B. (2002). Comparison of performance effects of adaptive automation applied to various stages of human-machine system information processing. *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 342 – 346).
- Cohen, M.S., Parasuraman, R., & Freeman, J.T. (1998). Trust in decision aids: A model and its training implications. *Proceedings of the 1998 Command and Control Research and Technology Symposium*. Washington, DC: CCRP
- Cotte, N., Meyer, J., & Coughlin, J. F. (2001). Older and younger drivers' reliance on collision warning systems. *Proceedings of the Human Factors Society, 45th annual meeting*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 277 – 280).

- Crocoll, W. M., & Coury, B. G. (1990). Status or recommendations: Selecting the type of information for decision aiding. *Proceedings of the Human Factors Society, 34th annual meeting*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 1524 – 1528).
- Dijkstra, J.J, Liebrand, W.B.G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour and Information Technology, 17*, 155 – 163.
- Dixon, S. R., & Wickens, C. D. (2003). *Imperfect automation in unmanned aerial vehicle flight control*. Technical Report AHFD-03-17/MAAD-03-2. Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Dixon, S. R., & Wickens, C. D. (2004). *Reliability in automated aids for unmanned aerial vehicle flight control: Evaluating a model of automation dependence in high workload*. Technical Report AHFD-04-05/MAAD-04-1. Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting misuse and disuse of combat identification systems. *Military Psychology, 13*, 147 - 164.
- Dzindolet, M. T., Pierce, L., Pomranky, R., Peterson, S., & Beck, H. (2001). Automation reliance on a combat identification system. *Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 532 – 536).
- Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for Kit of Factor-Referenced Cognitive Tests. Letter Sets*. Princeton, NJ: Educational Testing Service (pp. VZ-2, 174, 176, 179).
- Endsley, M. R. & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in dynamic control task. *Ergonomics, 42*, 462 – 492.
- Fitts, P. M. (1951). *Human engineering for an effective air navigation and air traffic control system*. Ohio State University foundation report, Columbus, OH.
- Green, D. M. & Sweets, J. A. (1966). *Signal Detection Theory and psychophysics*. New York: Wiley
- Gupta, N., Bisantz, A. M., & Singh, T. (2001). Investigation of factors affecting driver performance using adverse condition warning systems. *Proceedings of the 45th Annual Meeting of the Human Factors Society*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 1699 – 1703).

- Jackson, D. N., & Messick, S. (1978). Problems in human assessment. NY: Krieger
- Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1), 53-71.
- Johnson, J. D. (2004). *Type of automation failure: the effects on trust and reliance in automation*. Unpublished masters thesis, Georgia Institute of Technology, Georgia.
- Kantowitz, B. H., Becker, C. A., & Barlow, S. T. (1993). Assessing driver acceptance of IVHS components. *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 1062 – 1066).
- Kantowitz, B. H., Hanowski, R. J., & Kantowitz, S. C. (1997). Driver reliability requirements for traffic advisory information. In Y. I. Noy (Eds.). *Ergonomics and Safety of Intelligent Driver Interfaces* (pp. 1-22). Mahwah, NJ: Lawrence Erlbaum Associates.
- Lee, J. D., & Moray, N. (1992). Trust, control strategies, and allocation of function in human-machine systems. *Ergonomics*, 35, 1243-1270.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence and operator's adaptation to automation. *International Journal of Human Computer Studies*, 40, 153-184.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46, 50-80.
- Lehto, M. R., Papastavrou, J. D., Ranney, T. A., & Simmons, L. A. (2000). An experimental comparison of conservative versus optimal collision avoidance warning system thresholds. *Safety Science*, 36, 185 – 209.
- Maltz, M., & Shinar, D. (2004). Imperfect in-vehicle collision avoidance warning systems can aid drivers. *Human Factors*, 46, 357 – 366.
- McCarley, J. S., Wiegmann, D. A., Wickens, C. D. & Kramer, A. F. (2003). Effects of age on utilization and perceived reliability of an automated decision-making aid for luggage screening. *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society
- McDowd, J. M. & Shaw, R. J. (2000). Attention and aging: a functional perspective. In F. I. M. Craik & T. A. Salthouse (Eds.). *The Handbook of Aging and Cognition* (pp. 221- 292). Mahwah, NJ: Lawrence Erlbaum Associates.
- Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, 46, 196 – 204.

- Meyer, J., Bitan, Y., Shinar, D., & Zmora, E. (1999). Scheduling of actions and reliance on warnings in a simulated control task. *Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 251 – 255).
- Molloy, R., & Parasuraman, R. (1996). Monitoring an automated system for a single failure: Vigilance and task complexity effects. *Human Factors*, 38, 311-322.
- Moray, N., Inagaki, T. & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, 6, 44-58.
- Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27, 527 – 539.
- National Research Council (1997). *Flight to the future: Human Factors in air traffic control*. Washington D.C.: National Academy Press
- O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kauffman, & J. Thomas (Eds.), *Handbook of perception and human performance: Volume II. Cognitive processes and performance*. New York: John Wiley.
- Parasuraman, R. (1987). Human-computer monitoring. *Human Factors*, 29, 695-706.
- Parasuraman, R., Molloy, R., & Singh, I. (1993). Performance consequences of automation-induced “complacency.” *The International Journal of Aviation Psychology*, 3, 1-23.
- Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors*, 38, 665-679.
- Parasuraman, R. & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39, 230-253.
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, & Cybernetics*, 30, 286-297.
- Price, H. E. (1985). The allocation of functions in systems. *Human Factors*, 27, 33-45.
- Pritchett, A. R. (2001). Reviewing the role of cockpit alerting systems. *Human Factors and Aerospace Safety*, 1, 5-38.

- Riley, V. (1996). Operator reliance on automation: theory and data. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance* (pp. 19-36). Mahwah, NJ: Lawrence Erlbaum Associates.
- Rovira, E., Zinni, M., & Parasuraman, R. (2002). Effects of information and decision automation on multi-task performance. *Proceedings of the Human Factors & Ergonomics Society 46th Annual Meeting*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 327 – 331).
- Sanchez, J. (2005). Human-automation interaction: Factors that affect human behavior and system performance. Unpublished preliminary examination. Georgia Institute of Technology, Georgia.
- Sanchez, J., Fisk, A. D. & Rogers, W. A. (2004). Reliability and age-related effects on trust and reliance of a decision support aid. *Proceedings of the 48th Annual Meeting of the Human Factors Society*. Santa Monica, CA: Human Factors and Ergonomics Society (pp. 586 – 589).
- Sheridan, T. B. (2002). *Humans and automation: system design and research issues*. Santa Monica, CA: John Wiley & Sons, Inc.
- Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators (Man-Machine Systems Laboratory Report). Cambridge: MIT.
- Shipley, W. C. (1986). *Shipley Institute of Living Scale*. Los Angeles: Western Psychological Services.
- Sit, R. A., & Fisk, A. D. (1999). Age-related performance in a multiple-task environment. *Human Factors*, 41, 26 – 34.
- Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51, 991-1006.
- Skitka, L. J., Mosier, K. L., & Burdick, M. (2000). Accountability and automation bias *International Journal of Human-Computer Studies*, 52, 701 - 717.
- St. John, M., & Manes, D. I. (2002). Making unreliable automation useful. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, Santa Monica, CA: Human Factors and Ergonomics Society (pp. 332 – 336).
- Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, 58, 719 – 735.
- Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale – Revised*. New York: Psychological Corp.

- Wickens, C. D., & Dixon, S. R. (2005). *Is there a magic number 7 (to the minus 1)? The benefits of imperfect diagnostic automation: A synthesis of the literature*. Technical Report AHFD-05-01/MAAD-05-01. Savoy, IL: University of Illinois, Aviation Human Factors Division.
- Wickens, C. D., Helleberg, J., & Xu, X. (2002). Pilot maneuver choice and workload in free flight. *Human Factors*, 44, 171 – 188.
- Wickens, C. D., & Xu, X. (2002). Automation trust, reliability and attention. Institute of Aviation Tech. Report AHFD-02-14 / MAAD-02-2. Savoy: University of Illinois, Aviation Research Lab.
- Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: the effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomic Science*, 2, 352 – 367.
- Xu, X., Wickens, C. D. & Rantanen, E. (2004) *Imperfect conflicting alerting systems for cockpit display of traffic information*. Technical Report AHFD-04-8/NASA-04-2.