

Designing an Error Resolution Checklist for a Shared Manned-Unmanned Environment

by

Jacqueline M. Tappan

BASc Systems Design Engineering
University of Waterloo, 2008

Submitted to the Engineering Systems Division
in partial fulfillment of the requirements for the degree of

Master of Science in Engineering Systems
at the
Massachusetts Institute of Technology

June 2010

© 2010 Massachusetts Institute of Technology. All rights reserved.

Signature of Author _____
Jacqueline Tappan
Engineering Systems Division
May 2010

Certified by _____
Mary L. Cummings
Associate Professor of Engineering Systems, Aeronautics and Astronautics
Thesis Supervisor

Accepted by _____
Nancy Leveson
Professor of Engineering Systems, Aeronautics and Astronautics
Chair, Engineering Systems Division Education Committee

Report Documentation Page		Form Approved OMB No. 0704-0188
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.		
1. REPORT DATE MAY 2010	2. REPORT TYPE	3. DATES COVERED 00-00-2010 to 00-00-2010
4. TITLE AND SUBTITLE Designing an Error Resolution Checklist for a Shared Manned-Unmanned Environment		5a. CONTRACT NUMBER
		5b. GRANT NUMBER
		5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER	
	5e. TASK NUMBER	
	5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Massachusetts Institute of Technology,Engineering Systems Division,Cambridge,MA,02139		8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)		10. SPONSOR/MONITOR'S ACRONYM(S)
		11. SPONSOR/MONITOR'S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited		
13. SUPPLEMENTARY NOTES		
14. ABSTRACT The role of unmanned vehicles in military and commercial environments continues to expand resulting in Shared Manned-Unmanned (SMU) domains. While the introduction of unmanned vehicles can have many benefits, humans operating within these environments must shift to high?level supervisory roles, which will require them to resolve system errors. Error resolution in current Human Supervisory Control (HSC) domains is performed using a checklist; the error is quickly identified, and then resolved using the steps outlined by the checklist. Background research into error resolution identified three attributes that impact the effectiveness of an error resolution checklist: domain predictability, sensor reliability, and time availability. These attributes were combined into a Checklist Attribute Model (CAM) demonstrating that HSC domains with high levels of complexity (e.g. SMU domains) are illsuited to error resolution using traditional checklists. In particular, it was found that more support was required during such error identification, as data is uncertain and unreliable. A new error resolution checklist, termed the GUIDER (Graphical User Interface for Directed Error Recovery) Probabilistic Checklist, was developed to aid the human during the error identification process in SMU domains. Evaluation was performed through a human performance experiment requiring participants to resolve errors in a simulated SMU domain using the GUIDER Probabilistic Checklist and a traditional checklist tool. Thirty-six participants were recruited, and each was assigned to a single checklist tool condition. Participants completed three simulated error scenarios. The three scenarios had varying sensor reliability levels (low, medium, high) to gauge the impact of uncertainty on the usefulness of each checklist tool. The human performance experiment showed that the addition of error likelihood data using an intuitive visualization through the GUIDER Probabilistic Checklist improved error resolution in uncertain settings. In settings with high certainty, there was no difference found between the performances of the two checklists. While positive, further testing is required in more realistic settings to validate both the effectiveness of the GUIDER Probabilistic Checklist tool and the Checklist Attribute Model.		

15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 153	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

Designing an Error Resolution Checklist for a Shared Manned-Unmanned Environment

by

Jacqueline M. Tappan

Submitted to the Engineering Systems Division

On May 10th, 2010, in partial fulfillment of the requirements for the Degree of
Master of Science in Engineering Systems

Abstract

The role of unmanned vehicles in military and commercial environments continues to expand, resulting in Shared Manned-Unmanned (SMU) domains. While the introduction of unmanned vehicles can have many benefits, humans operating within these environments must shift to high-level supervisory roles, which will require them to resolve system errors. Error resolution in current Human Supervisory Control (HSC) domains is performed using a checklist; the error is quickly identified, and then resolved using the steps outlined by the checklist.

Background research into error resolution identified three attributes that impact the effectiveness of an error resolution checklist: domain predictability, sensor reliability, and time availability. These attributes were combined into a Checklist Attribute Model (CAM), demonstrating that HSC domains with high levels of complexity (e.g. SMU domains) are ill-suited to error resolution using traditional checklists. In particular, it was found that more support was required during such error identification, as data is uncertain and unreliable.

A new error resolution checklist, termed the GUIDER (Graphical User Interface for Directed Error Recovery) Probabilistic Checklist, was developed to aid the human during the error identification process in SMU domains. Evaluation was performed through a human performance experiment requiring participants to resolve errors in a simulated SMU domain using the GUIDER Probabilistic Checklist and a traditional checklist tool. Thirty-six participants were recruited, and each was assigned to a single checklist tool condition. Participants completed three simulated error scenarios. The three scenarios had varying sensor reliability levels (low, medium, high) to gauge the impact of uncertainty on the usefulness of each checklist tool.

The human performance experiment showed that the addition of error likelihood data using an intuitive visualization through the GUIDER Probabilistic Checklist improved error resolution in uncertain settings. In settings with high certainty, there was no difference found between the performances of the two checklists. While positive, further testing is required in more realistic settings to validate both the effectiveness of the GUIDER Probabilistic Checklist tool and the Checklist Attribute Model.

Thesis Supervisor: Mary L. Cummings

Title: Associate Professor of Engineering Systems, Aeronautics and Astronautics

Acknowledgements

I owe thanks to many people who contributed to the successful completion of this thesis.

First, a special thank you to Missy Cummings, my research advisor. Thank you for taking me on as a student, believing in my potential as a researcher, guiding me through the systems engineering process, and providing feedback (both positive and negative) week in and week out. I know that you have prepared me to achieve my career aspirations in academia and industry.

Thank you to Seth Teller and all other AR team members. Thanks for giving me a crash course in Linux and teaching me the basics of speech recognition and path planning. To William Li and Stephen Shum, thanks for inviting me along to Fort Hood to experience a real SSA. It was both educational, and thanks to your company, extremely enjoyable.

Thanks to Director, Defense Research & Engineering (DDR&E), which funded and supported this research.

I would like to thank Farzan Sasangohar, my thesis mentor, PhD mentor, and life mentor. Thank you for helping me to write and revise this thesis, and for your many thoughtful suggestions. It was great to share my office with a fellow Canadian, and to rejoice together at our team's performance during the Olympics.

Thank you to Birsen Donmez, who took an active role in my statistics instruction. I appreciate your time and patience in teaching me the finer details of "varsity statistics". Also, thank you for being so responsive through e-mail, even though you are now located in Toronto.

Thank you to my three UROPs, Scott Bezek, Peter Huang, and Rich Chan. Thanks for being my interface between HAL and CSAIL, and for all of your hard work in the development of the EIR simulation interface.

Thank you to all of the HAL graduate students past and present who welcomed me into the lab, and who I now consider friends: Farzan, Dave, Thomas, JR, Yves, Christin, Andrew, and many others.

To my family: Mom and Dave, Dad and She, Leo and Joanne. I am forever grateful for your unwavering support through even more years of schooling (and perhaps more yet to come). Thanks for the phone calls of encouragement, the care packages to boost my spirits, and the frequent visits to Boston.

I owe the most thanks to my best friend Daryl, who moved with me to Cambridge and has supported my ambitions from Day 1. There are not words enough to show how grateful I am to you; thank you does not seem to be enough. Hopefully one day I can repay the favor.

Table of Contents

Abstract.....	3
Acknowledgements	5
Table of Contents.....	7
List of Figures.....	11
List of Tables	13
List of Acronyms.....	15
Chapter 1. Introduction.....	17
1.1. Motivation.....	17
1.2. Problem statement.....	19
1.3. Research objectives.....	19
1.4. Representative SMU domain	20
1.5. Thesis overview	20
Chapter 2. Background	23
2.1. Classifying complex supervisory domains	23
2.1.1. Causal domains.....	23
2.1.2. Intentional domains	24
2.2. System sensor quality	26
2.3. Errors in SMU environments	27
2.4. Traditional checklists	28
2.5. HSC domain attributes for error identification	29
2.5.1. Domain predictability	30
2.5.2. Sensor reliability.....	30
2.5.3. Time availability.....	30
2.5.4. Checklist Attribute Model (CAM)	31
2.6. Alternative checklist design considerations.....	33
2.6.1. Defining role of automation	33
2.6.2. Human role in error identification.....	35
2.6.3. Visualization of error likelihoods.....	37
2.7. Summary	43
Chapter 3. GUIDER Probabilistic Checklist.....	45
3.1. Error identification.....	45
3.2. Error recovery	46
3.3. Application: Robotic forklift checklist	46
3.3.1. Current SSA operations.....	47
3.3.2. SSA domain with autonomous forklifts	50
3.3.3. Sources of error	53
3.3.4. The GUIDER representation	56
3.3.5. Error Identification and Recovery (EIR) display	57
3.4. Revised Checklist Attribute Model (CAM).....	63
3.5. Example domain classifications.....	65

3.5.1. Commercial aviation	65
3.5.2. Forklift domain	67
3.6. Summary	69
Chapter 4. Experimental Evaluation	71
4.1. EIR simulation	71
4.1.1. Error scenarios	71
4.1.2. EIR display	72
4.2. Hypotheses	79
4.2.1. Performance	79
4.2.2. Cognitive Strategies	80
4.2.3. Subjective feedback	81
4.3. Apparatus	81
4.4. Participants	82
4.5. Procedure	83
4.5.1. Pre-experiment interaction	83
4.5.2. Training	83
4.5.3. Error scenarios	84
4.5.4. Post-experiment questionnaire	85
4.6. Experiment design	85
4.7. Summary of performance metrics	86
4.8. Summary	87
Chapter 5. Results	89
5.1. Number of error confirmations	90
5.2. Cognitive strategies	93
5.2.1. Information collection	94
5.2.2. Information emphasis	94
5.3. Subjective feedback	96
5.3.1. Questionnaire data	96
5.3.2. General participant feedback	97
5.4. Discussion of experimental findings	97
5.5. Summary	100
Chapter 6. Conclusions and Future Work	101
6.1. Experimental results	102
6.2. Design recommendations	103
6.2.1. Certainty indicator	103
6.2.2. Combined pie chart graphic	104
6.2.3. Indication of selected errors	104
6.2.4. Limiting error sources	105
6.3. Experiment recommendations and future work	105
Appendix A: Descriptive Statistics	107
Appendix B: Consent to Participate	109
Appendix C: Demographic Questionnaire	113
Appendix D: Training Tutorials	115

Appendix E: Training Video Script 125

Appendix F: User Interaction Questionnaire..... 127

Appendix G: Randomization of Participants 129

Appendix H: Collected Data 131

Appendix I: Statistical Assumption Tests..... 141

Appendix J: Detailed Statistical Results 143

References 151

List of Figures

Figure 1: Shelves of emergency procedures at Chattanooga nuclear power plant simulator. ____	24
Figure 2: SMU domain with multiple unmanned vehicles. _____	26
Figure 3: Shadow 200 UAV checklist. _____	28
Figure 4: Human supervisory domain attributes for checklist design. _____	32
Figure 5: Decision tree visualization. _____	40
Figure 6: Treemap visualization. _____	41
Figure 7: Multi-level pie chart visualization. _____	42
Figure 8: Notional layout of a U.S. Army SSA. _____	48
Figure 9: Truck delivering pallets of materials to receiving area of an SSA. _____	48
Figure 10: Bulk storage of materials in an SSA. _____	49
Figure 11: Customer vehicles waiting to receive requested materials from SSA. _____	49
Figure 12: Tracking of items located in bulk area of SSA. _____	50
Figure 13: Entities operating within the forklift domain. _____	51
Figure 14: Tablet PC user interface for directing the RF in SSA. _____	52
Figure 15: Probabilistic error tree summarizing the potential forklift errors. _____	56
Figure 16: GUIDER representation of probabilistic error tree. _____	57
Figure 17: Identification screen in Error Identification and Recovery (EIR) display. _____	59
Figure 18: Recovery screen in Error Identification and Recovery (EIR) display. _____	59
Figure 19: <i>Multiple pallets designated</i> suggested by system as error source. _____	61
Figure 20: Diagnostic test inspecting Operator Tablet view. _____	62
Figure 21: <i>Non-pallet designated</i> selected as error source. _____	62
Figure 22: GUIDER incorporated into Error Identification and Recovery (EIR) display. ____	63
Figure 23: Revised Checklist Attribute Model (CAM). _____	64
Figure 24: Identification screen, GUIDER Checklist. _____	74
Figure 25: Identification screen, Traditional Checklist. _____	74
Figure 26: Map of forklift environment for Scenario 1. _____	76
Figure 27: Recovery screen after incorrect confirmation, GUIDER Checklist. _____	77
Figure 28: Recovery screen after incorrect confirmation, Traditional Checklist. _____	77
Figure 29: Recovery screen after correct confirmation, GUIDER Checklist. _____	78
Figure 30: Recovery screen after correct confirmation, Traditional Checklist. _____	78

Figure 31: Apparatus setup for experiment. _____	82
Figure 32: Contextual background screen for low reliability error scenario. _____	85
Figure 33: Effect of checklist on <i>number of error confirmations</i> . _____	91
Figure 34: Effect of reliability on <i>number of error confirmations</i> . _____	91
Figure 35: Effect of checklist and reliability on <i>number of error confirmations</i> . _____	93

List of Tables

Table 1: Generic error hierarchy with associated likelihood values._____ 39

Table 2: Summary of potential errors in robotic forklift field operations. _____ 54

Table 3: Sensor groups and related pallet approach errors. _____ 72

Table 4: Sensor reliability levels for simulated error scenarios. _____ 72

Table 5: Suggested and true error source for each error scenario. _____ 75

Table 6: Descriptive statistics for *number of error confirmations*. _____ 92

Table 7: Error resolution strategies of participants. _____ 95

Table 8: Summary of experimental results. _____ 98

List of Acronyms

AR	Agile Robotics
ATC	Air Traffic Control
CAM	Checklist Attribute Model
DOD	Department of Defense
EIR	Error Identification and Recovery
GPS	Global Positioning System
GUIDER	Graphical User Interface for Directed Error Recovery
HSC	Human Supervisory Control
HDP	High Domain Predictability
HSR	High Sensor Reliability
HTA	High Time Availability
LED	Light-Emitting Diode
LDP	Low Domain Predictability
LSR	Low Sensor Reliability
LTA	Low Time Availability
MDP	Medium Domain Predictability
MIT	Massachusetts Institute of Technology
MSR	Medium Sensor Reliability
MTA	Medium Time Availability
PRA	Probability Risk Assessment
RF	Robotic Forklift
RFID	Radio-Frequency Identification
SSA	Supply Support Activity
SMU	Shared Manned-Unmanned
UAS	Unmanned Aerial Systems
UAV	Unmanned Aerial Vehicle
UGV	Unmanned Ground Vehicle
UV	Unmanned Vehicle

Chapter 1. Introduction

The role of Unmanned Vehicles (UVs) is increasingly expanding in both simple and complex domains. The U.S. Department of Defense (DOD) plans to invest \$17 billion in Unmanned Aircraft Systems (UAS) between 2008 and 2013, while between 2000 and 2008, the inventory of UAS in DOD components rose from 50 aircraft to more than 6000 (GAO, 2008). Aviation, however, is not the only domain where such unmanned system expansion is occurring, with unmanned ground vehicles (UGV) being introduced to both commercial customers and individuals. Examples include KIVA Systems, which manufactures autonomous robots for warehouse operations for companies such as Amazon[®] and Zappos[®] (Scanlon, 2009). iRobot[®], which introduced the *Roomba*, a robot vacuum cleaner in 2002, has sold 3 million of the robots (The Economist, 2009) and continues to be a driving force in the growth of personal robots for the home (iRobot Corporation, 2009), as well as for the U.S. Army with the *PackBot* (iRobot Corporation, 2009).

In hostile environments where work tasks endanger human operators, the inclusion of unmanned vehicles to fulfill these duties may not only increase system safety, but also improve operating efficiency. Military equipment distribution warehouses are examples of hostile environments that are expanding to include unmanned vehicles. Currently, work is underway at the Massachusetts Institute of Technology (MIT) to develop a system of autonomous forklifts to distribute pallet-loaded supplies in these warehouses that are located in war zones (Chandler, 2009). The inclusion of robotic forklifts in these complex, unstructured, and sometimes hostile military environments has the potential to streamline activities and increase overall throughput, while reducing customer wait times and potentially saving human lives.

1.1. Motivation

Although the potential benefits of automation in such hostile environments are significant, there are many human factors issues that are associated with the introduction of autonomous vehicles into such complex environments. Chief among these is the changing role of human operators from direct and manual control of a system to being involved in higher-level planning and

decision-making (Cummings, Bruni, Mercier, & Mitchell, 2007). This shift to a supervisory role requires the human operator to undertake a number of new functions (Sheridan, 1992), including:

- Scheduling tasks and planning tasks
- Monitoring the actions of autonomous entities in the system and detecting failures
- Intervening when required to return the system to the desired state

All Human Supervisory Control (HSC) domains require monitoring for failures and overcoming error states to ensure high productivity while maintaining the safety of both humans and autonomous entities. Error resolution in supervisory control systems continues to take the form of serial checklists, either paper or electronic, that can serve as memory aid tools, ensuring that all required recovery steps have been executed (Gawande, 2009). Traditional checklists begin with an assumed error source and present recovery steps serially. These checklists suit domains where system behavior is predictable and there is consistent performance feedback. In such domains, the error source is relatively easy to identify and error resolution can (and should) begin immediately. Traditional checklists may be inappropriate for autonomous vehicle domains with high complexity, however, as these environments can be highly unpredictable and lack the clear information feedback loops present in the more predictable supervisory control domains of piloting aircraft and monitoring nuclear power generation systems.

Shared Manned-Unmanned (SMU) domains are a subset of HSC domains that incorporate both autonomous vehicles and human operators interacting within a single system. While a high level of complexity¹ typically characterizes all HSC domains (Cummings, Kirschbaum, Sulmistras, & Platts, 2006), SMU domains can have enhanced complexity levels due to the large number of distinct entities operating and interacting within the system. Uncertainty in SMU domains also exists, as human behavior is less deterministic and system boundaries are often undefined, resulting in unpredictable environmental factors acting on the domain. With such high complexity and uncertainty levels, error resolution in these systems can become complicated and could benefit from additional diagnostic information from various sources. As the prevalence of SMU domains will only continue to increase as technology advances, development of an error

¹ Complexity is defined by Merriam-Webster as “the quality or state of being hard to separate, analyze, or solve.”

resolution tool designed specifically for such environments is critical to ensure domain efficiency and the safety of humans operating within the system. Unfortunately, alternative error resolution tools, or checklists, designed specifically for highly complex SMU environments have not yet been developed.

1.2. Problem statement

Current error recovery checklists designed for use in traditional HSC domains, such as aviation and process control, are ill-suited to the unique characteristics of SMU domains, which are increasingly occurring in military, commercial, and various consumer environments. A new kind of error resolution tool, which allows human supervisors in SMU domains to overcome system errors efficiently, while maintaining domain safety, is required. This research proposes the development of this alternative checklist, which will be designed to satisfy the unique needs of SMU domains.

1.3. Research objectives

This research has three objectives:

- Identify the important attributes of HSC domains, their relationship to each other, and how they can be combined to establish a domain attribute model, which can be used to categorize HSC domains.
- Develop a new error resolution tool designed specifically for HSC domains that include autonomous vehicles. The design of this tool can be guided by previous research in serial checklists, complex work domains, automation, human decision-making, and information visualization.
- Evaluate the new error resolution tool against a traditional checklist tool to determine which is more effective in supporting error resolution in a representative SMU domain.

The objective of this new checklist is to improve the error identification and recovery process in SMU domains, ensuring that human supervisors within these environments can accurately and efficiently identify the source of an error, recover from the error, and transition the system back into an operational state.

1.4. Representative SMU domain

This research employs the previously discussed autonomous forklift domain to demonstrate the new, alternative checklist. Currently, military distribution warehouses, which store and distribute items required for U.S. Army active duty, utilize manually operated forklifts. Complexity and uncertainty in the domain is already high, as the operations of multiple manually operated forklifts must be coordinated to ensure that the warehouse environment runs efficiently. The war-zone environment introduces potential unpredictable events resulting from militant actions, and environmental factors, such as rain and wind, can negatively impact the open-air warehouse.

To increase warehouse efficiency, as well as to remove humans from the dangerous and exposed position of manually operating forklifts, the U.S. Army has proposed automation in the form of robotic forklifts. The introduction of these robotic forklifts will result in an SMU domain. This envisioned SMU domain will incorporate robotic forklifts (RFs), ground-level human operators who will interact with and direct the RFs in their tasks, and a high-level human supervisor who will monitor both the RFs and the human operators. While the current military warehouse environment has both high complexity and uncertainty, the addition of RFs only increases the unpredictability of the environment, and as a direct result, traditional error resolution checklists may not be appropriate. Therefore, the alternative checklist developed through this research will be applied to the autonomous forklift domain to determine whether error resolution efficiency in this representative SMU domain improves with this new tool.

1.5. Thesis overview

This thesis contains the following chapters:

- Chapter 1, *Introduction*, describes the motivation and research objectives of this thesis.
- Chapter 2, *Background*, outlines the current state of error resolution in supervisory control systems. This chapter identifies the characteristics of HSC domains that are important in error resolution, with three domain attributes identified: domain predictability, sensor reliability, and time availability. These attributes are combined into an attribute model that categorizes HSC domains, and identifies shortcomings of current

checklist tools. This chapter also presents relevant research material that guides the design and development of the alternative checklist tool.

- Chapter 3, *GUIDER Probabilistic Checklist*, uses the background research detailed in Chapter 2 to develop a new error resolution tool. The features of this probabilistic tool are demonstrated using the autonomous forklift project. Utilizing the domain attribute model developed in Chapter 2, two case studies are conducted to form an experimental hypothesis as to the best error resolution system for SMU domains.
- Chapter 4, *Experimental Evaluation*, describes the human performance experiment, incorporating a simulation of the SMU autonomous forklift domain, used to test the hypothesis of this research. Details include a discussion of participants, procedures, and experimental design.
- Chapter 5, *Results*, presents the findings of the human performance experiment using such metrics as number of error confirmations, cognitive strategy, and subjective appeal for both a traditional-style checklist and the GUIDER Probabilistic Checklist.
- Chapter 6, *Conclusions and Future Work*, compares the results of the human performance experiment with the research hypotheses. The chapter also provides a set of design and future experimental recommendations based upon the experimental results. The chapter concludes with a description of the future work necessary to generalize this research and integrate alternative error resolution methods into current and future practice.

Chapter 2. Background

In this chapter, common characteristics of HSC domains are investigated to determine the functionality that would be required for a new error resolution checklist in an SMU domain. An examination of current checklist systems is also performed in order to determine the shortcomings of current error resolution systems when applied to SMU environments.

Through this initial research, three HSC domain attributes are identified that have an impact on error resolution: domain predictability, sensor reliability, and time availability. These attributes are used to identify the current HSC domains that are well-suited to error resolution with traditional checklists, as well as those domains that are currently not well served, providing justification for the development of a new error resolution tool to be used in select HSC domains.

Relevant research in the fields of automation, human judgment under uncertainty, and information visualization is reviewed to guide the development of the alternative checklist.

2.1. Classifying complex supervisory domains

HSC domains can be grouped into two major categories: causal domains and intentional domains. By gaining an understanding of the characteristics of HSC domains, and which category individual domains fall into, the functionality required of an error resolution checklist can be better understood.

2.1.1. Causal domains

Causal domains are closed-loop systems that are isolated from their environment (Vicente, 1999). These domains have a direct feedback loop between the current state of the system and future system actions ensuring that the goals of the domain are continually met. Causal systems generally operate in predictable ways, as a result of clear constraints (Cummings & Guerlain, 2003; Wong, Sallis, & O'Hare, 1998). These constraints include behavior being dictated by laws of physics and the system having clear boundaries. An example of a causal domain is a power generation plant.

Supervisors in causal domains are responsible for monitoring the physical health status of the system, which can be closely observed through the extensive use of state sensors. If an error occurs in a causal system, the source of the error is relatively easy to pinpoint. With the error source identified, supervisors in these systems only need to recover from the failure and transition the causal system back to an operational state. Emergency checklists are usually used for error recovery, with the required steps printed on paper and stored in procedure books or included electronically as part of a computer system within the domain. An example of emergency checklist books within the Chattanooga nuclear power plant simulator is shown in Figure 1 (U.S. Nuclear Regulatory Commission, 2010).



Figure 1: Shelves of emergency procedures at Chattanooga nuclear power plant simulator.

2.1.2. Intentional domains

Intentional domains are considered to be open-loop and are subject to external influences (Vicente, 1999). Instead of goals being met by a clear feedback loop that dictates future system actions, outcomes are driven by motivations of individuals and groups that are part of the organization active in the domain: 1) the individual acting in a supervisory role, and 2) individuals and groups outside of the system whose actions can impact operations. An example of an intentional domain is command and control.

There is a high level of unpredictability and uncertainty within intentional domains, with unanticipated events likely to occur. According to (Cummings & Guerlain, 2003; Wong et al., 1998), this is in part due to:

- Human decisions directly dictating system behavior, as opposed to laws of physics dictating system behavior
- Systems not having obvious boundaries, and as a result, being influenced by highly uncertain environmental factors that cannot be controlled or anticipated

If an error occurs in an intentional system, the source of the error is more difficult to identify, due to high levels of uncertainty. Therefore, basic emergency checklists, which focus on error recovery once the error source has been identified, may not be a viable option within these domains.

SMU domains

SMU domains fall under the categorization of intentional domains, but can have increased levels of unpredictability due to the inclusion of autonomous vehicles within the environment. In traditional intentional domains, the main system entities are humans, who perform manual operating tasks within the system or high-level supervision tasks of human operators and statically located automation. In SMU domains, however, there is the addition of mobile autonomous vehicles. As a result, the number of distinct entities increases, and may include:

- Autonomous vehicles
- Human operators that direct autonomous vehicles
- Human operators that manually operate vehicles
- High-level human supervisor monitoring the entire system

With an increase in the number of distinct domain entities, there is also an increase in the number of different interactions occurring between the entities within the system, as represented in Figure 2 (Naval Research Laboratory, 2006). As a result, SMU domains have increased complexity over general intentional domains, and are often far more complex than causal-based systems.

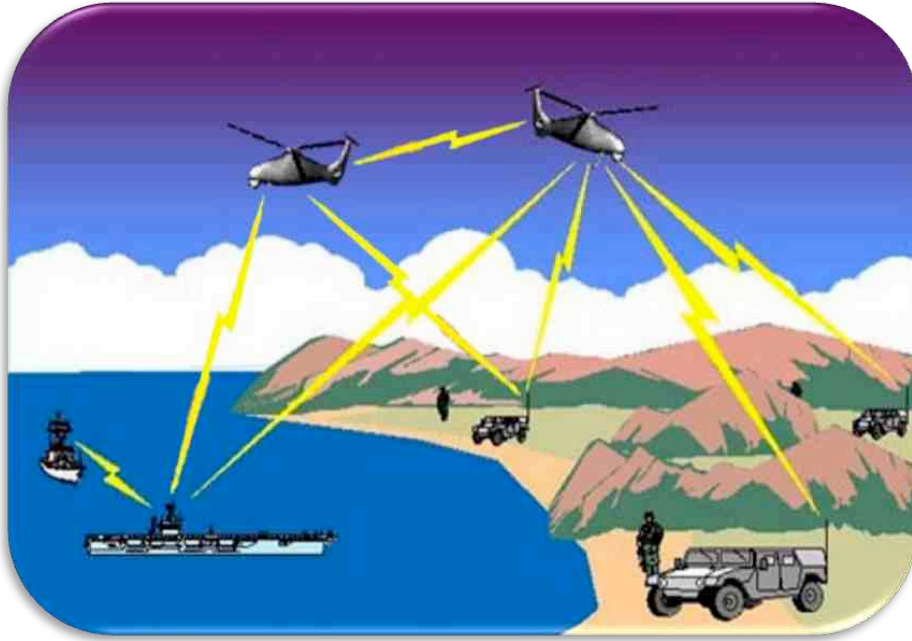


Figure 2: SMU domain with multiple unmanned vehicles.

2.2. System sensor quality

In order to operate autonomous vehicles within an unpredictable environment, it is vital that autonomous vehicles can detect environmental cues that guide their actions and behavior. The ability to detect cues from the environment ensures that these vehicles do not have to be consistently teleoperated by a human operator, but can instead independently select behaviors in order to fulfill mission objectives.

A system has reliable sensors if they consistently and accurately measure intended system parameters and states. Unfortunately, in many complex work environments there can be uncertainty associated with the data available to operators, resulting in the current state of the system being unclear (Vicente & Rasmussen, 1990; Vicente, 1999). For example, LIDAR (Light Detection and Ranging) sensors, which find the range of a distant target and allow an autonomous vehicle to sense potential obstacles in the environment, could provide erroneous data due to both systematic errors (e.g., laser detector bias) or random errors (e.g., signal-to-noise ratio, type of terrain, transmission properties of the atmosphere) (Huising & Pereira, 1998).

Low sensor reliability impacts the accuracy of the data that may be used by both automation and a high-level human supervisor for error resolution. If current system data transmitted by automation to the human supervisor is inaccurate, identification of the source of system failure could become more difficult and may lead to errors. In addition, if the source of system failure is incorrectly identified, domain efficiency and human safety within the domain could be compromised.

2.3. Errors in SMU environments

Reason (1990) argues that human error occurs when “a planned sequence of mental or physical activities fails to achieve its intended outcome” (p. 9). Unfortunately, in SMU systems, the source of system failure may not only be human-related. In such domains, autonomous entities present in the system can experience logic-based errors (resulting from pre-programmed coding mistakes) and component failures, with both potentially resulting in undesirable system behavior. When supervising SMU domains, it is important for human supervisors to be able to resolve all errors, both human and autonomy generated, as quickly as possible in order to return the system to normal operating conditions. This process, which includes identification of the error source and recovery from the identified error, can be grouped together using the term error resolution.

When a failure occurs in an HSC domain, it is always important to identify the source of the failure and recover from the failure as promptly as possible, in order to transition the system back into an operational state. In some time-critical domains, however, both human life and the integrity of the system depend on efficient resolution of the error state. If an error is not resolved within a short, limited duration, planes can crash, nuclear reactors can meltdown, and patients can die. To support supervisory-level error recovery, as well as maintain efficiency in error resolution, checklists are often implemented in an assistive role. These checklists guide the supervisor, step by step, through the recovery process. As supervisory control systems are utilized more frequently for monitoring complex work domains, checklists for error resolution have been implemented widely in workstations (Commission on Engineering and Technical Systems, 1997).

2.4. Traditional checklists

Traditional checklists present procedural steps serially as an aid to memory, ensuring that all required steps in some process are executed. An example of a traditional checklist, in paper-based form, is presented in Figure 3 (Department of the Army, 2004). This checklist is for the Shadow 200 unmanned aerial vehicle (UAV), which is flown by the U.S. Army and Marine Corps for surveillance, targeting, and reconnaissance (AAI Corporation, 2010). Even though the UAV is flown remotely, and therefore, the loss of the aircraft does not translate directly into the loss of human life, it is vital that system failures do not result in damage to the structural integrity of these very costly machines. Therefore, it is necessary that ground-control pilots be given procedures to resolve all conceivable Shadow 200 emergencies, including engine failure, fuse failure, and high engine temperature.

AV High Engine Temperature

CAUTION
Climb to a safe altitude may be continued for no more than 2 minutes. Once at safe altitude and level, if over-temperature condition persists for more than 2 minutes RTB. If poor WOT RPM and climb performance is observed RTB. Failure to comply could result in loss of AV.

CAUTION
Two or more indications (High CHT, High RAO, Low WOT RPM, Low Rate of Climb) are present RTB for engine investigation. Failure to comply could result in loss of AV.

1. Altitude Level (AVO)

2. Airspeed 70 Knots (AVO)

3. Temps..... Monitor (AVO)

Temps remain high

4. Land..... As Soon As Possible (AVO)

Temps normal

5. Mission..... Continue (AVO)

Figure 3: Shadow 200 UAV checklist.

Checklists are generally implemented in HSC domains in two capacities:

- *Normal checklists*: used as a memory aid for completing routine procedures.
- *Emergency checklists*: used during error situations to recover from one or more system failures and transition the system back into an operational state.

Human supervisors are typically well trained on how and when to utilize Normal checklists. During system operation, there are predetermined time slots when particular tasks need to be completed before operations can progress. For example, airplane pilots must complete a “Pre-Landing” checklist, which includes lowering the landing gear, extending the landing spoilers, and braking as required (Transport Canada, 2001). Interaction with Emergency checklists is less structured, however, as failure occurrences are often difficult to foresee. When using an Emergency checklist, the supervisor not only has to complete predetermined recovery steps, but also needs to be able to identify the source of the error so that the appropriate checklist can be selected.

Often, traditional checklists will begin with an assumed error source that has been identified through automated sensor and/or human feedback. The human supervisor then proceeds serially through recovery steps. Traditional checklists are therefore appropriate for domains where system behavior is predictable and there is consistent system state feedback. In such causal systems, the error source is relatively straightforward to identify, system complexity is relatively understood, and sensor reliability is high. With this straightforward error identification process, error recovery can (and should) begin immediately.

If the environment is intentional with the enhanced complexity of SMU interactions, the source of the error may be difficult to identify due to the unpredictability and uncertainty within the domain. The reliability of the sensors located on the autonomous vehicles in SMU environments can also lead to uncertainty in error identification. Attention will likely need to focus on the error identification process. Hence traditional checklists, which focus only on error recovery, may need to be modified to be appropriate for such domains.

2.5. HSC domain attributes for error identification

Based on this background research, three HSC domain attributes deemed to have an impact on the error resolution process were identified: domain predictability, sensor reliability, and time availability.

It is important to note that these HSC attributes predominantly impact the error identification portion of error resolution, as uncertainty and inaccuracy will make the identification of the error source more complicated. Error recovery, on the other hand, is not impacted by these domain attributes. Once the error has been accurately identified, the required recovery steps to resolve the error will remain the same, regardless of the level of domain predictability, sensor reliability, and time availability in the system.

2.5.1. Domain predictability

The two HSC domain classifications, causal and intentional, dictate the predictability of domain behavior. A system is predictable if it has well-defined boundaries; inputs and outputs into the system are known and documented, making unanticipated events unlikely (Vicente, 1999). Low Domain Predictability (LDP) could be seen as a characteristic of intentional domains, while High Domain Predictability (HDP) could be seen as a characteristic of causal domains. As previously noted, error identification may be difficult in intentional domains, as there are high levels of uncertainty and complexity, which can be further enhanced in SMU environments.

2.5.2. Sensor reliability

Sensor reliability assesses how accurately the system sensors measure intended system parameters and states. An HSC domain could have sensors whose reliabilities range from Low Sensor Reliability (LSR) to High Sensor Reliability (HSR). For example, inherent characteristics of the domain environment, such as blowing sand in war zones located within desert climates, can negatively impact the accuracy of domain sensors, resulting in the feedback of low reliability data to the system supervisor. This inaccurate data may complicate error identification when a system failure occurs, as the supervisor will be uncertain whether they can trust the data provided from the sensors, and utilize it during identification of the error source.

2.5.3. Time availability

Time availability is an important factor when recovering from an error in some HSC domains, as the inability to resolve an error within a restricted time window may result in harm to system entities. A domain has restricted time availability when system failure or human safety will be

compromised if the error is not resolved within a limited duration, which varies for different HSC domains (Inagaki, 2006), but can range from seconds to hours. An HSC domain could range from Low Time Availability (LTA) to High Time Availability (HTA). If there is uncertainty associated with the source of a system failure, error identification can be difficult in domains with LTA, with time pressure potentially having negative effects on human judgment and decision-making.

An example of an LTA system is a nuclear reactor plant. The Chernobyl nuclear plant accident had a restricted time window for error resolution, as can be seen by the devastation left behind after the operational errors went unresolved and a series of consequences led to the explosion of a reactor. UGV systems, on the other hand, typically have medium to high time availability. While it is important that the error is resolved efficiently, failure to resolve the error is unlikely to result in the loss of human life. Damage to system integrity, however, is likely.

2.5.4. Checklist Attribute Model (CAM)

The three HSC domain attributes were combined into a graphical display, termed the Checklist Attribute Model (CAM), with each attribute represented as the edge of a tetrahedron (Figure 4). In the model, each attribute edge can be broken down into three interval scales, ranging from low, to medium, to high (i.e. the edge of the tetrahedron corresponding to sensor reliability ranges from LSR at the bottom of the tetrahedron to HSR at the top of the tetrahedron).

This graphical representation categorizes HSC domains by their need for decision-support during error identification. Error identification in HSC domains with HDP, HSR, and HTA is not mentally demanding, as uncertainty is low, data is reliable, and time is available for selecting the source of system failure. Traditional checklists, which have been utilized in HSC domains for decades, can be useful in such domains, as their limited assistance during error identification and primary focus on error recovery steps is suitable for the low mental demands associated with such domains.

Error identification in HSC domains with LDP, LSR, and LTA is mentally demanding as uncertainty is high, data is unreliable, and time is restricted for identifying the source of system

failure. Traditional checklists are not appropriate for these domains, as decision-support during error identification is not provided to assist the human supervisor in error selection. SMU domains, which are intentional and highly complex, are examples of environments with such characteristics, and therefore, are not well served by traditional checklist systems. An alternative checklist, which assists the supervisor during error identification, is therefore required.

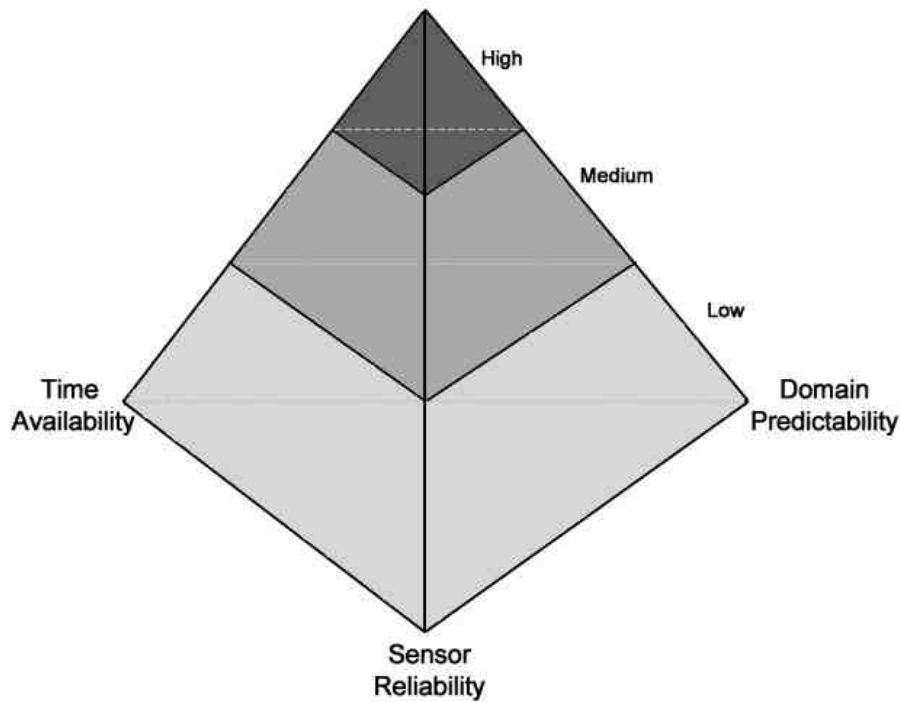


Figure 4: Human supervisory domain attributes for checklist design.

Error identification in HSC domains with Medium Domain Predictability (MDP), Medium Sensor Reliability (MSR), and Medium Time Availability (MTA) falls into a gray zone that is difficult to categorize. The appropriate checklist for these domains may be either a traditional checklist or the alternative checklist, depending on the overall level of complexity and uncertainty. As any uncertainty can negatively impact the error resolution process, assistance during error identification in such HSC domains, provided through the alternative checklist, may prove beneficial.

The CAM visualization represents all HSC domains. As traditional checklists are best suited to HSC domains with HDP, HSR, and HTA, the graphical model indicates that these checklists are

ideal for domains falling into the *high* section of the tetrahedron. As technology advances and environments increase in size and complexity, fewer HSC domains will have such characteristics. Thus, the proportion of HSC domains that are suited to error resolution with traditional checklists will become (and is already becoming) smaller, as depicted in Figure 4. Complex HSC domains, including SMU systems, will continue to increase in frequency, highlighting the need for an alternative error resolution tool. Considerations in the design of such a checklist tool are discussed in the following section.

2.6. Alternative checklist design considerations

The need for a checklist specifically designed for SMU domains has been identified. This checklist will incorporate a decision support tool to guide the human supervisor in accurately identifying the source of the error, enhancing efficiency and safety in SMU environments during error identification. There are many considerations that must be made in the design of such a checklist. These considerations are discussed in the following subsections.

2.6.1. Defining role of automation

Parasuraman and Riley (1997) define automation as “the execution by a machine agent (usually a computer) of a function that was previously carried out by a human” (p. 231). Automation can be incorporated into a system to various degrees, ranging from the human in complete control of the system to the automation in complete control of the system. While high levels of automation can result in a lower taskload for the human supervisor (Parasuraman, Sheridan, & Wickens, 2000), if the level of automation is too high, supervisors may experience a loss of situation awareness as a result of being out of the decision-making loop (Kaber, Endsley, & Onal, 2000). In addition, consistently relying on automation during decision-making can result in skill degradation (Parasuraman et al., 2000).

While automation can be applied to all aspects of HSC domains, there are two particularly relevant applications in SMU domains: 1) the unmanned vehicles that operate within the environment, and 2) the decision-support tool provided to the high level supervisor in the system. While determining the tasks and capabilities of the unmanned vehicles in an SMU domain is

outside the scope of this thesis, it is essential to consider the contributing roles that the supervisor and the automation will play during error resolution in order to achieve efficient and safe failure recovery. Once an appropriate automation level has been determined, this can be built into the alternative checklist system.

When automating an error resolution tool, four different categories related to the distinct stages of human information processing must be considered: information acquisition, information analysis, decision and action selection, and action implementation (Parasuraman et al., 2000). Each of these categories can be automated to varying degrees, with the appropriate automation range determined by considering the human performance consequences of the automation, as well as the reliability of the automation and the potential costs of incorrect decisions/actions. For error resolution support in the alternative checklist, the categories of information acquisition and decision and action selection were identified as the areas where automation could be of the greatest assistance to the human supervisor.

In SMU domains, information acquisition is predominantly performed by automation through the use of sensors (Parasuraman et al., 2000). During error resolution, it will be useful to the human supervisor to have this information organized by context, location, and other criteria, absolving the supervisor from having to dedicate cognitive processes to such tasks. By applying automation to information acquisition and organization, the focus of the human supervisor can be shifted from low-level sensory activities to high-level reasoning about the collected data, or decision selection.

While automation of the information acquisition process will likely be beneficial during error resolution, automating decision and action selection may not be appropriate due to the high level of complexity and uncertainty associated with SMU environments. As the sensor-collected data may be inaccurate, the deductive reasoning abilities of automation may be ill-suited to decision-making and error source identification. The inductive reasoning abilities of the human supervisor, on the other hand, may be better matched (Fitts, 1951), and therefore, the human should be responsible for error source identification during error resolution. This source

identification can be assisted by the automation, however, through the environmental data that it collects and organizes.

2.6.2. Human role in error identification

By including the human in decision and action selection, or error identification, the overall efficiency of the error resolution process then relies on the judgment of the human supervisor. Unfortunately, human judgment under uncertainty is not perfect, due to incomplete knowledge about the problem space and limited computational abilities. Further, time availability within an HSC domain can negatively impact judgment during error identification. Under time pressure, human decision-makers cannot always employ the decision-making strategy that determines the best alternative, as they may not have the time or attentional resources to consider and evaluate multiple hypotheses (Sarter & Schroeder, 2001). In such conditions, accuracy may be traded for time savings, and decision-making heuristics may need to be employed (Tversky & Kahneman, 1974).

Humans use a number of heuristics when making decisions in situations where there are time constraints, as well as incomplete knowledge and a bound on computing abilities. Three well-known decision-making heuristics that may impact the error identification process are:

- *Representativeness*: The probability of event B being of type A is evaluated by the degree to which B resembles A (Tversky & Kahneman, 1974). This, unfortunately, neglects the prior probability of the type A event occurring in the world. Without knowledge of past system performance, a human supervisor in an SMU domain may be apt to overestimate the likelihood of unlikely errors, due to their similarity with available system data.
- *Availability*: The probability of an event is based on the ability to retrieve similar events from memory (Tversky & Kahneman, 1974). Decision-makers are inclined to believe that an event is more likely to happen in the world if that event can be easily retrieved, although this ease of retrieval may not accurately reflect the true probability of occurrence. As error identification in SMU intentional domains is difficult, this decision-making bias could negatively impact error source identification, with the supervisor basing identification on availability instead of collecting data that confirms or refutes the believed error source.

- *Fast and frugal*: A subclass of decision-making heuristics that employ a minimum of time and computation to make judgments (Todd & Gigerenzer, 2000). Fast and frugal heuristics limit the search through options using stopping rules, one of the most basic of which is one-reason decision-making, where the selection between choices is based on a single metric. This is likely to be employed during error identification in situations with LTA, and would be particularly detrimental to error source selection in times of LSR. If the human supervisor were to use sensor data as the metric for error selection, inaccurate sensor data could result in incorrect error identification.

Humans employ decision-making heuristics as coping mechanisms, reducing complex tasks to more simple judgments (Tversky & Kahneman, 1974). These heuristics are therefore quite useful, as they allow for time and computational savings and often produce results that are good enough. Sometimes, however, the employment of heuristics can lead to severe reasoning errors. If these errors occur in non-critical environments, then the repercussions are not as far reaching. For example, if a student bases team selection for a project on how well candidates represent the “ideal” teammate, consequences from a bad team member selection will only be felt for the duration of the project, resulting in limited annoyance and frustration. If a human supervisor in a complex environment (e.g. nuclear power generation, aviation) employs bad judgment, however, negative consequences can be much farther reaching, and may include loss of system integrity and human life.

Error identification in SMU domains, as previously discussed, is difficult due to the high level of complexity and uncertainty in these environments. As the consequences of poor decision-making in critical HSC domains can be grave, it is crucial that a decision-support tool be provided to assist the human through the error identification process during error resolution. While the functions supported by such a tool were discussed in Section 2.6.1, with a focus on information acquisition and organization, how this tool will provide support has yet to be determined.

Decision-making heuristics often involve subjective assessments of probabilities, with inaccuracies sometimes resulting. Specifically looking at error identification in SMU domains, heuristics could result in the incorrect identification of an error source. In order to prevent (or

limit) incorrect error identifications, historical error occurrence data could be collected by a decision-support tool and presented to the supervisor, providing them with real probabilistic data. This error likelihood data would assist the supervisor in framing the current error using concrete measures, as opposed to subjective, heuristic-based judgments, which can lead to judgment errors. The error likelihood data could be collected by the automated decision-support tool and presented to the supervisor as part of an organized display. The likelihood data could then be aggregated with other available system cues to guide the supervisor in identifying the most likely source of error.

This compilation of error probabilities is similar to the Probability Risk Assessment (PRA) approach (Kirwan, 1992), or the more focused Human Reliability Assessment (HRA) approach (Gertman & Blackman, 1994), where the likelihood of potential system or human failures is quantified and used to predict how frequently each event will occur. These approaches are used to assess the safety of a system, often before the system has been constructed, and therefore, the error likelihood values are usually best estimates. For the envisioned decision-support tool in the alternative checklist, the error likelihoods could be derived from historical error occurrence data, providing a more accurate picture of the actual error landscape. A description of the algorithm that would be needed to collect, analyze and calculate these error likelihoods is outside the current scope of this research.

Even if accurate probabilistic data is collected and provided as part of the decision-support tool, humans are poor at interpreting probabilistic information (Tversky & Kahneman, 1974). If error likelihood data were to be included in the alternative checklist system as part of a decision-support tool during error identification, an intuitive display method of presenting that information to the human supervisor would need to be developed. The following subsection discusses the benefits of such a graphical representation, as well as potential options for representing the error likelihood data.

2.6.3. Visualization of error likelihoods

In order to prevent potential errors in reasoning resulting from the utilization of decision-making heuristics, a decision-aid tool incorporating probabilistic data could be developed as part of an

alternative checklist. This tool would lend computing power and supply much-needed knowledge to decision-makers. The need to develop an intuitive decision aid for the supervisor is paramount to ensure that the addition of probabilistic data to the decision-making task does not overwhelm the supervisor. As visualizations take advantage of the natural abilities of the human vision system, they continue to be the best method to communicate data to human operators (Schroeder, Martin, & Lorensen, 2006).

Visualizations are graphical representations of data or concepts that support decision-making (Ware, 2004). An appropriately designed visualization minimizes the cognitive complexity of a task (Guerlain, Jamieson, Bullemer, & Blair, 2002) and takes advantage of ecological perception, allowing users to directly perceive relationships within presented data (Gibson, 1979). In the case of probabilistic error data, it is important that users are able to compare and contrast the relative likelihoods of all possible errors quickly so that they can utilize this information in error source selection. In developing a graphic to represent the probabilistic error data, two key characteristics of the data need to be embodied: 1) hierarchical information, conveying the categorization of errors within a system (e.g., mechanical failures, automation errors, etc.) and, 2) the relative likelihoods of each system error.

Many graphical representations of likelihood data could convey both of these characteristics. Three options, namely tree structures, treemaps, and multi-level pie charts, are presented below. Each graphical representation is demonstrated using a generic error hierarchy that groups system errors into overall categories and subcategories, and includes associated error likelihood data, as summarized in Table 1.

An example error hierarchy can be demonstrated using the HSC domain of aviation and the overall error category *Landing*. This category can be divided into subcategories, including *Mechanical Failure*, *Fuel Shortage*, and *Human Error*. These subcategories would then include all related errors; an engine failure would fall into the Mechanical Failure subcategory, while forgetting to lower landing gear would fall into the Human Error subcategory. Each error during landing could have an error likelihood associated with it, based on historical event occurrence.

Table 1: Generic error hierarchy with associated likelihood values.

Overall category	Subcategories	Errors	Likelihoods
Error Category 1	Subcategory 1a	Error 1a.1	0.25
		Error 1a.2	0.125
		Error 1a.3	0.0625
	Subcategory 1b	Error 1b.1	0.25
		Error 1b.2	0.125
	Subcategory 1c	Error 1c.1	0.125
	Subcategory 1d	Error 1d.1	0.03125
		Error 1d.2	0.03125

Tree structures

The most basic graphical representation for depicting hierarchical data is a tree structure that starts with the overall error category at the top of the visualization. This root node is broken down into child nodes, or subcategories of errors, which are then broken down into actual error types or events. While the example hierarchical structure referenced in Table 1 and graphically depicted in Figure 5 only contains three levels, there is no limit on the number of levels of data that can be represented using a tree structure. Tree structures have been used in different tasks within many different fields, including operations research, computer science, business management, biology, and linguistics.

While probabilistic data could be textually listed at each node to indicate the likelihood of occurrence for each error or category of errors, this representation would not intuitively convey the relative likelihoods of each error and would require data integration and increased mental workload on the part of the human supervisor. As well, the probabilistic breakdown for all errors in the overall error category is difficult to discern from such a representation. Due to the required data integration when using this representation, the tree structure graphic is not a viable option for the decision-support tool as part of error identification in the alternative checklist.

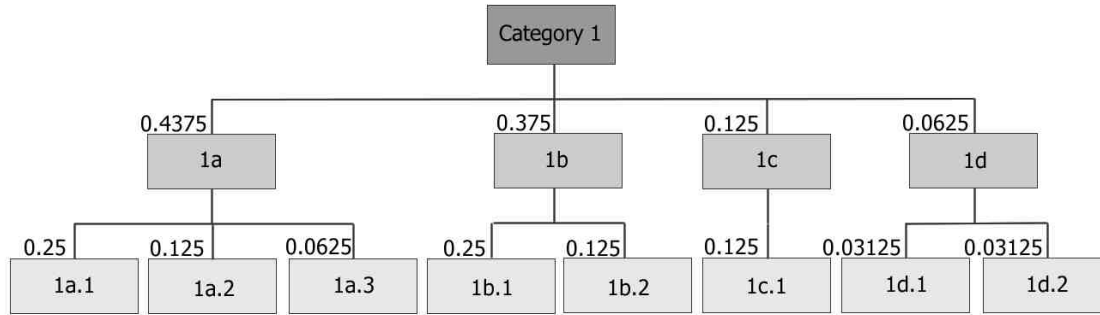


Figure 5: Decision tree visualization.

Treemaps

A treemap (Human-Computer Interaction Lab, 2003) is a graphical representation that depicts a hierarchical tree through the repeated subdivisions of a rectangular shape into nested rectangles. The outer rectangle represents the root of the tree, or the overall error category. This rectangle is divided into its children, or error subcategories, which can then be divided into error events or sources. Once again, while the example hierarchy from Table 1 has only three levels, this division of rectangles could continue, and multiple levels of data could be conveyed. As each further layer of data subdivides the space even further, however, there often needs to be a limit on the number of levels of data depicted. Treemaps have been commonly utilized in the fields of business and portfolio management for presenting both high-level overviews and low-level details of stock market activity (Cable, Ordonez, Chintalapani, & Plaisant, 2004; Smart Money, 2010).

In a treemap, each node or rectangle has an area proportional to a specific dimension of the data. For the generic error hierarchy data from Table 1, this dimension would be the associated error likelihood data, with the size of each rectangle representing the probability associated with each system error. By using object size to represent magnitude (in this case likelihood), the human observer can directly perceive that the larger rectangle is greater than the smaller rectangle, and therefore, immediately comprehend the likelihood data being conveyed (Guerlain et al., 2002).

The treemap resulting from the generic hierarchy data is shown in Figure 6. As both the hierarchical and likelihood aspects of the error data can be depicted using this graphical representation, it is a potential option for the error identification support tool.

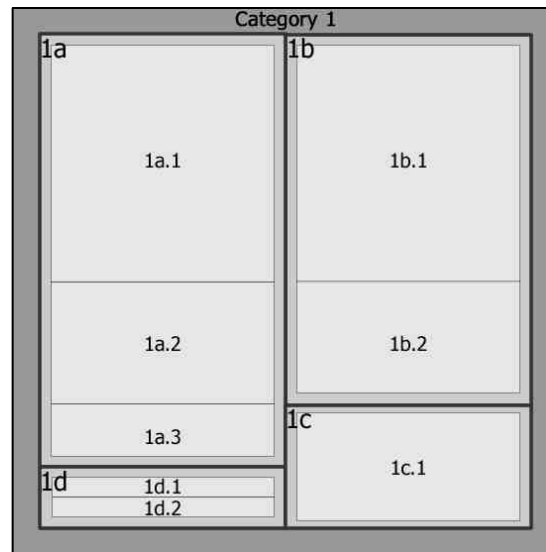


Figure 6: Treemap visualization.

Multi-level pie charts

In the multi-level pie chart depiction (Andrews & Heidegger, 1998; Stasko, Catrambone, Guzdial, & McDonald, 2000), the root of the hierarchical tree, or overall error category, is shown as the center of a pie chart. The next level of the tree, the error subcategories, is shown as the first layer of the pie chart, with the final layer of the pie chart representing the error events or sources. As with the other graphical representations, the pie chart graphic is capable of showing many hierarchical layers of data, but has been limited to three in the generic error hierarchy used for example purposes (Table 1). Like the tree structure, the pie chart graphic grows outwards with the addition of layers, unlike the treemap. The pie chart graphic is similar to the treemap representation, however, in that it inherently conveys a further dimension of the data: proportionality. In the case of the error data, this proportionality is likelihood. Each layer of the pie chart graphic can be seen as representing 100 percent probability, and therefore, the size of each error slice in the pie chart is directly proportional to its error likelihood. Once again, by representing magnitude (or likelihood) data through the size of the object, the human observer can directly perceive this information (Guerlain et al., 2002).

The pie chart graphic resulting from the generic hierarchy data is shown in Figure 7. As both the hierarchical and likelihood aspects of the error data can be depicted using this graphical representation, it is a potential option for the error identification support tool.

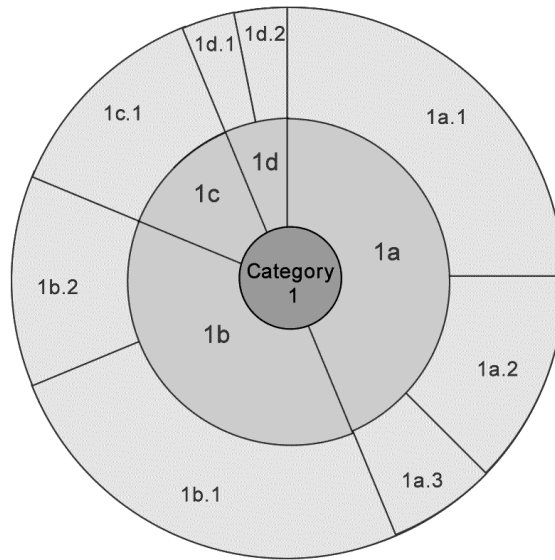


Figure 7: Multi-level pie chart visualization.

Selection of visualization

It is vital that probabilistic data is presented graphically in order for the data to be intuitively understood by humans. Of the three graphical representations considered, only treemaps and multi-level pie charts can represent both properties of the error likelihood data in a manner that does not cause undue mental workload on the human supervisor. However, the multi-level pie chart has a representational advantage over the treemap graphic: the ease with which it can be scaled for single or multiple layers of data. With the pie chart, additional layers of data are added by attaching an additional external ring to the pie chart (i.e., the graphic begins with a central circle and builds out from this central point). With the treemap visualization, additional layers of data are added by further compartmentalizing an overall rectangle (i.e., the graphic begins with an external rectangle and builds in from this outer point). Due to this inherent property of the treemap representation, individual data points can become small and difficult to comprehend with the addition of further data layers, as can already be seen in Figure 6.

Due to this identified disadvantage, treemaps will not be considered as a potential visualization method for the likelihood data in this effort. Therefore, the pie chart graphic is the best method of visualizing this data. The pie chart could be included as part of an alternative checklist that supports human supervisors in SMU domains during error resolution.

2.7. Summary

Three HSC domain attributes (domain predictability, sensor reliability, time availability) were identified that can be used to classify the needs of human supervisors during error identification in emergency events. Combining these into an attribute model for checklist design, it was identified that few HSC domains are suited to error resolution using traditional checklist tools. In order to properly support error resolution in domains not suited to traditional checklist use, including SMU environments, an alternative checklist must be developed. The design of this new checklist is discussed in the next chapter.

In this chapter, design considerations for this new checklist were discussed, including the roles of the human and of automation during error identification in the new checklist, with the determination that the automation should be responsible for organizing useful diagnostic data for error identification, while the human supervisor should be responsible for identification and selection of the error source. The information collected, organized, and presented by an automated decision-support tool was also considered, leading to the decision that probabilistic error likelihood data should be used to support supervisors during identification of an error source. Finally, methods of graphically depicting the error likelihood data were evaluated, with the multi-level pie chart representation found to be the most suitable depiction for use in the checklist, as it can convey both the hierarchical and proportional characteristics of the likelihood information.

Chapter 3. GUIDER Probabilistic Checklist

This chapter discusses the development of the new error resolution checklist tool. To overcome uncertainty and data inaccuracy, this new checklist includes probabilistic error data to guide the user in the error identification process, and a traditional serial presentation of steps to recover from the identified error. The new error resolution system was termed the GUIDER (Graphical User Interface for Directed Error Recovery) Probabilistic Checklist. The format of the two Probabilistic Checklist components, error identification and error recovery, is detailed.

Operations in current U.S. Army warehouse environments is described, as well as how these environments could change with the introduction of autonomous forklifts. A prototype design of the checklist for the autonomous forklift domain, termed the Error Identification and Recovery (EIR) display, is developed to use in testing of the new checklist tool. The GUIDER Probabilistic Checklist is then characterized using the Checklist Attribute Model (CAM) presented in Section 2.5. Two HSC domains are analyzed using CAM in order to predict the appropriate checklist, GUIDER or traditional, for the environment. First, a commercial aviation environment is analyzed, and second, the autonomous forklift domain is analyzed. These analyses are used as a basis to form hypotheses for the human performance experiment discussed in Chapter 4.

3.1. Error identification

When a failure occurs in an HSC system, it is critical that the source of the error is identified as quickly as possible so that error recovery can begin and the system can be transitioned back to an operational state. This not only ensures that efficiency in the system is maintained, but also that the probability of threats to human safety posed by system errors is reduced. To aid the human supervisor in SMU domains, it was deemed that a decision-support tool incorporating probabilistic error data should be included in the error identification portion of the new checklist tool, in order to support the selection of the error source. To intuitively convey the error likelihood data to the high-level supervisor, a graphical decision aid was deemed necessary. Of the possibilities evaluated in Section 2.6.3, the pie chart graphic was selected over the other graphical representations.

In the GUIDER Probabilistic Checklist, error identification will be performed using various information sources. As in traditional error resolution checklists, feedback concerning the current state of the system would be provided through system sensors. Supplementing this feedback will be the graphical decision aid conveying error likelihood data, as well as the supervisor's situation awareness of the present system state. This combination of data should contribute to more effective error identification under the uncertain conditions of the SMU domain, with the error likelihood data, and the pie chart representation of this data, overcoming many negative consequences related to human judgment under uncertainty, which was discussed in Section 2.6.2.

3.2. Error recovery

Once the supervisor has selected an error from the error identification portion of the GUIDER Probabilistic Checklist, error resolution will transition to error recovery. This portion of the checklist would consist of a traditional serial presentation of the recovery steps specific to the identified error. As the domain predictability, sensor reliability, and time availability attributes of HSC domains only impact the error identification portion of error resolution, the error recovery methodology used in traditional HSC checklists can be used in the new error resolution tool designed for SMU domains.

If the human supervisor identifies the error source correctly and the system failure is successfully resolved, the system would shift back to normal operations. If the error was incorrectly identified, however, and the recovery steps do not resolve the system failure, the error resolution process would shift back to the error identification portion of the GUIDER Probabilistic Checklist. The human supervisor would then be required to identify an alternative error source, continuing the process until the failure state has been resolved.

3.3. Application: Robotic forklift checklist

Military distribution warehouses, referred to in the U.S. Army as an SSA (Supply Support Activity) warehouse, store and maintain items (packed together into pallets) required for field operations. While manually driven forklifts are currently used to transport materials within the

SSA, a current project at MIT is proposing the introduction of autonomous forklifts into this domain. A background of the current SSA environment and an introduction to the autonomous forklift project are presented below. In addition, a prototype version of the GUIDER Probabilistic Checklist, termed the Error Identification and Recovery (EIR) display, is developed for the domain.

3.3.1. Current SSA operations

Pallets are transported between the different areas of the SSA environment, depicted in Figure 8, using manually operated forklifts. Pallets arrive in the SSA via truck bed in a reception area (Figure 9), are moved to a bulk lot (Figure 10), and get transported to the pickup area when customers arrive for requested items (Figure 11). In order to run the warehouse environment efficiently, multiple manually operated forklifts are used to move pallets. Currently, a high-level human supervisor in the system is responsible for monitoring human operators and maintaining efficiency. The supervisor does not have access to real-time system data, however, such as current locations of inventory, forklifts, or human operators, and performs monitoring duties by moving around the different areas of the SSA and observing operations.

Errors in current SSA systems predominantly revolve around the inventory stored in the warehouse, which is not tracked automatically using a Radio-Frequency Identification (RFID) system, but is instead manually tracked through SSA personnel. When inventory arrives at the SSA, human operators manually enter the inventory into a database system and then it is placed in bulk storage until the customer arrives for pick up. Every morning, a single employee also records all current items in the bulk area of the SSA and updates this information on a large summary board (Figure 12).

Using this inventory process, errors are common. Receiving trucks arriving at the wrong SSA are unloaded without verifying the contents of the shipment. When the items are processed, the mistake is realized and the items need to be reloaded onto the truck bed and shipped to the correct location. Items are also misplaced and cannot be located when the customer arrives. Both of these errors, while not compromising system safety, negatively impact the efficiency of the SSA, as well as customer opinion of the operation.

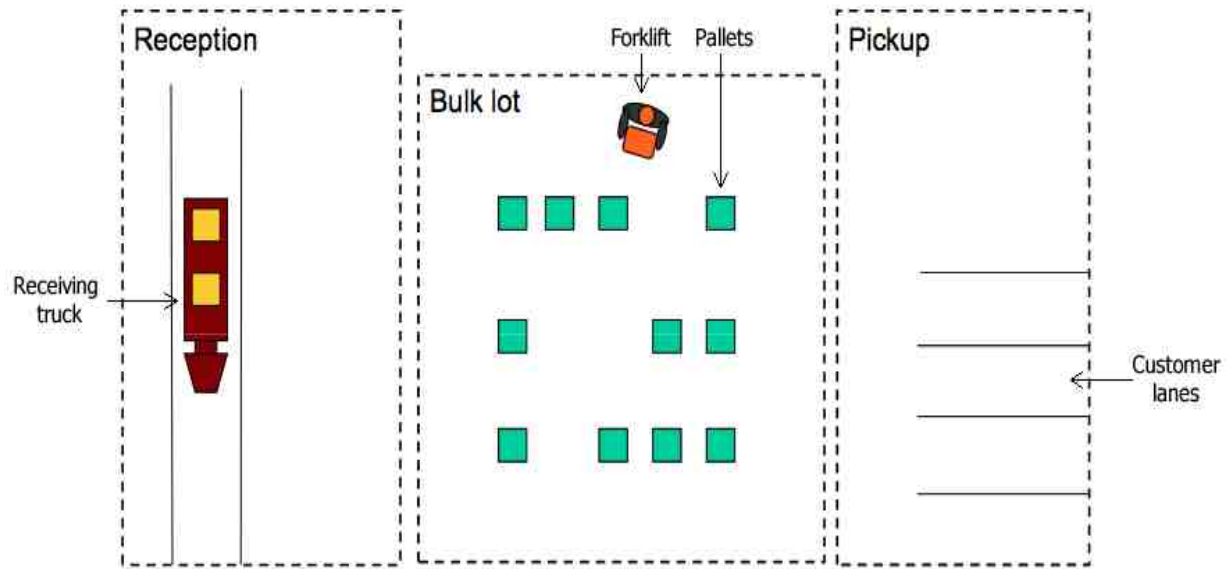


Figure 8: Notional layout of a U.S. Army SSA.



Figure 9: Truck delivering pallets of materials to receiving area of an SSA.



Figure 10: Bulk storage of materials in an SSA.



Figure 11: Customer vehicles waiting to receive requested materials from SSA.

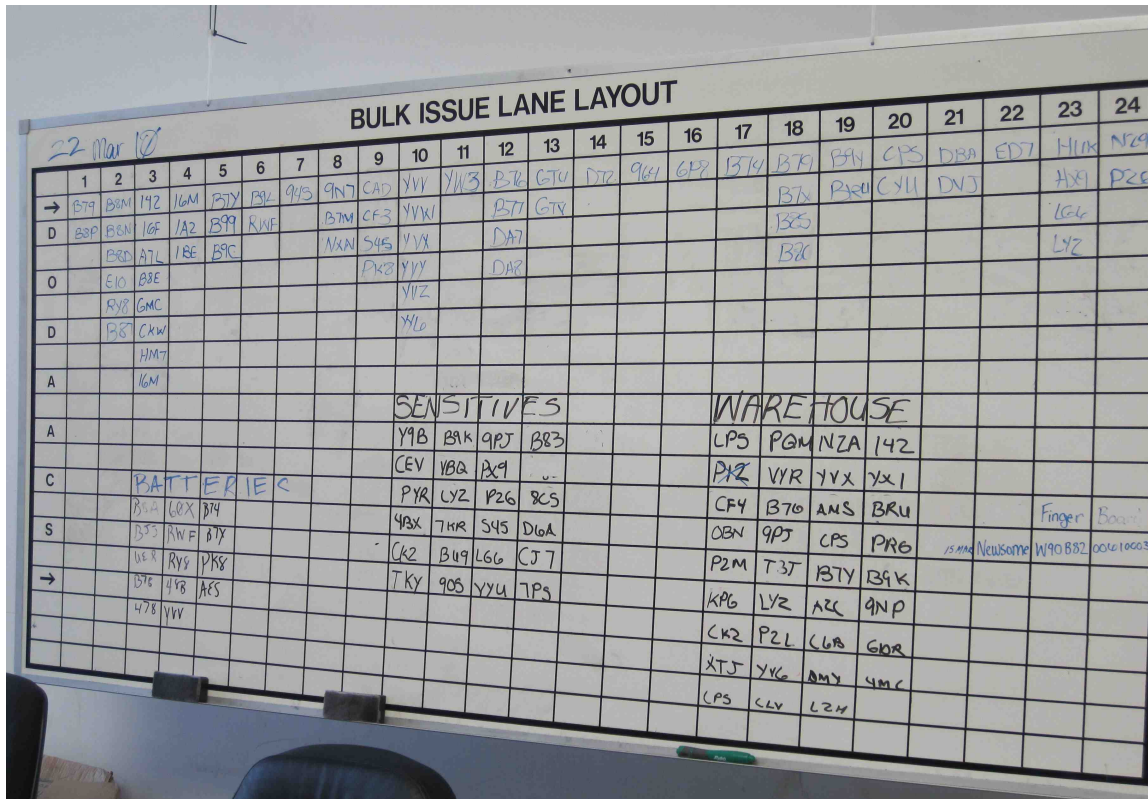


Figure 12: Tracking of items located in bulk area of SSA.

3.3.2. SSA domain with autonomous forklifts

Due to the location of some SSAs in active warzones, as well as a first step towards increased efficiency in the warehouse environment, a group of researchers at MIT are building an autonomous forklift that will replace the manually operated forklifts that are currently employed in the SSA domain (Chandler, 2009). This autonomous forklift project hopes to introduce unmanned robotic forklifts (RFs) into military equipment distribution warehouses and has the potential to streamline activities in these complex and sometimes hostile environments. It is envisioned that the RFs will move pallets within the SSA environment based on directions from ground-level human operators working alongside, and interacting with, the RFs. A human supervisor will also be included in the SSA environment to perform real-time monitoring of both RFs and human operators, to carry out planning and scheduling for the system entities, and to resolve system failures. The three distinct entities that would operate within this envisioned robotic SSA are shown in Figure 13. The following subsections discuss the requirements for each of these entities in the autonomous forklift environment.

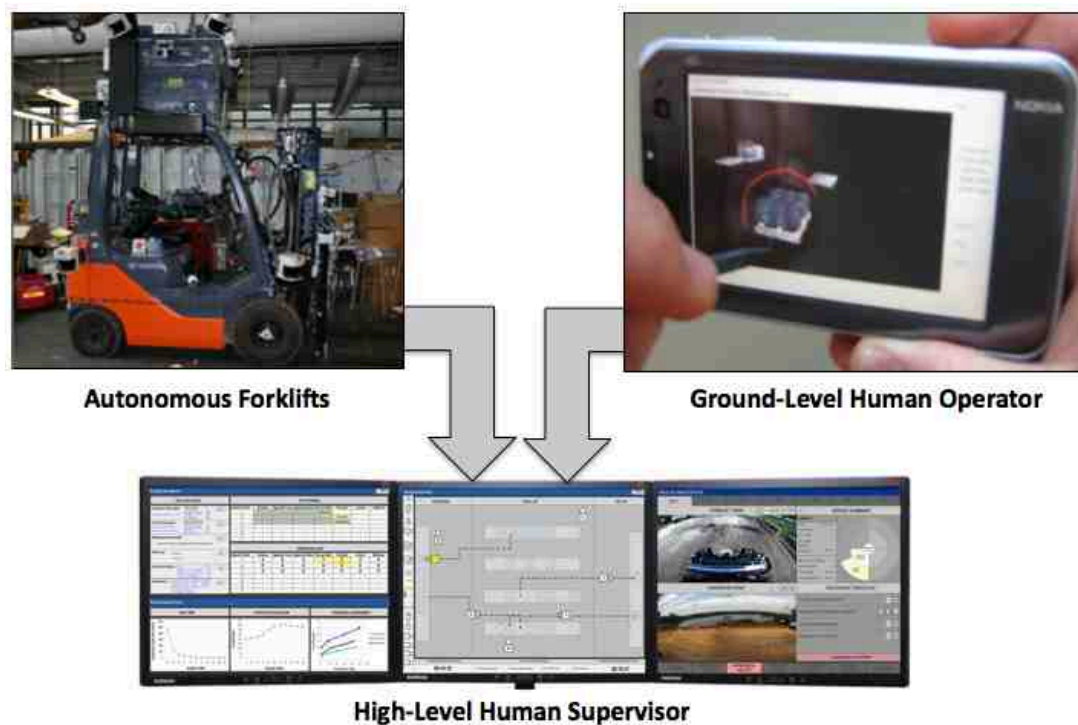


Figure 13: Entities operating within the forklift domain.

Robotic forklifts

In order to accomplish the pallet pickup and delivery task, the RFs would need to have a number of capabilities. These include (Walter, 2009):

- Detecting pallets with variable structure and load
- Inferring the geometry of a priori unknown trucks
- Avoiding obstacles within the SSA
- Retaining a common world model of the SSA
- Safely and smoothly interacting with human operators in the environment

To meet these needs, the prototype RF (altered Toyota 3-ton forklifts) has been outfitted with Sick[©] (<http://www.sick.com/>) LIDARs for sensing objects, Hokuyo[©] (<http://www.hokuyo-aut.jp/>) LIDARs for detecting pallets and trucks, four cameras (facing forward, backward, right, and left) for human operator and high-level supervisor situation awareness, and Light-Emitting Diodes (LED) signs to indicate the current task the RF is undertaking for the benefit of nearby human operators. Further modifications to the forklift are expected in order to achieve higher

reasoning levels by the RF, cooperative task allocation among multiple RFs, and localization within an environment without the use of GPS (Global Positioning System) (Walter, 2009).

Human operators

The role of human operators within the SSA would be to direct the RFs within the warehouse environment. This direction would be given using a handheld tablet PC, allowing the human operators to circle target pallets in the environment using the tablet touch screen and stylus. An example screenshot of the tablet PC can be seen in Figure 14, with a pallet circled. Once an RF has picked up a pallet, the operator would circle a drop-off location (e.g., in bulk storage if the item has just arrived at the SSA or in the issue area if a customer has arrived for an item). It is envisioned that operators would also give directions through voice commands and gesturing.

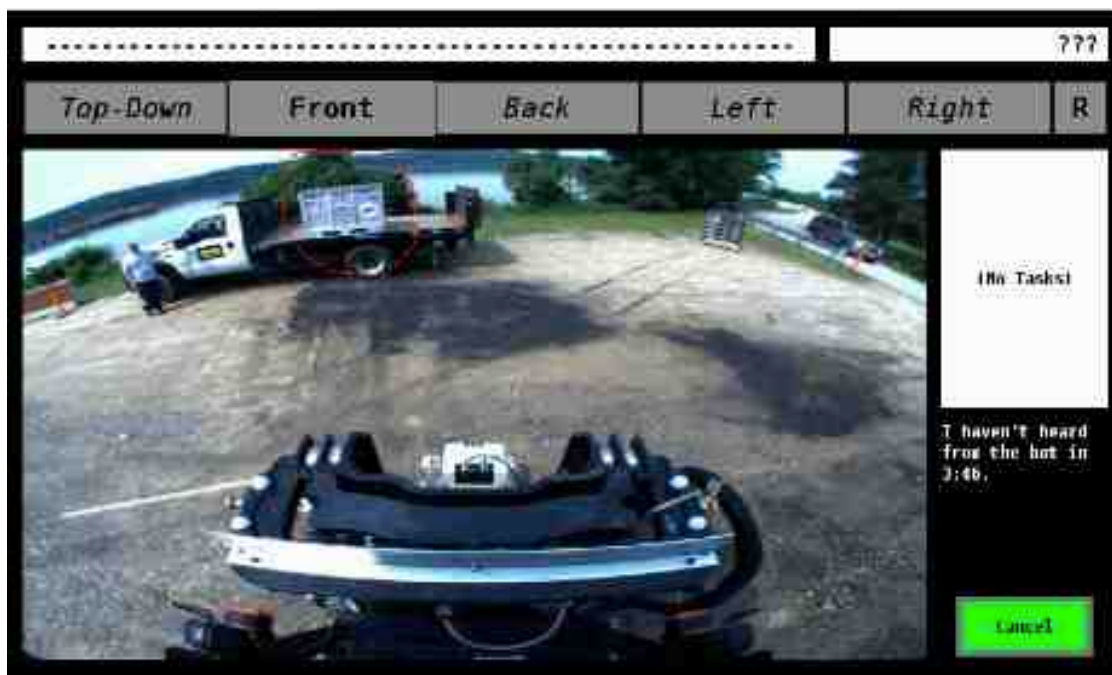


Figure 14: Tablet PC user interface for directing the RF in SSA.

Human supervisor

The human supervisor would be responsible for monitoring the RFs and the human operators working within the SSA warehouse environment. The supervisor would be responsible for high-level tracking of system entities, planning and scheduling tasks, and error resolution. The needs of the high-level supervisor for this final task would be met through the Error Identification and

Recovery (EIR) display, which would allow the supervisor to identify an error in the system, recover from the identified error, and transition the SSA back into an operational state. This EIR display would need to incorporate a checklist system to assist the supervisor through the error resolution task.

The envisioned SSA of the future, incorporating autonomous forklifts, has many of the characteristics of an SMU domain. There would be multiple unmanned forklifts, human operators, and potentially manned forklifts driven by the human operators in this shared environment. The unmanned forklifts and human operators would work within close proximity of one another. The result would be a high-level of complexity, due to the large number of potential interactions within the environment. Sensor reliability would also be a concern, as the forklifts would be operated in an environment where weather, blowing sand, and militant actions could negatively impact sensor accuracy. As this envisioned SSA is representative of an SMU domain, it was selected as a representative environment for demonstrating the GUIDER Probabilistic Checklist.

3.3.3. Sources of error

The first step of applying a checklist to a domain is to identify all potential errors within the environment. This step was undertaken for the representative domain, with the identification of all errors that could occur in robotic forklift field operations. An error model was developed that grouped system errors by functional step or point of occurrence in pallet pickup and delivery (Table 2). Seven distinct steps were identified in the functional process that involved a human operator summoning a robotic forklift to pick-up a pallet in the receiving area of the SSA (Figure 8) and delivering it to the bulk lot:

- Summon: Human operator calls an RF to the reception area to begin task.
- Approach truck: RF approaches truck that is carrying target pallet.
- Approach pallet: RF zones in on the location of the target pallet on the truck.
- Pick up pallet: RF inserts tines into pallet slots and lifts pallet off of the truck.
- Transport pallet: RF transports pallet to location designated by the human operator.
- Unload pallet: RF lowers pallet into drop-off location and removes tines from pallet slots.
- Withdraw: RF withdraws from drop-off location and waits for further instructions.

Table 2: Summary of potential errors in robotic forklift field operations.

Functional Step (FS)	Errors
Summon	-Voice command misunderstood by forklift -Gesture command misunderstood by forklift -Forklift does not receive command -Mechanical failure
Approach truck	-Operator designates path to wrong truck -Forklift takes wrong path -Forklift path blocked -Mechanical failure
Approach pallet	-Operator designates wrong pallet -Operator designates multiple pallets -Operator designates non-pallet -Operator designates pallet slots incorrectly -Forklift detects wrong pallet -Forklift detects non-pallet -Forklift cannot detect pallet slots -Forklift path blocked -Mechanical failure
Pick up pallet	-Forklift cannot find pallet slots -Pallet too heavy -Forklift picks up wrong pallet -Mechanical failure
Transport pallet	-Operator designates path to wrong location -Forklift takes wrong path -Obstacle in approach path -Forklift drops pallet/distribution unstable -Forklift transports wrong pallet -Mechanical failure
Unload pallet	-Obstacle in unloading location -Forklift cannot unload pallet -Forklift unloads pallet incorrectly -Forklift drops off wrong pallet -Mechanical failure
Withdraw	-Obstacle in withdraw path -Mechanical failure

There are a number of distinct errors, as can be seen in Table 2, that could occur during each of the seven functional steps. These errors could result from forklift failure, human error, or as a result of an interaction between the two entities (RFs and operators). To limit the scope, only those errors that could occur during the Approach Pallet functional step were considered for the

development of the EIR display for the forklift domain. There are nine errors that were identified at this functional step, with the errors grouped into four categories:

1) Pallet identification

- Operator designates wrong pallet: Using the tablet PC, the operator identifies an incorrect pallet for pickup. The result is the RF approaching the wrong pallet.
- Operator designates multiple pallets: Using the tablet PC, the operator identifies multiple pallets for pickup. The result is uncertainty as to which pallet the RF should pickup.
- Operator designates non-pallet: Using the tablet PC, the operator identifies an object for pickup that is not a pallet. The result is the RF approaching the wrong object.
- Forklift detects wrong pallet: While the operator identifies the correct pallet, the forklift approaches the wrong pallet.
- Forklift detects non-pallet: While the operator identifies the correct pallet, the forklift approaches an object for pickup that is not a pallet.

2) Slot identification

- Operator designates pallet slots incorrectly: Using the tablet PC, the operator incorrectly identifies the two slots of the pallet. As a result, it may not be possible for the RF to insert its tines into the pallet slots.
- Forklift cannot detect pallet slots: While the operator correctly identifies the two pallet slots, the forklift cannot detect the slots.

3) Obstacle

- Forklift path blocked: An object blocks the path of the forklift during pallet approach.

4) Mechanical failure

- Forklift mechanical failure could occur during any step of the pallet pickup and delivery process. For simplicity in design of the EIR display, motor failure and structural failure (e.g., wheel damage, forklift frame damage) were considered as the primary sources of mechanical failures.

3.3.4. The GUIDER representation

To demonstrate the GUIDER Probabilistic Checklist for error identification, the *approach pallet* functional step was chosen. The *approach pallet* step involves the forklift moving toward a pallet and inserting its tines into the pallet slots. A probabilistic error tree (Figure 15) was developed to attach likelihood data to each of the ten error sources that could occur during the *approach pallet* step, with each of the errors grouped into the four categories presented in Section 3.3.3. As the autonomous forklift domain is still in development, historical probability data is not available. Instead, preliminary estimates were chosen for the ten possible errors.

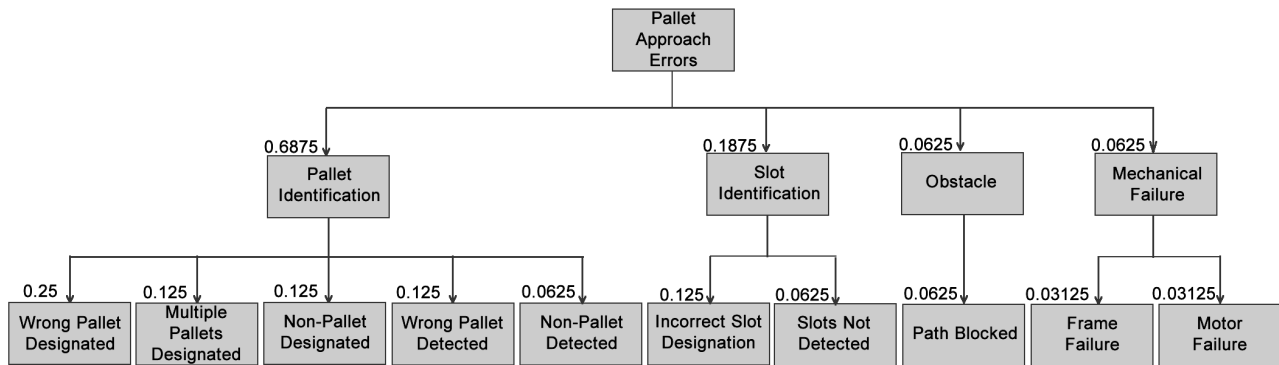


Figure 15: Probabilistic error tree summarizing the potential forklift errors.

This probabilistic decision tree was then transformed into a more intuitive graphical form, adapting the decision tree into the pie chart graphic presented in Section 2.6.3, simultaneously conveying to the supervisor both the hierarchical structure of the data and the relative likelihood of occurrence of each possible error source during an error event (Figure 16). The graphic, appearing when the forklift encounters an error, summarizes all possible error sources at the current functional step. The current functional step (*pallet approach*) is displayed in the center of the GUIDER graphic. The high-level error categories at this stage are shown in the next ring of the graphic, with their portion of the pie chart equivalent to the combined probabilistic occurrence of all errors in that category. Finally, each individual error is shown in the outermost ring, with individual likelihood data conveyed through the overall portion of the full pie chart.

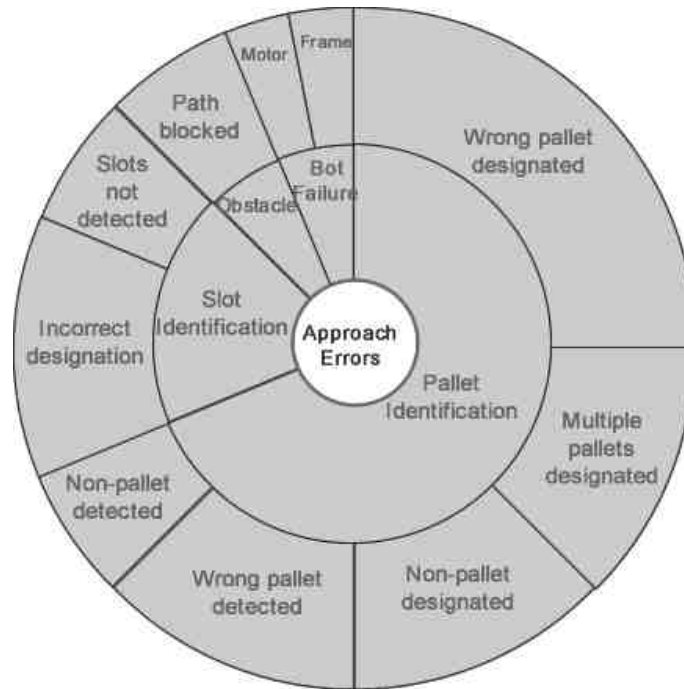


Figure 16: GUIDER representation of probabilistic error tree.

In order to utilize the GUIDER Probabilistic Checklist within the autonomous forklift domain, the error resolution tool needed to be incorporated into an EIR display. This display would be a primary component of a decision-support system for the high-level supervisor operating within the forklift environment. A prototype version of the EIR display is discussed in the next subsection.

3.3.5. Error Identification and Recovery (EIR) display

GUIDER was incorporated into an EIR display as part of a decision-support tool for supervisors in the robotic forklift domain. This EIR display allows the supervisor to track error occurrences for each unmanned forklift, identify the source of the error, and recover from the error.

The *identification* screen of the interface consists of the following six components (Figure 17):

1. Forklift tabs that allow the supervisor to select which forklift information he would like to view in the interface. Once a forklift has been selected, the other screen components would be specific to that selected forklift. In Figure 17, the supervisor is currently viewing information concerning Robotic Forklift 1 (RF1).

2. Forklift View, which provides the supervisor with ground level perspective of the RF domain through the real-time camera on the forklift. Each forklift is outfitted with four cameras: *front*, *right*, *back*, and *left*, with the Forklift View always showing the footage from the *front* camera. In Figure 17, the Forklift View shows a pallet, outlined in blue, which has been identified by the forklift.
3. Operator View, which shows the supervisor the current view of the handheld operator tablet used to control the forklift. This view is a combination of one of the four cameras on the forklift (the current view is selected by the human operator), as well as any annotations that have been made to the view using the tablet stylus. In Figure 17, the operator is currently monitoring the *left* camera on the forklift.
4. Identification portion of GUIDER Probabilistic Checklist. In Figure 17, *multiple pallets designated* is the error selected in the GUIDER pie chart graphic.
5. Additional diagnostic tests (e.g., Check Operator Tablet) to confirm or refute error sources. In Figure 17, the diagnostic tests related to *multiple pallets designated* are present.
6. Function bar indicating process the forklift is in during the pickup/delivery phase. In Figure 17, RF1 is currently undertaking the *approach pallet* functional step.

The *recovery* screen of the interface consists of the following five components, many identical to those of the *identification* screen (Figure 18):

1. Forklift tabs that allow the supervisor to select which forklift information he would like to view in the interface.
2. Forklift View, which gives the supervisor ground level perspective of the robotic forklift domain through the real-time camera on the forklift.
3. Operator View, which shows the supervisor the current view of the handheld tablet used to control the forklift.
4. Recovery portion of GUIDER Probabilistic Checklist. In Figure 18, the recovery checklist for *multiple pallets designated* is shown.
5. Function bar indicating process the forklift is in during the pickup/delivery phase.

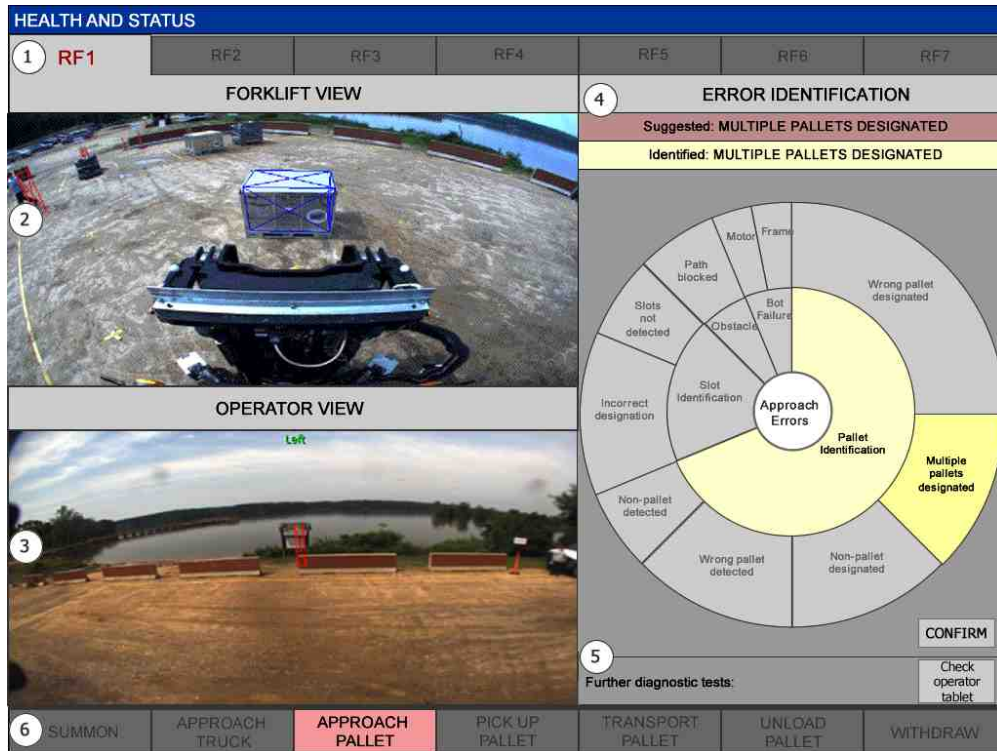


Figure 17: Identification screen in Error Identification and Recovery (EIR) display.

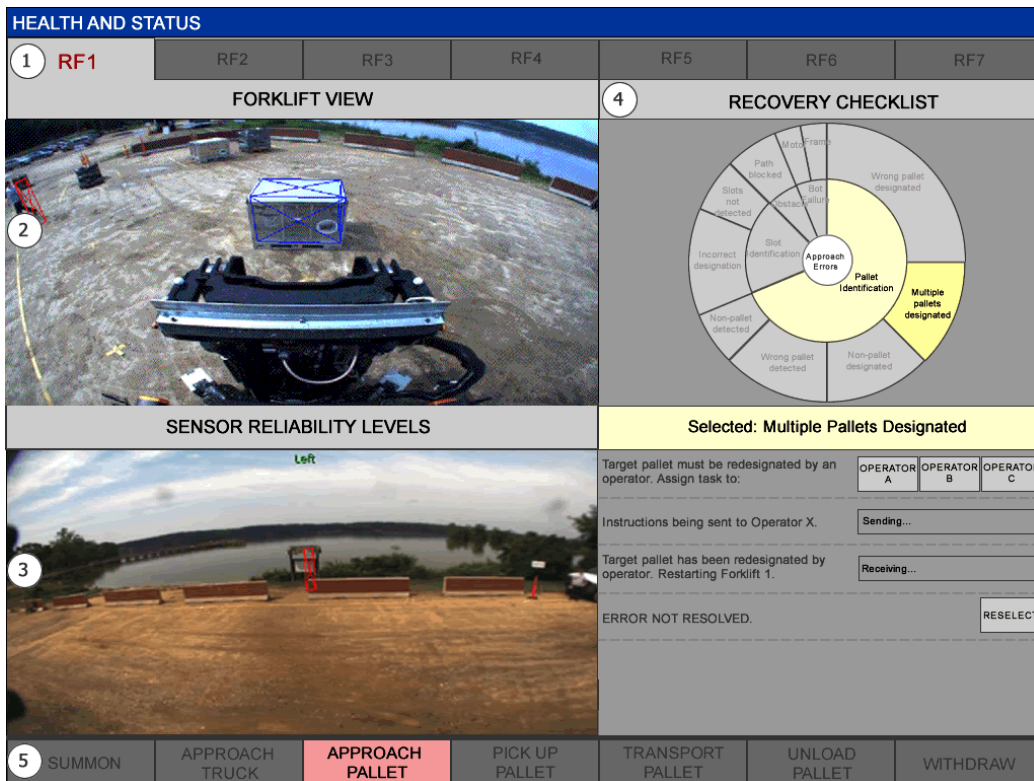


Figure 18: Recovery screen in Error Identification and Recovery (EIR) display.

When an error occurs in the robotic forklift domain, the error is indicated in the EIR display in four ways:

1. The tab representing the forklift experiencing the error changes to red (Component 1 in Figure 17).
2. The functional step that the forklift was undertaking when the error occurred is highlighted in red (Component 6 in Figure 17).
3. The source of error suggested by the system through state sensors is displayed (Component 4 in Figure 17).
4. The source of error suggested by the system is highlighted in yellow in the GUIDER graphic (Component 4 in Figure 17).

As the information provided by system sensors may not be accurate, the human supervisor has two options available: 1) to confirm the error source suggested by the system sensors, or 2) to refute this suggestion and select a different error source. If the supervisor agrees with the error suggested by the system, error resolution begins immediately by confirming the error source. The error resolution process would then shift to error recovery. It is possible, however, that information available to the supervisor does not coincide with the suggested error source. This inconsistency could be due to unpredictability in the SMU domain or inaccurate sensor feedback. For example, blown sand in the environment, which is likely to be an operational factor for the RF in current war zones (e.g., Iraq, Afghanistan), could impact LIDAR performance and make the detection of pallets within the forklift domain unreliable.

The human supervisor can use the information included in the EIR interface to identify the source of error. On the left of the display, real time forklift and operator views are included to supply the supervisor with additional information concerning the cause of the error (Components 2 and 3 in Figure 17). A second source of information is the GUIDER pie chart graphic, which summarizes all possible errors that could be the source of the failure, as well as their likelihood of occurrence (Component 4 in Figure 17). If the suggested error has a very small likelihood, the supervisor may be unwilling to accept the suggested error source as the true error source. Moreover, the supervisor may have information not available to the system, such as unusual environmental conditions (e.g., muddy areas). Finally, once an error has been selected using the

pie chart graphic, diagnostic tests, such as checking the pallet identification number (to verify if the identified pallet is the target pallet) and current position of the forklift in the domain (to compare this position to the location of the target pallet) are available to provide additional information to either confirm or refute the selected error source (Component 5 in Figure 17).

To illustrate an example where disagreement between supervisor and the RF sensor system could occur, consider the situation where the sensors suggest that multiple pallets have been designated for pickup, as opposed to a single pallet. The EIR interface for such a situation is shown in Figure 19.

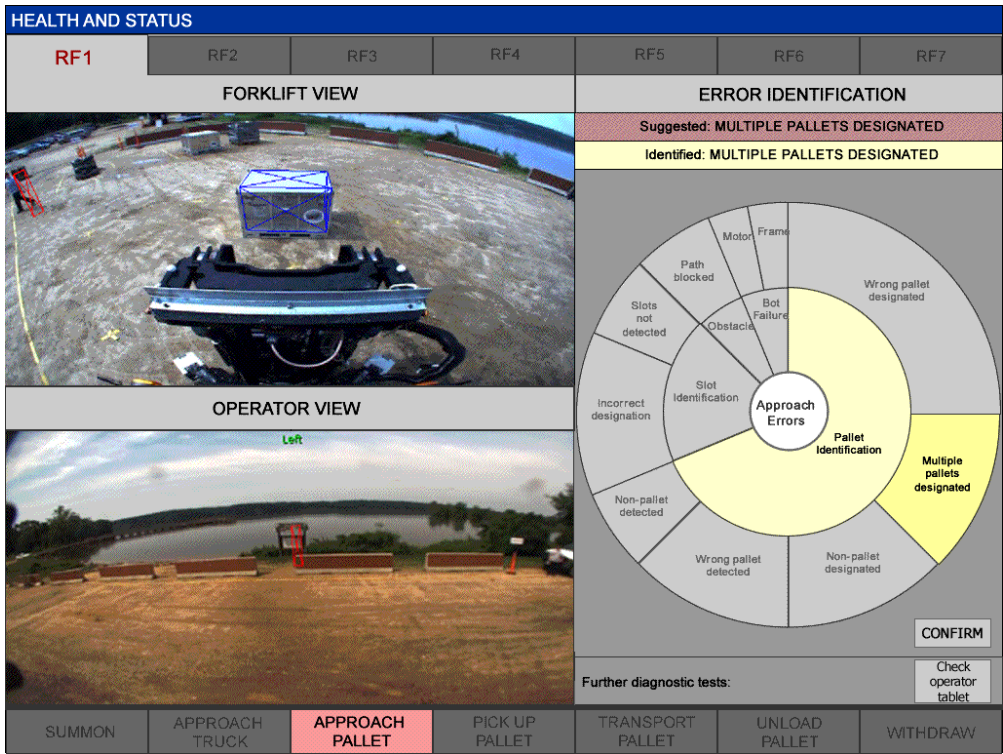


Figure 19: *Multiple pallets designated* suggested by system as error source.

The supervisor, after performing the further diagnostic test of viewing the operator tablet (Figure 20), might instead decide that the actual source of error is that the human operator has circled the target pallet incompletely. The supervisor would therefore disagree with the system suggestion and instead conclude that a *non-pallet designation* has occurred (Figure 21). By clicking CONFIRM, the error resolution process would proceed to error recovery for the *non-pallet designation* error. The associated *recovery* screen of the EIR interface is shown in Figure 22.

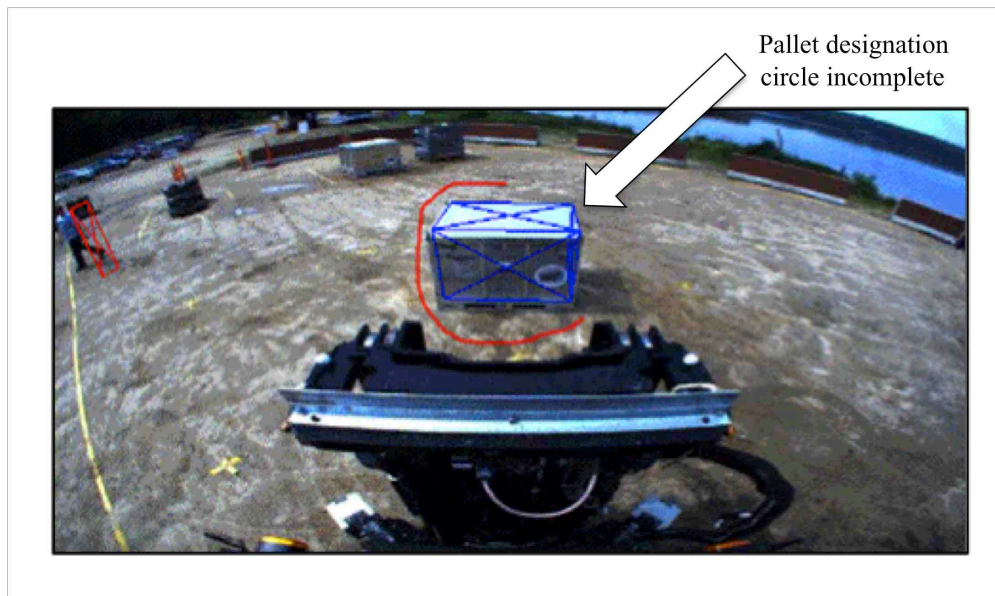


Figure 20: Diagnostic test inspecting Operator Tablet view.

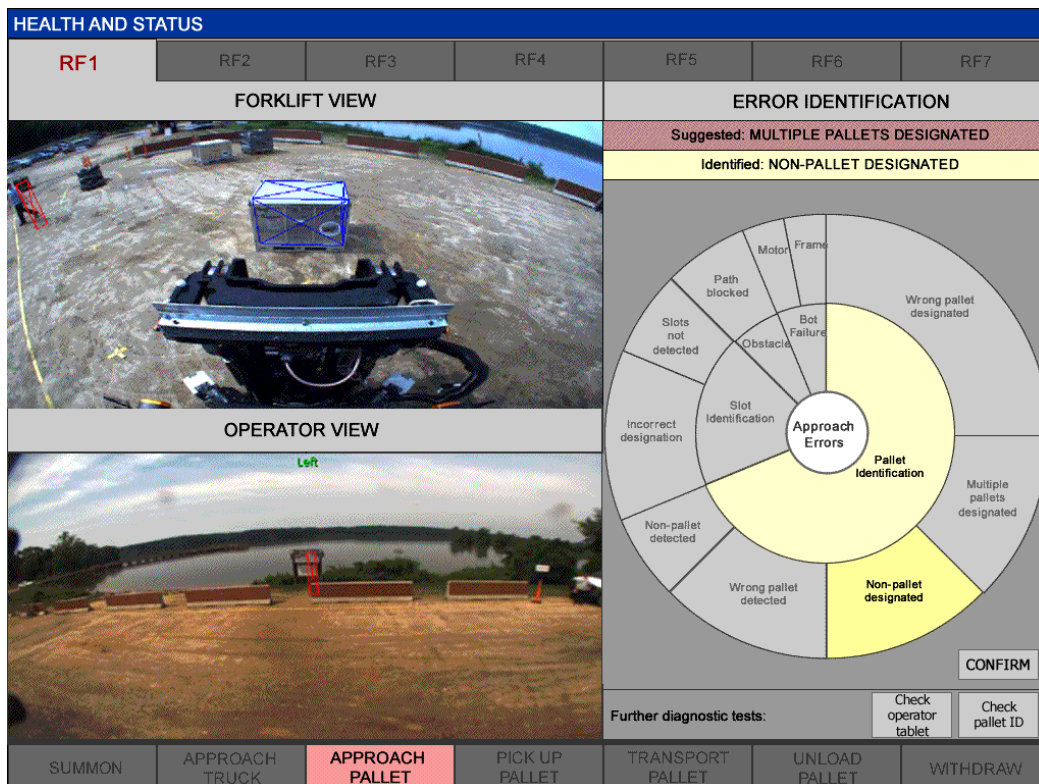


Figure 21: *Non-pallet designated* selected as error source.

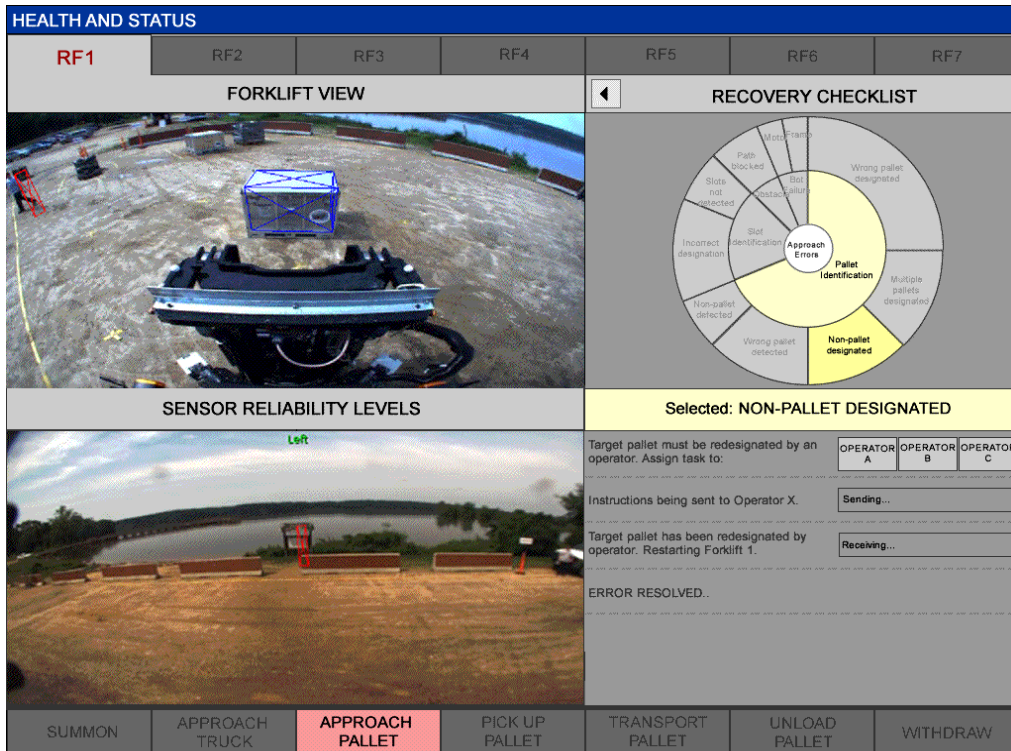


Figure 22: GUIER incorporated into Error Identification and Recovery (EIR) display.

The human supervisor would then complete the required recovery steps to return the system to an operational state. For the *non-pallet designated* error, these steps could include the supervisor sending instructions to the closest operator to re-designate the target pallet. If the recovery steps resolve the system failure, the supervisor correctly identified the error. If the failure persists, the supervisor incorrectly identified the error and reiteration of the error identification process would be required.

3.4. Revised Checklist Attribute Model (CAM)

The GUIDER Probabilistic Checklist was designed to meet the needs of error resolution within SMU domains. Therefore, the features of the checklist make it applicable for the following domain attributes:

- Low Domain Predictability (LDP): SMU domains are intentional environments (Section 2.1.2) that are highly complex and uncertain. Therefore, SMU domains could benefit from a combination of contextual system data and supervisor perspective to overcome uncertainty during error identification, which is the premise of the probabilistic checklist.

- **Low Sensor Reliability (LSR):** The knowledge of the supervisor can guide the error resolution process, instead of error resolution being solely dependent on the sensor information being received by the system, which is imprecise.
- **Low Time Availability (LTA):** If a domain has a high level of uncertainty and complexity, decision-support could help to streamline the error identification process, making the probabilistic checklist a viable option when there is low time available for resolving the failure.

Considering the characteristics of traditional checklists and the GUIDER Probabilistic Checklist, the CAM presented in Section 2.5.4 (Figure 4) can be revised to include checklist tool suggestions (Figure 23); the top of the tetrahedron would represent the end of the spectrum best suited to traditional checklist systems, and the base of the tetrahedron would represent the end of the spectrum best suited to the GUIDER Probabilistic Checklist. The middle of the tetrahedron (or the medium part of the attribute scale) would remain a gray zone that does not allow for the immediate classification of an HSC domain.

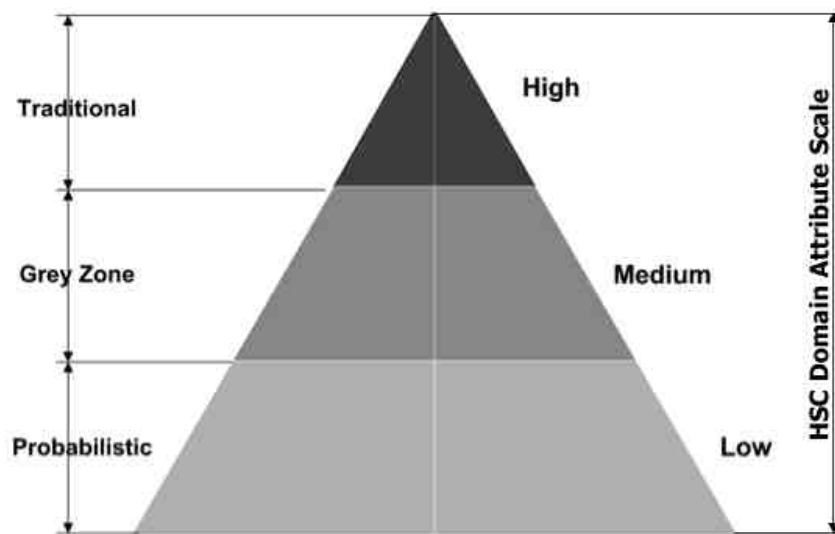


Figure 23: Revised Checklist Attribute Model (CAM).

As a result of this revision, classification of an HSC domain using CAM would allow a system designer to predict the best checklist tool for that domain. If a majority of the classifications were low (at least two of the three ratings on the domain attribute scale are low), the GUIDER

Probabilistic Checklist would be recommended. If the majority of the classifications were high, the Traditional Checklist would be recommended. If the majority of the classifications were medium, or each domain attribute was rated uniquely (i.e. one low rating, one medium rating, one high rating), the recommendation would be unclear.

This rating technique, while intuitive, is a subjective classification method that would need validation before utilization as part of an error resolution design process. The classification of existing HSC domains using CAM could assist in this validation process. The best checklist tool for error resolution within a domain could be hypothesized through CAM classification, with this hypothesis either accepted or rejected through experimental testing within a simulated version of the environment. If error resolution performance were best using the predicted checklist tool, there would be further evidence to validate CAM. On the other hand, if error resolution performance were best using the tool not predicted, there would be evidence against the validity of CAM.

Classifications for two example HSC domains are presented in the following section to demonstrate the use of CAM. These classifications are then used as the basis for the formation of hypotheses in a human performance experiment to test the GUIDER Probabilistic Checklist.

3.5. Example domain classifications

Two representative HSC domains were examined to demonstrate the use of CAM (Figure 23) for checklist selection. The first domain is a traditional system: commercial aviation. The second domain is the robotic forklift SMU environment.

3.5.1. Commercial aviation

The domain of commercial aviation was selected because it has many contrasting characteristics with SMU domains. Traditional checklists have been used in commercial aviation for decades and while accidents resulting from checklist errors still occur, these assistive devices continue to be the primary method used for error resolution. Given the three previously identified domain

attributes and applying them to commercial aviation, a traditional checklist is shown to be the appropriate checklist selection.

Domain predictability

The commercial aviation domain is an example of a generally causal-system in that the system has clear constraints, the behavior of the system is well understood, and there is a clear feedback loop between the state of the system and the pilot supervising the flight. As a result of these clear boundaries, the aircraft almost always acts and responds in an expected manner. It should be noted, however, that there might be unexpected environmental factors that may act on the system, including severe weather and other objects in flight (birds, aircraft). These may result in unpredictable system behavior, and therefore impact human performance in this environment. As commercial aircrafts are rarely impacted by unpredictable environmental factors, it is concluded that this domain has high domain predictability (HDP).

Sensor reliability

In the domain of commercial aviation, there is limited uncertainty associated with the sensor data provided to the pilot. Due to the clear boundaries of the system and high reliability of sensors, it is possible to provide feedback data to the pilot for almost all system components for most situations. These sensors have been refined over the last half-century, resulting in high-functioning and reliable equipment systems (Dismukes, Berman, & Loukopoulos, 2007). As a direct result, aircraft accidents have decreased substantially. We can conclude that the sensor reliability in commercial aircraft systems is very reliable. Using the domain attribute scale, it is evaluated as having high sensor reliability (HSR).

Time availability

The low time availability of the commercial aviation domain is immediately apparent, as an aircraft accident can result in not only the loss of the system, which is quite costly, but also the passengers. If the proper steps are not followed during aircraft operation, or error events are not resolved in a timely manner, serious consequences can result. It should be noted, however, that while it is vital that errors in commercial aviation be resolved quickly, the time available for error resolution could vary greatly between error events. It can be concluded that error resolution

in the commercial aviation domain is almost always time critical, as system and operator safety could be compromised. Using the determinant scale in Figure 23, it is evaluated as having low time availability (LTA).

Overall evaluation

Receiving two high ratings for domain predictability and sensor reliability, using the checklist classification model we can form the hypothesis that a traditional checklist system will better aid error recovery in the HSC domain of commercial aviation.

3.5.2. Forklift domain

A high-level supervisor in the robotic forklift domain will be responsible for monitoring human and autonomous entities in the system, maximizing workflow efficiencies, and monitoring for and recovering from system errors. An examination of the domain using the identified attributes can help to determine which checklist style is most appropriate for this system.

Domain predictability

The robotic forklift domain falls into the general classification of an intentional command and control environment, with unclear constraints. The domain has dynamic entities, both in the form of human operators and unmanned vehicles, and is highly influenced by unforeseen environmental factors present in war environments. Moreover, human decision-makers, whose behavior is guided by decision-making heuristics, impact these domains. With such unclear system boundaries, the overall system response would be quite variable for unexpected events. Using the domain attribute scale, it can be concluded that the robotic forklift domain has low domain predictability (LDP).

Sensor reliability

In the robotic forklift domain, many sensors will be required to guide the forklift during the pallet pickup and delivery process, as well as provide feedback to the supervisor during this task. For example, during pallet approach, sensors will be needed that locate the pallet, identify the pallet, locate the slots of the pallet, and guide the forklift tines into those pallet slots. In addition,

other sensors would be required to operate at all times, such as obstacle detectors and sensors that monitor for mechanical and system failures.

Due to the system having unclear boundaries and a wide range of environmental factors acting on it, it is impossible to incorporate sensors to provide feedback information for all system entities for all situations. The operating conditions of the military warehouse environment also complicate information feedback loops. In war zones, weather, uneven ground, and militant actions may result in sensors becoming damaged or inoperable, leading to inaccurate data being communicated to the supervisor.

Due to the difficult operating conditions and unclear boundaries of the robotic forklift domain, sensor reliability may be compromised. Using the domain attribute scale, this domain is evaluated as having low sensor reliability (LSR).

Time availability

In the robotic forklift domain, it is important that orders are fulfilled in a timely manner. Delivery trucks must be unloaded efficiently, the unloaded pallets taken from the reception area and moved into bulk lot storage promptly, and with quick delivery to customers. While it is important for time inefficiencies to be avoided, errors that cause system delay seldom impact overall system health or operator safety if they are not resolved immediately. Instead, these errors will likely lead to longer wait times and negative customer feedback. In isolated situations, however, the efficient movement of equipment may be vital to military operations. In such cases, quickness of processing becomes more critical.

It can be concluded that error resolution in the robotic forklift domain is usually not time critical on the order of seconds, although time will occasionally be a factor. Using the determinant scale, it is evaluated as having medium time availability (MTA).

Overall evaluation

Receiving two low ratings for system predictability and sensor reliability and one medium rating for time availability, using the checklist classification model we can form the hypothesis that the

GUIDER Probabilistic Checklist system will more effectively assist error identification and recovery in this SMU domain.

3.6. Summary

The GUIDER Probabilistic Checklist was designed for highly complex and uncertain SMU domains, supplying the human supervisor with error likelihood data to aid them during the error identification process. This likelihood data was included in the checklist to provide contextual information about past system performance, which when combined with the present situation awareness of the supervisor, could combine to improve error identification. The pie chart graphic was selected as the graphical representation of the error likelihood information, with the graphic intuitively conveying the hierarchical and proportional characteristics of the likelihood data.

The GUIDER Probabilistic Checklist was applied to the RF SSA environment as a representative SMU domain. The GUIDER Checklist, including graphical likelihood representation, was incorporated into a prototype EIR display. For demonstration purposes, only a small subset of the potential errors in the RF domain was considered during the development of this EIR interface. The resulting EIR display included the GUIDER Probabilistic Checklist (with pie chart graphic), video footage from the on-forklift cameras, and a serial presentation of recovery steps once an error source had been identified and confirmed.

Based on the revised Checklist Attribute Model (CAM), it was determined that the newly designed GUIDER Probabilistic Checklist could better aid error resolution in an SMU domain when compared to a traditional checklist. This hypothesis will be tested through a human performance experiment that is described in the next chapter. In order to test the two error recovery systems, error scenarios will be simulated using the robotic forklift domain and an EIR interface developed from the prototype screenshots included in Section 3.3.5.

Chapter 4. Experimental Evaluation

A human performance experiment was conducted to evaluate the effectiveness of the GUIDER Probabilistic Checklist compared to a more traditional checklist in an SMU environment. The forklift domain was utilized as a representative SMU domain, with simulated forklift error scenarios created for the EIR display discussed in Section 3.3.5. This chapter describes the setup of the experiment, including apparatus used in the simulations, hypotheses, and the experimental procedure.

4.1. EIR simulation

In order to evaluate research hypotheses through a human performance experiment, a specific EIR interface was developed and potential error scenarios were simulated. The two components of the EIR simulation, the simulated error scenarios and the EIR display are described in the following subsections.

4.1.1. Error scenarios

In order to assess the hypotheses formulated using the Checklist Attribute Model (CAM), which was presented in Section 3.4, one of the three attributes identified in the model had to be incorporated into the experiment as an independent variable. By adjusting this variable in different error scenarios, the predictions of CAM could be verified or refuted. The attribute that was selected as an independent variable was sensor reliability. This attribute was chosen because it could be easily varied in the simulated testing conditions utilized for the experiment.

It was decided that the error scenarios would only involve a single robotic forklift (identified as RF1), and all failures would occur at the *approach pallet* functional step. The limiting of the failure to a single process step reduced the potential error sources in the simulated forklift environment to ten. Each of the ten errors was grouped into one of six system sensor groups. The potential errors for the simulated scenarios, as well as the associated sensor groups, are summarized in Table 3.

Table 3: Sensor groups and related pallet approach errors.

Sensor Group	Potential Error
Mechanical	Frame failure
	Motor failure
Obstacle detection	Path blocked
Pallet designation	Multiple pallets designated
	Non-pallet designated
	Wrong pallet designated
Pallet detection	Non-pallet detected
	Wrong pallet detected
Slot designation	Incorrect slot designation
Slot detection	Slots not detected

Three error scenarios were created to represent each of the three levels of sensor reliability that were included in CAM: low reliability, medium reliability, and high reliability. The related reliability level for each sensor group for the three simulated scenarios is shown in Table 4.

Table 4: Sensor reliability levels for simulated error scenarios.

Sensor Group	Scenario 1 (Low)	Scenario 2 (Medium)	Scenario 3 (High)
Mechanical	Low	High	High
Obstacle detection	Low	Low	High
Pallet designation	Medium	Medium	Medium
Pallet detection	Low	Medium	High
Slot designation	Low	Medium	Medium
Slot detection	Low	Low	High

4.1.2. EIR display

The design of the simulated EIR display is based on the prototype EIR interface that was presented in Section 3.3.5. Two unique simulated displays were used representing: 1) the GUIDER Probabilistic Checklist, and 2) a Traditional Checklist, which did not include a decision-support tool to assist the participant through error identification, and instead listed all potential error sources alphabetically. The adoption of two EIR displays allowed for the testing

and comparison of the two distinct checklist tools. Both simulated interfaces consisted of six components:

- Component 1: Forklift tabs that allow the supervisor to select which forklift information they would like to view in the interface.
- Component 2: Forklift View, which gives the supervisor ground level perspective of the robotic forklift domain through the real-time camera on the forklift.
- Component 3: Sensor reliability levels (instead of the Operator View that was included in the original EIR interface).
- Component 4: Error resolution checklist tool (GUIDER or Traditional).
- Component 5: Additional diagnostic tests to confirm or refute error sources.
- Component 6: Process bar indicating process the forklift is in during pickup/delivery.

The EIR interface designed for the GUIDER Probabilistic Checklist is shown in Figure 24 and the EIR interface designed for the Traditional Checklist is shown in Figure 25.

For each simulated error scenario, the EIR display was customized in a number of ways. The Forklift View (Component 2) was unique to each of the three error scenarios, with the camera image aiding the participant in developing a fuller perspective of the current system state. Diagnostic tests were also available for each scenario, providing additional information to confirm or refute an error source. As part of the sensor feedback to the system, an error source was suggested to the participant at the top of the error resolution checklist tool (Component 4), which could be correct or incorrect (Table 5). The accuracy of this suggestion, or the level of trust that should be placed in this information, was indicated to the participant through the reliability level of the sensor group responsible for detection of the suggested error source (Table 3). For example, the suggestion of motor failure in Scenario 1 is likely to be inaccurate as the mechanical sensor group responsible for the detection of that error source has a low reliability level. As can be seen in Table 5, the only scenario for which the suggested error source and the true error source matched was Scenario 3, which had predominantly high reliability sensors.

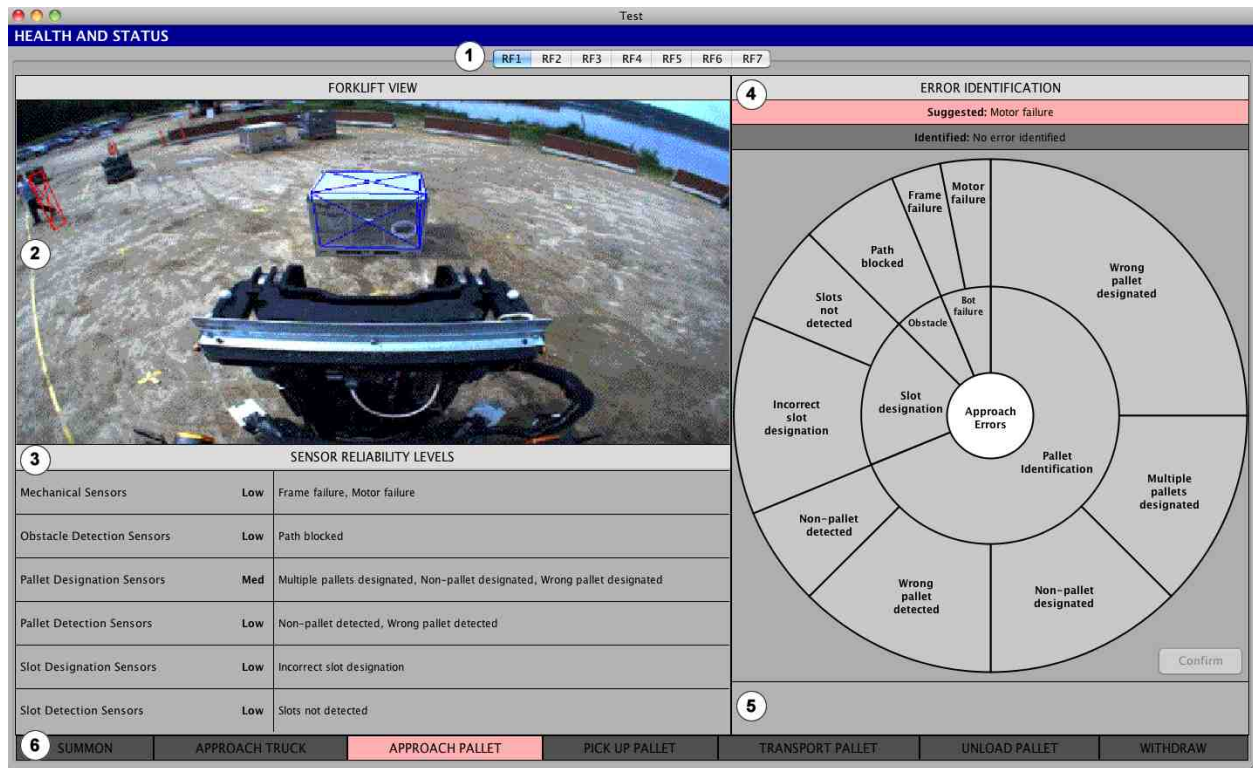


Figure 24: Identification screen, GUIDER Checklist.

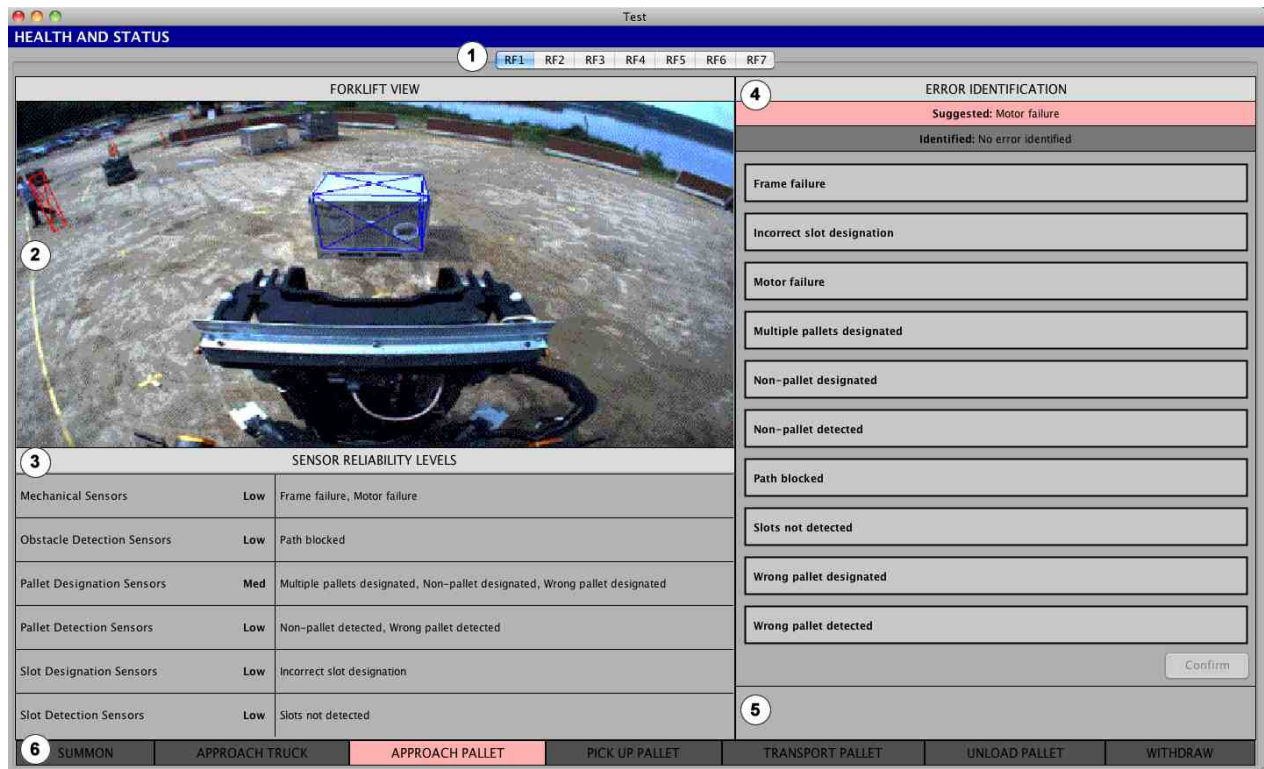


Figure 25: Identification screen, Traditional Checklist.

Table 5: Suggested and true error source for each error scenario.

	Scenario 1 (Low)	Scenario 2 (Medium)	Scenario 3 (High)
Suggested Error	Motor failure	Incorrect slot designation	Wrong pallet detected
True Error	Slots not detected	Wrong pallet designated	Wrong pallet detected

The participants were required to aggregate the provided diagnostic information presented in the error identification portion of the simulated EIR interface to identify what they believed to be the true source of the system failure. This information included the Forklift View, the sensor reliability levels, the suggested error based on the sensor feedback, error likelihood information (GUIDER Checklist only), and the available diagnostic tests. The diagnostic tests provided additional information to confirm or refute each of the potential error sources. For example, one diagnostic test available within the EIR display allowed a participant to check the pallet identification number of the detected pallet against the goal pallet. A second diagnostic test option allowed a participant to view the operator tablet to see the pallet that was designated by the operator, as well as the method of designation. The diagnostic test information provided was identical for both checklist types (GUIDER, Traditional).

Once participants identified an error, they needed to confirm the error selection by clicking on the CONFIRM button located at the bottom of the GUIDER checklist. Up until the confirmation of an error, participants could continue to collect diagnostic data for any of the potential error sources. After clicking CONFIRM, they transitioned into error recovery. In the error recovery portion of the simulated EIR display, participants were presented with a recovery task that needed to be completed in order to resolve the identified error. For example, if *non-pallet detected* was the confirmed error source, the first step in the recovery checklist would state: “*Forklift 1 path planning software needs to be validated. Assign task to.*” While the error recovery task was unique to the ten potential error sources, each task needed to be assigned to one of three operators within the simulated environment: Operator A, Operator B, or Operator C. A physical printout of the current location of the forklift and operators was provided to the participants, so that they could determine which operator to assign the task to in the environment, based on proximity to the failed forklift. This environment map was unique for each of the simulated error scenarios. The map for Scenario 1 is shown in Figure 26.

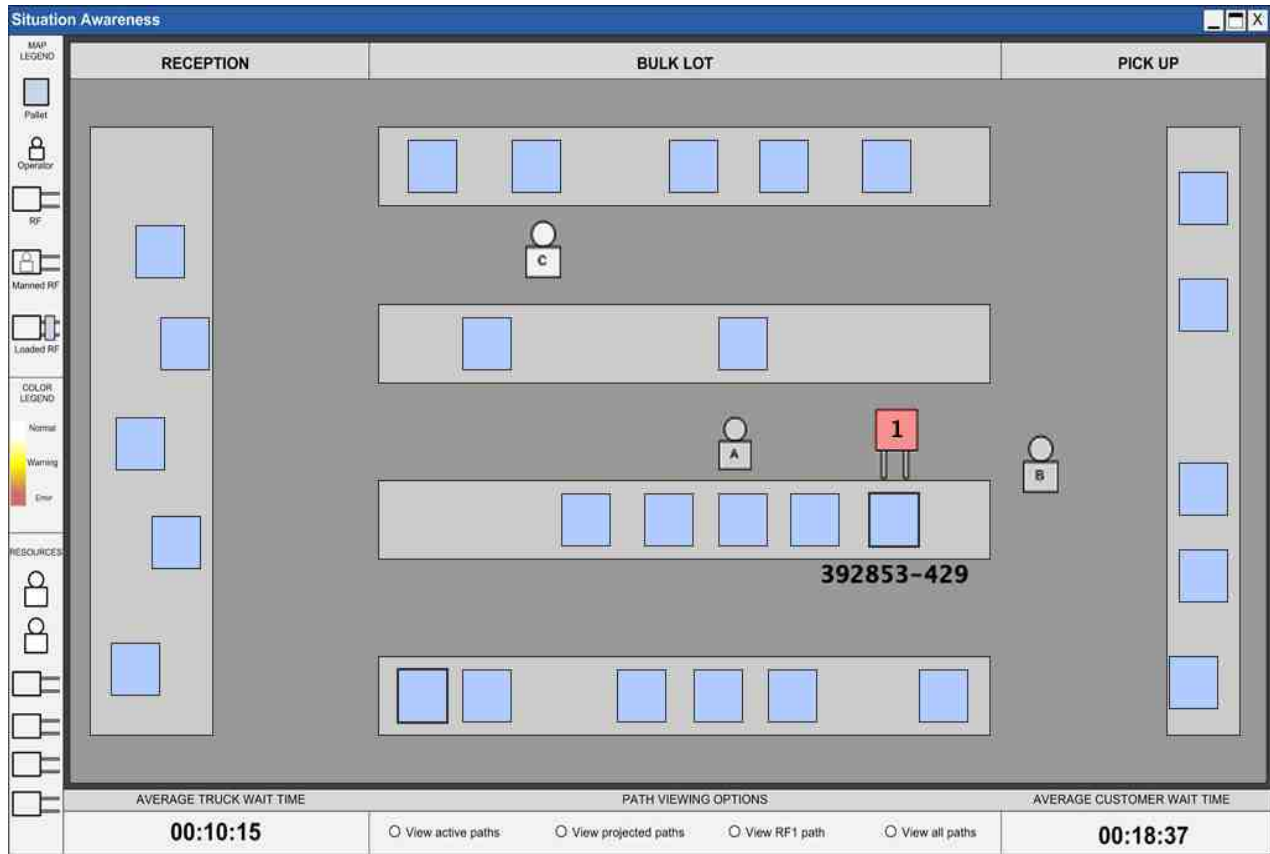


Figure 26: Map of forklift environment for Scenario 1.

Once a participant assigned the task to an operator, the assignment would be processed and the simulation would assess whether the participant had correctly or incorrectly identified the error source. If participants had selected the incorrect error source, they would be alerted, *“Incorrect error identified. Please try again”*. This situation is shown in Figure 27 for the GUIDER Probabilistic Checklist and in Figure 28 for the Traditional Checklist, with participants transitioning back into the error identification portion of the checklist by clicking the *Reselect* button. If the correct error was selected, participants were alerted, *“Error identified”* and then continued to the next simulated error scenario by clicking the *Continue to next scenario* button. This situation is shown in Figure 29 for the GUIDER Probabilistic Checklist and in Figure 30 for the Traditional Checklist.

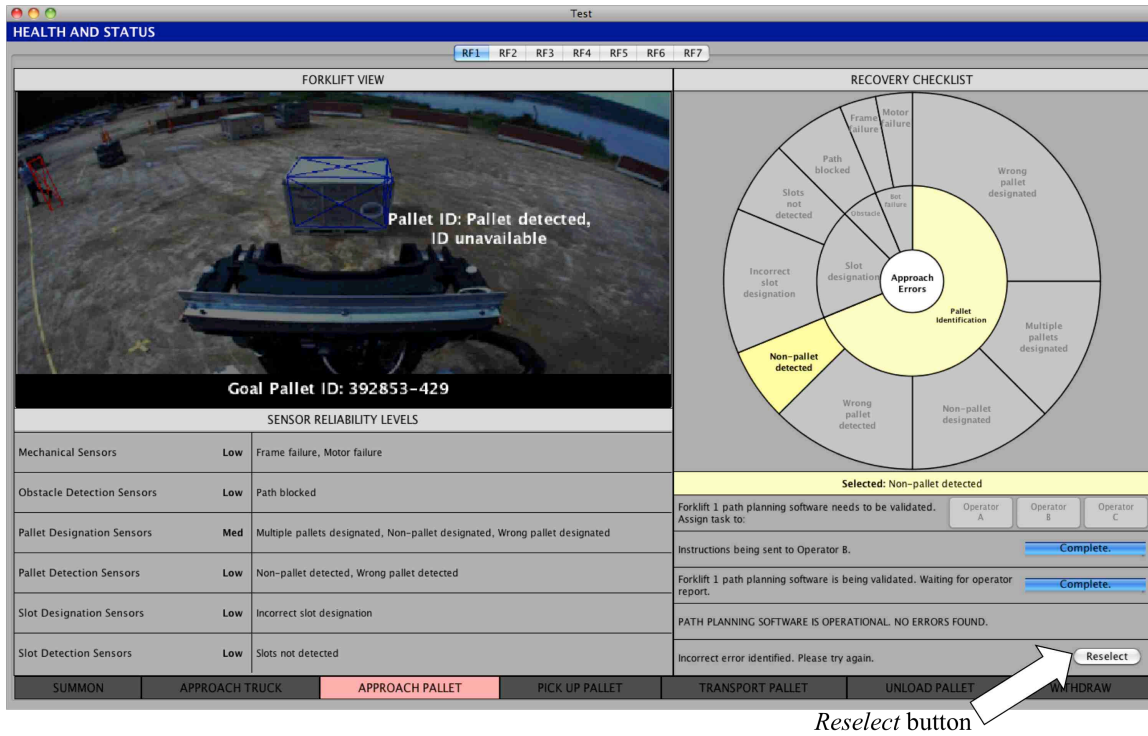
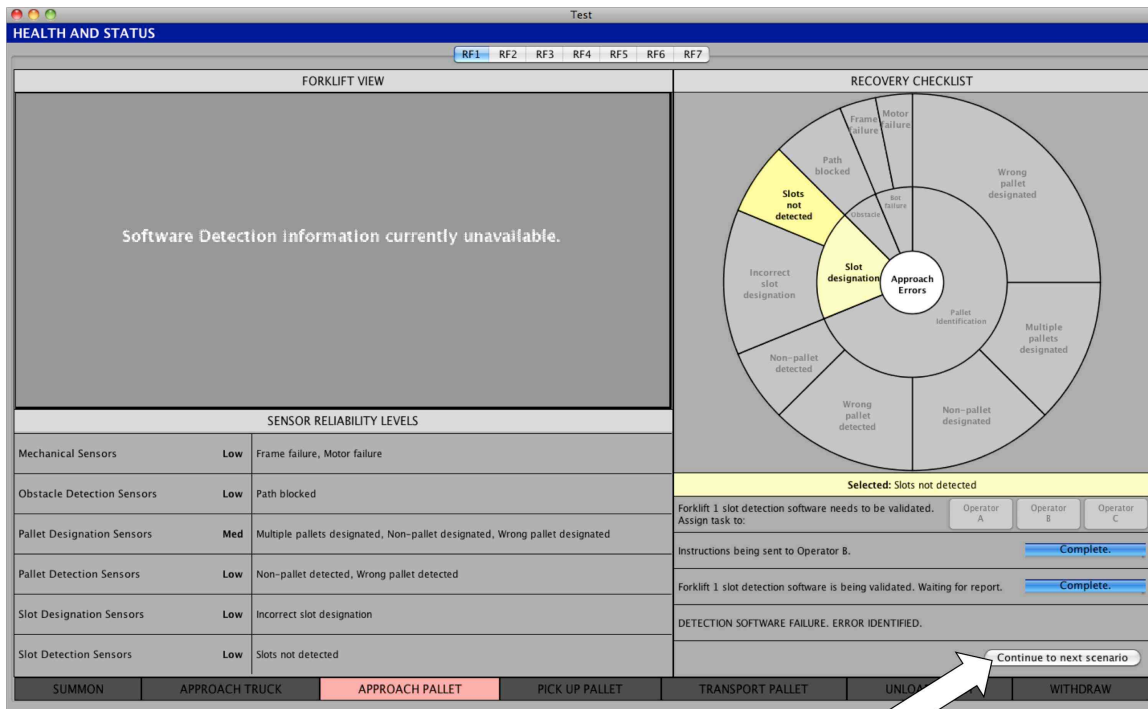


Figure 27: Recovery screen after incorrect confirmation, GUIDER Checklist.

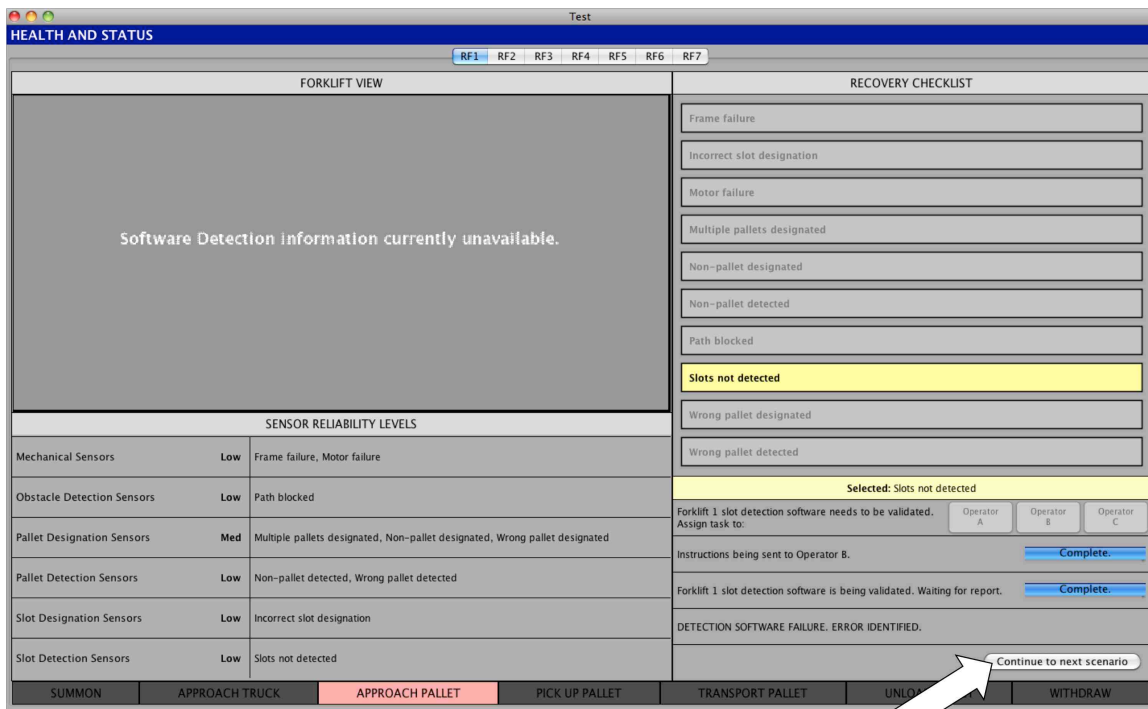


Figure 28: Recovery screen after incorrect confirmation, Traditional Checklist.



Continue to next scenario button

Figure 29: Recovery screen after correct confirmation, GUIDER Checklist.



Continue to next scenario button

Figure 30: Recovery screen after correct confirmation, Traditional Checklist.

4.2. Hypotheses

The Checklist Attribute Model (CAM), presented in Section 3.4 and Figure 23, classifies traditional checklist systems and the newly designed GUIDER Probabilistic Checklist system by three HSC domain attributes: domain predictability, sensor reliability, and time availability. If a domain receives a majority of high ratings on the three domain attribute scales, it is hypothesized that a traditional checklist system is most appropriate for error resolution within that domain. If a domain receives a majority of low ratings on the three domain attribute scales, it is instead hypothesized that the GUIDER Probabilistic Checklist system with decision-support during error identification, is more appropriate. Finally, if a domain receives a majority of medium ratings on the three domain attribute scales, or a single rating each of low, medium, and high, a hypothesis as to the best error resolution system cannot be made. As a result of the case study of the forklift domain (Section 3.5.2), it is hypothesized that the GUIDER Probabilistic Checklist is most appropriate and that gains in performance over a traditional checklist tool would be greatest at levels of low sensor reliability (LSR).

A human performance experiment tested this hypothesis in terms of participant performance on the three simulated error scenarios, the cognitive strategy of the participants, and subjective user feedback concerning the appeal of each error checklist system.

4.2.1. Performance

It was hypothesized that human supervisors would have higher performance in low sensor reliability HSC domains when using the GUIDER Probabilistic Checklist compared to the Traditional Checklist. Conversely, it was hypothesized that in HSC domains with high sensor reliability, performance would be better with the Traditional Checklist. In general, it was hypothesized that performance in high reliability settings would be better than performance in medium reliability settings, and that performance in medium reliability settings would be better than performance in low reliability settings.

Performance was measured using two metrics. The first was the *number of error confirmations* made before completing an error scenario. An error confirmation was made once the participant

clicked the CONFIRM button on the interface and transitioned into the error recovery portion of the checklist. This metric evaluated the ability of the participant to combine the diagnostic information presented in the EIR display to make a correct error diagnosis. As the GUIDER Probabilistic Checklist provides error likelihood data that the Traditional Checklist system does not, it was hypothesized that the error identification process would be improved when using the GUIDER Probabilistic Checklist under low sensor reliability settings.

The second performance metric was *time to complete scenario*. This metric was predicted to be highly correlated with the *number of error confirmations* made by the participant. This relationship is evident, as the more incorrect errors confirmed, the longer it would take for the error scenario to be completed. Therefore, it was hypothesized that *time to complete scenario* would be lower when using the GUIDER Probabilistic Checklist when compared to the Traditional Checklist, due to the predicted positive impact of error likelihood data on the error diagnosis process. The time difference between the two checklist systems would be most apparent in low sensor reliability settings, where data uncertainty is high, and was predicted to be undetectable in high sensor reliability settings, where data uncertainty is low.

4.2.2. Cognitive Strategies

It was hypothesized that error-resolving performance in the error scenarios would be improved when the provided diagnostic information was used to identify and confirm the true error source. Therefore, a participant that is able to identify the error correctly within one or two error confirmations is likely to have more fully utilized the diagnostic data than a participant that ends up selecting many errors before correct selection. In this manner, it could be assumed that the more time participants spend in the error identification phase before making their first error confirmation, the more diagnostic data collection that was performed.

It was hypothesized that performance with both the GUIDER Probabilistic Checklist and the Traditional Checklist would be improved when the participant spends more time in error identification before making the first error confirmation. This hypothesis was evaluated using the metric, *time to first error confirmation*. This metric was hypothesized to show greater statistical difference in the low sensor reliability scenarios where more data uncertainty is present, than in

high reliability settings where there is little to no data uncertainty. It was assumed that this metric would be negatively correlated with *number of error confirmations*.

A second metric that assesses error identification strategy is *use of diagnostic tests*. The use of diagnostic tests during the error identification process indicates that the human supervisor is undergoing the process of confirming or refuting an error source, and is presently uncertain about the true source of the system failure. It was hypothesized that in low reliability settings, participants that utilize the diagnostic tests more frequently would be more successful during error identification. Once again, it was assumed that this metric would be negatively correlated with *number of error confirmations*.

4.2.3. Subjective feedback

It was hypothesized that participants would prefer the error identification and recovery task when using the GUIDER Probabilistic Checklist, as opposed to the Traditional Checklist. This prediction was made because the GUIDER Probabilistic Checklist would provide more information to guide the user in identifying the source of a system failure than would the Traditional Checklist. Therefore, the participant should feel more confident in the error identification process when utilizing the GUIDER Probabilistic Checklist.

It must be noted that as participants only interacted with a single checklist type, a direct comparison of the two checklists by a single participant could not be made. Therefore, the subjective user interaction questions were isolated appraisals of the tool that had been used, as opposed to a contrast of the tools. This is a limitation of the experimental design, and in future studies, it would be beneficial to have participants utilize each checklist so that a direct comparison can be made.

4.3. Apparatus

The experimental platform was developed using Sun Microsystems *Java* programming language. Java was chosen mainly to leverage its portability property. For the experiment, the simulations

were run on a MacBook Pro (Intel Core 2 Duo, 2.8 GHz, 8GB RAM, 256GB SSD and a 15.2" monitor) laptop equipped with an Apple Magic Mouse and Altec-Lansing speakers (Figure 31).

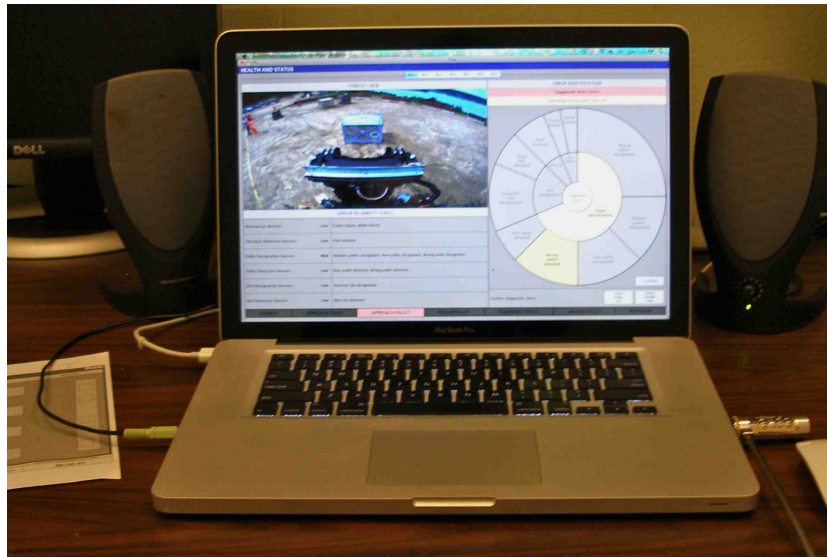


Figure 31: Apparatus setup for experiment.

4.4. Participants

Thirty-six computer literate participants between the ages of 18 and 31 were recruited, and were reimbursed with two movie tickets for their time. The mean age of participants was 22.81 years, with a standard deviation of 2.92 years. All of the participants were undergraduate students, graduate students, or researchers at the Massachusetts Institute of Technology (MIT) and came from a variety of disciplines, including computer science, business, aerospace, and biomedicine. Gender of the participants was balanced, with 18 male participants and 18 female participants. Eight of the participants had utilized some form of checklist in a formal setting, including checklists for aviation, Air Force satellite operations, and medical environments. A summary of the descriptive statistics is included in Appendix A.

As the GUIDER Probabilistic Checklist is a new conception of error identification and recovery systems, it was determined that a general user base should first be used to verify the potential of the new checklist. If positive results are found through this preliminary testing, subject matter experts could be recruited to evaluate the two checklist systems.

4.5. Procedure

The experimental procedure consisted of four parts: pre-experiment interaction, training, error scenarios, and post-experiment interaction. Each experimental component is discussed in the following section. The experiment lasted approximately 50 minutes.

4.5.1. Pre-experiment interaction

Upon arrival, participants were asked to fill out the Consent to Participate Form (Appendix B). After giving consent, participants completed a brief demographic survey documenting age, gender, previous checklist experience, and video gaming experience (Appendix C). The survey was conducted online using free online survey software (www.surveymonkey.com), with the data automatically saved online, organized by participant. Once the survey was finished, participants progressed into the training portion of the experiment. The time to complete the pre-experiment component was 5 minutes.

4.5.2. Training

A PowerPoint tutorial was developed for both the Traditional Checklist system and the GUIDER Probabilistic Checklist system (Appendix D). The tutorial slides began with an overview of the autonomous forklift domain, introducing the entities within the system, including forklifts, human operators, and human supervisor. The tutorial also introduced the components of the EIR interface and the information contained in the interface that would be helpful to the participant during the error resolution task. This included the forklift view picture, the reliability level of the six sensor groups in the forklift domain, the error likelihood data (for those participants in the GUIDER treatment), and the diagnostic tests for each of the ten potential error sources. In order to prepare those participants assigned to the GUIDER treatment to understand the pie chart graphic, four additional slides detailing the error likelihood graphical representation were included in that tutorial.

Near the end of the tutorial slides, a training video was included. The video was approximately four minutes in length, and reinforced the information that was presented in the tutorial slides using an example error scenario and voice over (Appendix E). Interaction with the interface was

demonstrated, as well as the specifics of how the available diagnostic data could be utilized to pinpoint the error source. The video concluded with the selection of the true error source and recovery from the identified error.

With the video complete, participants were given final experiment instructions, which described the number of error scenarios included in the experiment, the goal of the experiment, and a description of the performance metrics that would be measured. Before beginning the experimental trials, participants were also given the opportunity to interact with the simulation through the same error scenario that was used in the training video. Once they felt comfortable with the system and all lingering questions had been addressed, the experiment began. Overall, training took approximately 20 minutes to complete.

4.5.3. Error scenarios

At the beginning of each scenario, participants read a contextual background summary discussing the reliability of sensors in the environment, as well as a summary of recent system behavior (Figure 32). This background varied for each of the three error scenarios (low reliability, medium reliability, high reliability), but was identical for each of the two checklist types (GUIDER, Traditional).

The participants then began the experimental trials. Participants were asked to use the available information to identify the source of the failure in the system, confirm the error, and recover from the error by assigning the recovery task to one of three operators working within the forklift environment. The last task was used to promote cognitive effort during error recovery; there was no correct operator assignment and the assignments made by participants were not analyzed. Participants iterated on this process until they had identified the true source of error.

When participants completed a scenario, they transitioned directly into the next scenario. Once the true error source had been identified for each scenario, the testing portion of the experiment was complete. For each error scenario, a number of items were tracked and collected in unique files with CSV format. This information included the time the scenario began, time of error selections, time and location of diagnostic test selections, the error sources confirmed, time of

error source confirmations, operator assignment to recovery tasks, and whether the error was correctly or incorrectly identified. The overall time to complete the three error scenarios was 20 minutes.

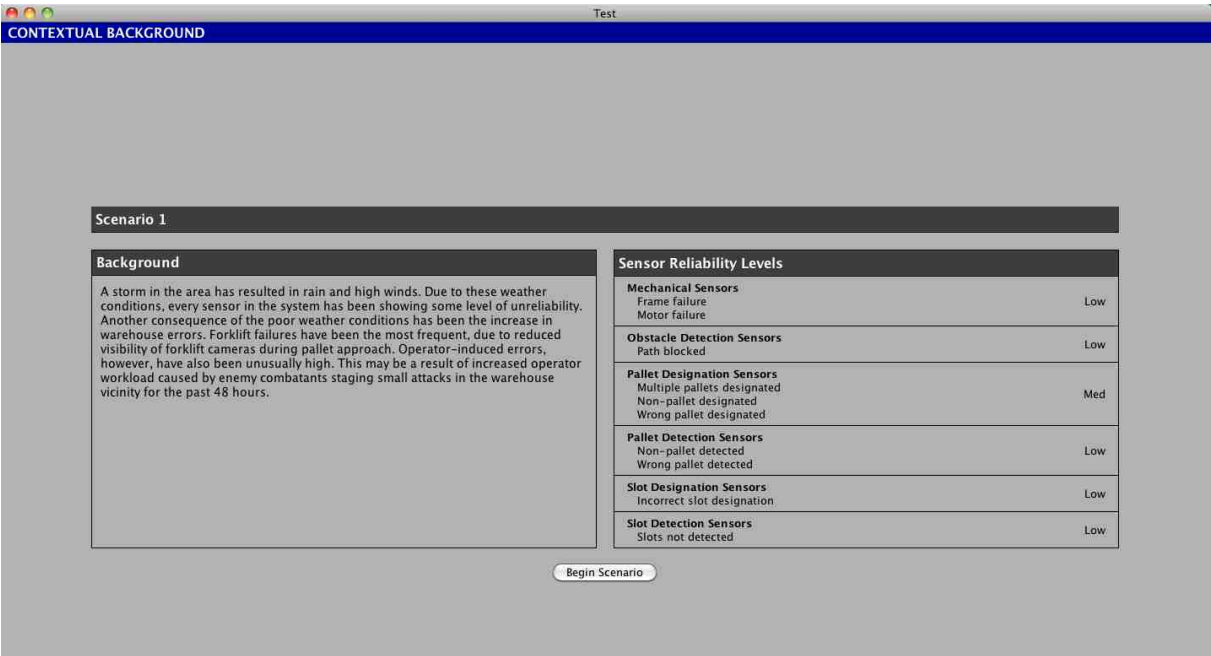


Figure 32: Contextual background screen for low reliability error scenario.

4.5.4. Post-experiment questionnaire

After completing the three error scenarios, the experiment concluded with the participants completing a subjective user interaction questionnaire (Appendix F). Once again, this survey was conducted online using free online survey software (www.surveymonkey.com), and probed participants about satisfaction with the checklist system, level of confusion, and overall performance. Workload-related questions were also included, such as mental workload level and frustration level.

4.6. Experiment design

The experiment was a 2x3 fixed factor design with two independent variables: Checklist System (GUIDER, Traditional) and Sensor Reliability Level (Low, Medium, High). There were repeated measures on the Sensor Reliability Level factor. Therefore, participants received three experimental treatments, undertaking the error resolution task for each reliability level for a

single assigned checklist. Participants were randomly assigned to one of the two error recovery systems and the three trials were randomized (www.randomization.com) and balanced, ensuring that the order of sensor reliability level was varied (Appendix G).

4.7. Summary of performance metrics

A variety of performance metrics were measured in order to verify or refute the hypotheses presented in Section 4.2. Each of these metrics is described below:

- **Number of error confirmations:** The number of errors identified and confirmed by the participant during the error scenario before recovering from the failure. While there were only ten errors to select from, this number could be greater than ten if participants re-confirmed a previously confirmed error source.
- **Time to complete scenario:** The total time from the start of the scenario (participant clicking “Begin Scenario” button in contextual background) up until the system error had been resolved.
- **Use of diagnostic tests:** The number of diagnostic tests utilized by participants before their first error confirmation. This metric could range in value from zero (if no diagnostic checking was performed) to ten (if diagnostic tests for each error source were checked). This metric was assumed to be an estimate of the amount of probing and information gathering the participants performed before identifying and confirming their first error source.
- **Time to first error confirmation:** The time from the start of the scenario (participant clicking “Begin Scenario” button in contextual background) up until participants made their first error confirmation. This metric was assumed to be an estimate of the amount of probing and information gathering the participants performed before identifying and confirming their first error source.
- **Error resolution strategy:** The information source that was emphasized during error confirmation. This metric was difficult to assess for many of the available information sources, and was therefore limited to two sources: suggested error and likelihood data. A participant was counted as basing error confirmation on the suggested error if the first error confirmation made by the participant matched the suggested error. A participant

was counted as basing error confirmation on likelihood data if the first error confirmation made by the participant matched the error with the highest likelihood.

- **Subjective user interaction:** Subjective data that was collected using a five-point Likert scale. There were eight questions on the user interaction questionnaire (Appendix F), and therefore, eight data points for each participant.

4.8. Summary

An experiment was conducted to evaluate the effectiveness of the new error recovery tool, termed the GUIDER Probabilistic Checklist. This experiment consisted of three distinct simulated errors in the forklift domain, with each scenario varying in the reliability level of system sensors (low, medium, high). Participants were either assigned to perform the error resolution tasks using the GUIDER Probabilistic Checklist or a Traditional Checklist. When presented with an error, participants were tasked with identifying the source of the error and recovering from the error. The metrics used to assess performance were *number of error confirmations* and *time to complete scenario*.

Once the experiment was complete, data had been collected for each of the performance metrics for all 36 participants. In order to confirm or refute the hypotheses presented in this chapter, the data needed to be formally analyzed using appropriate inferential statistical tests. The statistical tests utilized, and the results of those tests, are presented in the next chapter.

Chapter 5. Results

Statistical analyses were conducted to compare the GUIDER Probabilistic Checklist with the Traditional Checklist within the complex and uncertain autonomous forklift domain. The two primary dependent variables were: *number of error confirmations* and *time to complete scenario*. As *number of error confirmations* and *time to complete scenario* were highly positively correlated ($\rho = 0.810$, $p < .001$), only the results relating to *number of error confirmations* are reported. Secondary performance metrics measured in the experiment were *use of diagnostic tests*, *time to first error confirmation*, *error resolution strategy*, and *subjective checklist assessment*.

An analysis of variance (ANOVA) test was initially utilized for the analyses of the primary dependent variable data. If the data did not meet normality and homogeneity of variance assumptions (or the transformation of this data did not meet these assumptions), nonparametric tests were instead implemented. The effects of gender, age, and video gaming experience were found to have no significant impact on this performance data, and were therefore excluded from any further analysis. The results for the first error scenario undertaken by each participant was excluded from analysis of the primary variable data to limit the impact of learning effects.

Correlations and qualitative analysis of the cognitive strategies were used to discern relationships between error recovery methodology and the two independent variables (checklist type, reliability level). Finally, a Mann-Whitney U test was used to compare the user interaction feedback data obtained for the GUIDER tool and the Traditional tool. An alpha of 0.05 was used for all statistical tests. It must be noted that due to the high number of statistical tests that were performed during data analysis, the family-wise alpha value would be much lower than 0.05, and therefore, tests needed a high level of statistical significance before their results could be considered meaningful.

For a summary of the collected data, assumption tests, and detailed test results, see Appendix H, Appendix I, and Appendix J, respectively.

5.1. Number of error confirmations

During the error identification process, participants had to determine which of the ten potential errors was the source of the failure. Once they believed that they had identified the correct error source, participants would lock-in this selection by clicking the CONFIRM button at the bottom of the checklist, transitioning them into error recovery.

The number of error confirmations metric measured the number of error confirmations that were made by a participant in order to identify the true error source and complete an error scenario. This count data could range from one, if the participant identified the correct error source on the first error confirmation, to ten, if the participant identified each of the available error sources before correctly identifying the true error source. There was also potential for more than ten error confirmations, if the participant re-identified an already confirmed error source. In the collected data set this value never exceeded 10, although some participants did repeat certain error confirmations. An ANOVA analysis was initially employed to analyze the collected data. Unfortunately, the data did not meet the assumptions of the ANOVA test, and therefore, nonparametric analysis was utilized. The results of the normality and homogeneity tests are included in Appendix I.

The Pearson's chi-square test of independence was used to assess whether the number of error confirmations made in each reliability level, and when using each checklist type, were independent. For example, the test assessed whether participants differed in their number of error confirmations when using GUIDER compared to the Traditional tool, or in the low reliability setting compared to the high reliability setting. There was no significance found between the two checklist types across all reliability levels ($\chi_{8,72} = 9.817, p = .278$), with the relationship between *number of error confirmations* and checklist type graphically depicted in Figure 33. However, there was significance found between the three reliability levels across both checklist types ($\chi_{16,72} = 47.123, p < .001$). The relationship between *number of error confirmations* and reliability level is graphically depicted in Figure 34.

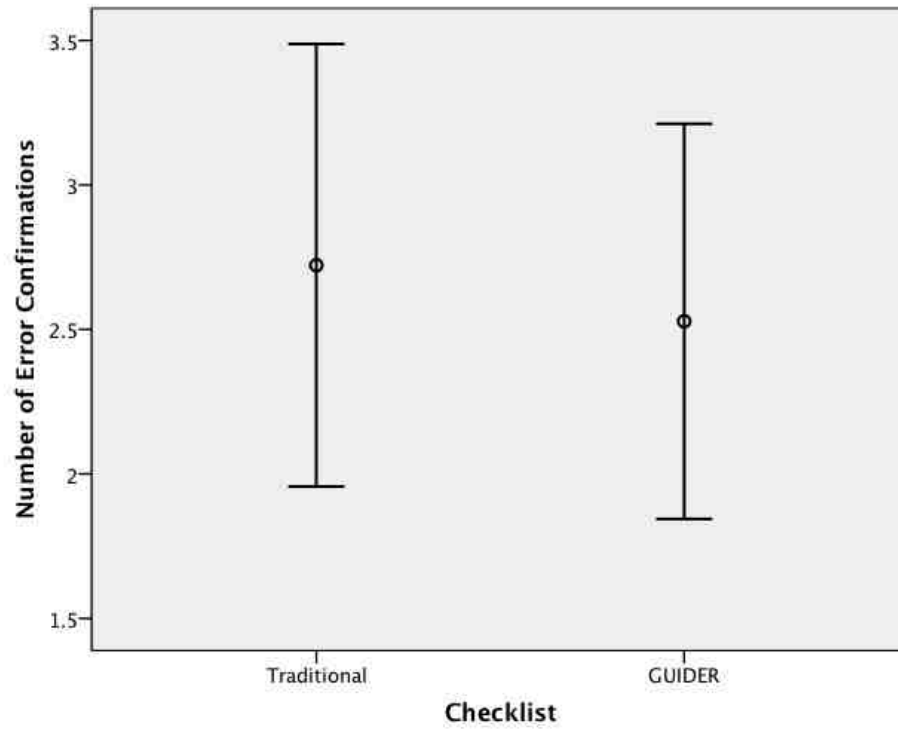


Figure 33: Effect of checklist on *number of error confirmations*.

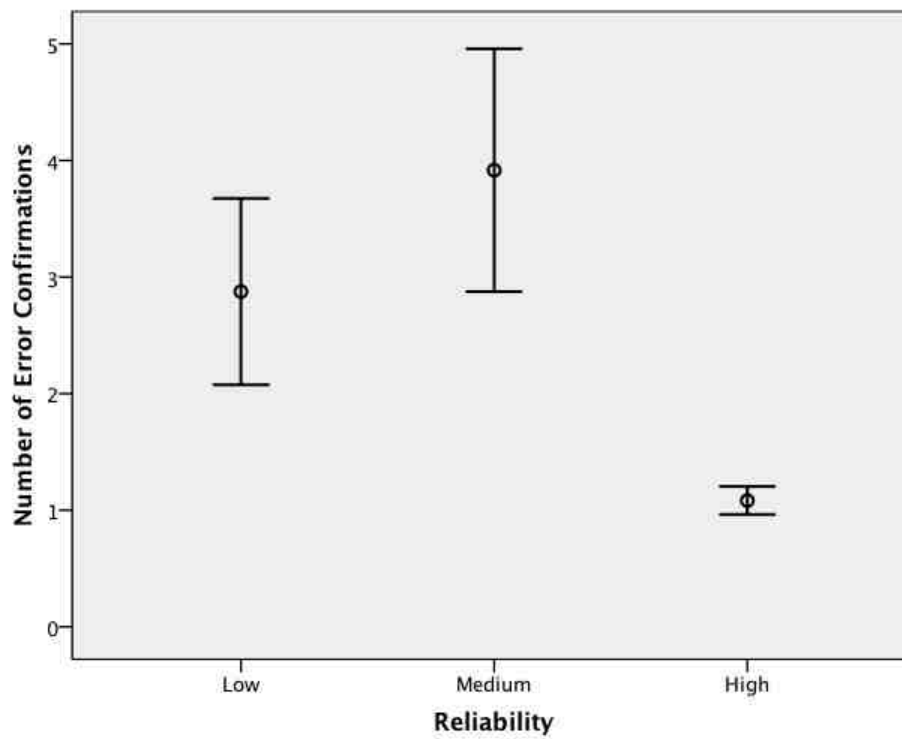


Figure 34: Effect of reliability on *number of error confirmations*.

As a significant effect of reliability on *number of error confirmations* was found, it was necessary to perform further testing to determine the nature of the relationship. The comparison between the three reliability levels was made using a Wilcoxon Signed Ranks test. A significant difference was found between the low and high reliability levels ($Z = -4.086, p < .001$) and the medium and high reliability levels ($Z = -4.143, p < .001$). No difference was found between the low and medium reliability levels ($Z = -1.229, p = .219$).

It was also important to assess the differences in *number of error confirmations* between the two checklist types at each of the three reliability levels, as it was hypothesized that GUIDER would result in significantly better performance at the low reliability level compared to the Traditional tool. The factor level means comparison was performed using a Mann-Whitney U test (mean and standard deviation data is included in Table 6). The relationship between *number of error confirmations* and the interaction of checklist type and reliability level is graphically depicted in Figure 35.

Table 6: Descriptive statistics for *number of error confirmations*.

OVERALL	Low Reliability	Medium Reliability	High Reliability
Mean	3.14	4.50	1.14
Median	2.00	4.00	1.00
Mode	2	5	1
STDEV	2.26	2.59	0.54
Min	1	1	1
Max	9	10	4
TRADITIONAL CHECKLIST	Low Reliability	Medium Reliability	High Reliability
Mean	2.28	5.17	1.11
Median	2.00	5.00	1.00
Mode	2	5	1
STDEV	1.74	2.81	0.32
Min	1	1	1
Max	8	10	2
GUIDER CHECKLIST	Low Reliability	Medium Reliability	High Reliability
Mean	4.00	3.83	1.17
Median	4.00	3.50	1.00
Mode	4	4	1
STDEV	2.43	2.23	0.71
Min	1	1	1
Max	9	8	4

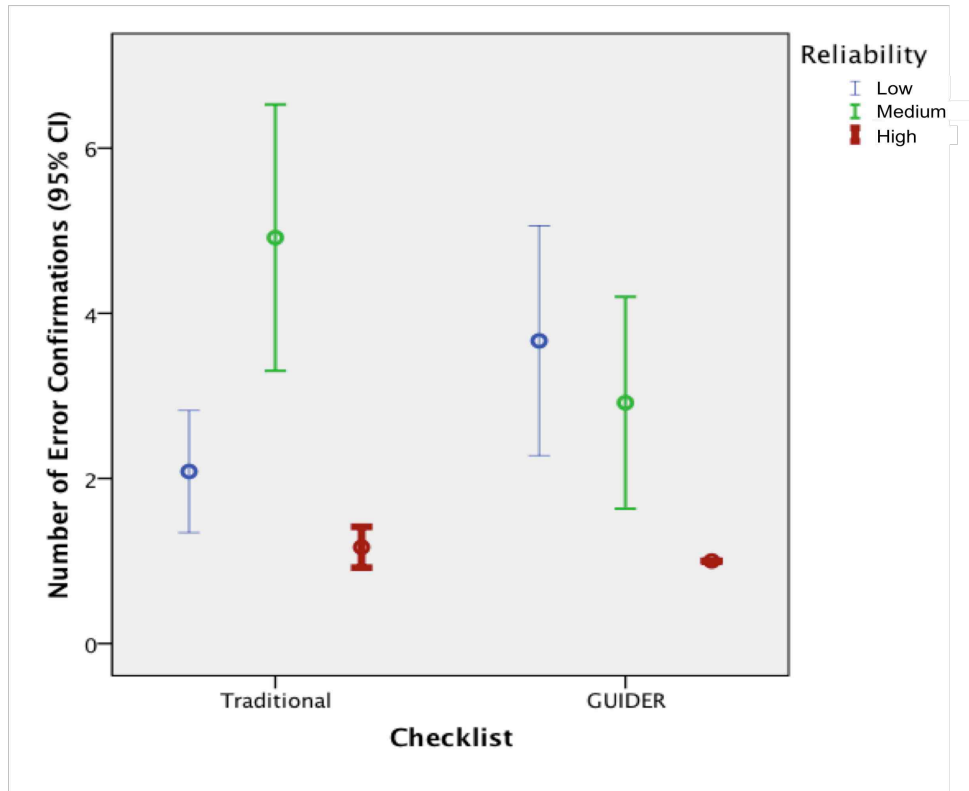


Figure 35: Effect of checklist and reliability on *number of error confirmations*.

A significant difference was detected at the medium reliability level ($Z = -2.114$, $p = .039$), with the GUIDER Checklist resulting in significantly fewer error confirmations in comparison with the Traditional Checklist. No statistical difference was detected at the low reliability level ($Z = -2.011$, $p = .052$) or the high reliability level ($Z = -1.446$, $p = .514$). Full statistical results have been included in Appendix J.

5.2. Cognitive strategies

The error identification strategies used by participants impacted their performance on the primary dependent variable, *number of error confirmations*. The cognitive strategy utilized by each participant was broken into two categories: information collection during error identification and information emphasis during error confirmation.

5.2.1. Information collection

Two performance metrics were used to determine how much information participants collected during the error identification process. The first metric was *use of diagnostic tests* and the second metric was *time to first error confirmation*. These metrics are described in Section 4.7.

A Spearman's rank correlation was used to assess the statistical relationship between the metrics that described the amount of information collected and the primary dependent variable, *number of error confirmations*. The metric *use of diagnostic tests* had a significant negative correlation with *number of error confirmations* in the low reliability scenario ($\rho = -0.565, p = .004$). This indicated that as diagnostic test use increased, the number of errors confirmed by participants decreased. In the high reliability scenario, however, there was a significant positive correlation ($\rho = 0.507, p = 0.012$), indicating that the *number of error confirmations* rose with the increased use of diagnostic tests. Therefore, those participants that collected more data performed worse than those participants that did little data collection under the high reliability condition. No significant correlation was found in the medium reliability scenario ($\rho = -0.129, p = .547$).

Significant correlations were also found between *time to first error confirmation* and *number of error confirmations*. In the low reliability scenario, there was a significant negative correlation ($\rho = -0.538, p = .007$), while in the high reliability scenario there was a trend towards a positive correlation ($\rho = 0.349, p = .095$). In other words, spending more time collecting diagnostic data had a positive impact in the low reliability condition, but negatively impacted performance in the high reliability condition. Once again, there was no significant correlation found at the medium reliability level ($\rho = -0.172, p = .422$). All associated statistical tests are included in Appendix J.

5.2.2. Information emphasis

While it was important to determine how much information participants gathered during the error identification process, it was also important to determine the information source on which they based their error confirmation. The metric that described this was *error resolution strategy*. As it was very difficult to discern the exact cognitive strategy of participants during the error identification process, this metric was limited to only two information sources for which concrete

evidence could be collected. The first was the suggested error information. A participant was said to base *error diagnosis* on the suggested error source if the first error selected from the available error list during diagnostic data collection was the suggested error, while a participant was said to base *error confirmation* on the suggested error source if the first error source confirmed by the participant matched the suggested error.

The second information source for which concrete evidence was collected was likelihood data. A participant was said to base *error diagnosis* on likelihood if the first error selected from the available error list during diagnostic data collection was the most likely error source (based on the GUIDER pie chart graphic), while a participant was said to base *error confirmation* on likelihood if the first error confirmation made matched the most likely error source. This information source data was only gathered for those participants assigned to the GUIDER Probabilistic Checklist, as they were the only participants that were provided with error likelihood information. Data on the number of participants basing error diagnosis and error confirmation on one of these two information sources is summarized in Table 7, grouped by reliability level.

Table 7: Error resolution strategies of participants.

Basis	Checklist	Low		Medium		High	
		Diagnosis	Confirmation	Diagnosis	Confirmation	Diagnosis	Confirmation
Suggestion	Traditional (N = 18)	5	2	16	7	16	16
	GUIDER (N = 18)	6	2	11	8	13	16
	Total (N = 36)	11	4	27	15	29	32
Likelihood	GUIDER (N = 18)	2	6	1	0	1	0

There was great disparity between the number of participants that based error identification on the suggested error for each of the three reliability levels. In the high reliability scenario, 29 of 36 participants started the error identification process by clicking on the diagnostic test information associated with the suggested error source. In the medium reliability scenario, the number was similar, with 27 of 36 participants beginning the diagnostic process with the suggested error source. In the low reliability scenario, however, only 11 of 36 participants based

their error resolution strategy on the suggested error. A similar trend was observed for the first error confirmation made by participants, with 32 of 36 confirming the suggested error in the high reliability scenario, 15 of 36 participants confirming the suggested error in the medium reliability scenario, and only 4 of 36 participants confirming the suggested error in the low reliability scenario. This trend did not seem to be affected by the type of checklist tool used for error resolution, with similar numbers observed in both the Traditional and GUIDER settings.

The reverse trend was seen for the likelihood data in that more participants relied on this information in the low reliability scenario compared to the medium and high reliability scenarios. When beginning the error identification process, 2 of 18 participants in the low reliability scenario started error resolution by clicking on the diagnostic tests associated with the most likely error source. In the medium and high reliability scenarios, only 1 of 18 participants followed this strategy. When confirming their first error source, 6 of 18 participants in the low reliability scenario selected the most likely error, compared to zero participants in the medium and high reliability settings.

5.3. Subjective feedback

Checklist tool preference was found by comparing the collected responses from the subjective user interaction questionnaire (Appendix F) using a Mann-Whitney U test. As each participant only interacted with a single checklist system, a direct comparison between the two tools could not be made. Subjective preference was also deduced from comments made by participants at the end of the user interaction questionnaire. Both of these sources of feedback are discussed in the following subsections.

5.3.1. Questionnaire data

No significance was found between the responses for the two checklists on any questions. Full results are included in Appendix H and all associated statistical tests are included in Appendix J.

5.3.2. General participant feedback

Of the 36 participants that took part in the experiment, 17 chose to leave additional comments at the end of the user interaction questionnaire. Six participants (2 using GUIDER Checklist, 4 using Traditional Checklist) mentioned some difficulties in initially understanding the forklift domain and felt that additional training time should have been provided before beginning the three simulated error scenarios. Two participants (both using GUIDER Checklist) also stated that the first scenario they partook in was difficult (low reliability scenario for one participant, medium reliability for other participant), which provides further evidence that a longer or enhanced training session might have been beneficial to participants.

Most other comments were related to the design of the checklist tools. Three participants mentioned a desire to have already-selected error sources removed or highlighted to ensure that these errors were not reselected in the future. One such participant stated that he was very focused on the diagnostic information available and did not devote mental resources to tracking the errors he had already selected. By the time the participant realized the need to perform the tracking of already confirmed errors, it was too late. Two other participants assigned to the Traditional Checklist noted their desire to have the error sources listed by highest likelihood of occurrence in the system, as opposed to an alphabetical listing. A single participant noted a desire to have the sensor reliability levels incorporated into the GUIDER pie chart graphic as a combined display. Finally, three participants stated a general liking for their checklist tool: one participant from the Traditional Checklist setting, and two participants from the GUIDER Checklist setting.

5.4. Discussion of experimental findings

The human performance experiment collected experimental data to evaluate two error resolution checklist tools at three sensor reliability levels. Important statistical findings that were identified through this experiment have been summarized in Table 8.

The collected data followed a somewhat predictable pattern for the primary performance metrics. Performance, as measured by *number of error confirmations*, was significantly better when in

the high reliability setting compared to the medium and low reliability setting. There was no significant difference between the medium and low reliability setting, however, as had been hypothesized in Section 4.2.1.

Table 8: Summary of experimental results.

	Traditional	GUIDER	Overall
Low	- Cognitive strategy difficult to discern	- Cognitive strategy based on likelihood data	- More confirmations than high scenario; no difference to medium scenario - Improved performance with increased use of diagnostic tests
Medium	- More error confirmations compared to GUIDER tool	- Fewer error confirmations compared to Traditional tool	- More confirmations than high scenario; no detected difference to low scenario - Cognitive strategy based on suggested error; not based on likelihood data
High	- Indistinguishable difference in primary performance compared to GUIDER tool	- Indistinguishable difference in primary performance compared to traditional tool	- Less confirmations than low and medium scenarios - Worse performance with increased use of diagnostic tests - Cognitive strategy based on suggested error; not based on likelihood data

The interaction effects for the six checklist/reliability pairs (Traditional/high, Traditional/medium, Traditional/low, GUIDER/high, GUIDER/medium, GUIDER/low) were also investigated, with interesting findings. It was hypothesized that the GUIDER Checklist system would provide its biggest performance gains during error resolution when uncertainty was the greatest. In other words, at the low reliability setting, the GUIDER Checklist was predicted to have significantly better performance when compared to the Traditional Checklist system. Performance was only found to be statistically different for the two checklists at the medium reliability setting, with improved performance occurring with the use of the GUIDER Checklist. Insignificant differences between the two checklists were found for both the low and high reliability settings.

These results can be explained by considering the scenario uncertainty for each of the three reliability levels. While the data provided to the participants in the low reliability setting was uncertain and potentially inaccurate, the participants were told about this uncertainty and as a result, their cognitive strategy for error resolution was appropriately adjusted. Participants in the low reliability setting were more inclined to collect data during the error identification portion of error resolution, and statistical data indicated improved performance in the low reliability setting with increased use of diagnostic tests and increased time spent collecting data during error identification. Participants in the low reliability setting were also less likely to base their diagnostic process and error confirmation on the suggested error source, and for those in the GUIDER Checklist, were more likely to base confirmation on the provided likelihood data.

In the high reliability setting, there was little to no data uncertainty, and once again, participants were completely informed about the high certainty level of the data. As a result, participants were inclined to use fewer diagnostic tests and spend less time collecting data during the error identification portion of error resolution. This behavior did not negatively impact the performance of participants, as the high level of certainty did not warrant a cautious approach to error identification. Participants in the high reliability scenario were also likely to use the suggested error source to guide their diagnostic approach, and to select the suggested error as their first error confirmation.

Interestingly, in the high reliability scenario, performance was hindered with additional data collection and diagnostic test use. This result could be attributed to the level of trust participants placed in the diagnostic data included as part of the EIR display. From the cognitive strategy data, it is known that 29 of 36 participants began the error identification process by checking the diagnostic data for the suggested error source, which in this scenario, was the true error source. Participants that were trusting of this diagnostic data didn't need to perform any other data gathering, as the diagnostic information indicated that the suggested error was in fact the failure source. Participants that were not trusting of this data, however, continued collecting data, and it appears, eventually made an incorrect error identification. This result indicates that in domains with high data certainty, it is vital that humans interacting with the system have a high-level of trust in the data collected by automation, or error resolution performance can suffer.

In the medium reliability setting, data uncertainty was more limited compared to the low reliability setting, but overall uncertainty was at its maximum as participants were unsure whether they could trust the information provided to them by the system. While the impact of this uncertainty on information collection was unclear, it was evident that the information emphasis of participants more closely mirrored the strategy employed in the high reliability setting instead of the low reliability setting. The large number of participants (41.6%) that based their error resolution strategy on the suggested error, and the low number of participants in the GUIDER Checklist setting that based their error resolution strategy on the likelihood data (0%), provide evidence for this approach. This suggests that participants had difficulty creating an independent strategy for the medium reliability setting, and as a result, performance suffered. However, this performance decrement was significantly improved by the inclusion of the error likelihood data in the GUIDER Checklist, with this added information significantly reducing *number of error confirmations*. The inclusion of error likelihood data as part of the GUIDER Probabilistic Checklist reduced general scenario uncertainty during error resolution. This reduction in uncertainty impacted performance levels for the three scenarios. This change in human behavior supported the hypothesis that performance would be worst in the low reliability condition and best in the high reliability condition.

5.5. Summary

The results of the human performance experiment indicate that humans have difficulty creating independent decision-making strategies for more ambiguous certainty levels. Participants in the experiment were uncertain how much to trust system data in the medium reliability scenario, and performance suffered as a result. The addition of error likelihood data through the GUIDER Checklist tool appears to improve error resolution performance in highly uncertain settings. Due to these positive findings, this checklist design should be investigated further for use in SMU supervisory domains, where the reliability of sensor data is often uncertain.

Considerations for further research, including design recommendations and limitations of these findings, will be discussed in the next chapter.

Chapter 6. Conclusions and Future Work

The goal of this research was to design an alternative checklist, or error resolution tool, for resolving errors in Shared Manned-Unmanned (SMU) domains. This research began with an overview of current checklist systems used in traditional Human Supervisory Control (HSC) domains and through this research, three attributes that impact the effectiveness of a checklist within such environments were identified: domain predictability, sensor reliability, and time availability. These attributes were combined into the Checklist Attribute Model (CAM) that indicated that a large portion of HSC domains, including SMU environments, are not currently well served by traditional checklist systems. This finding justified the need for the development of an alternative checklist tool that would redesign the error identification portion of error resolution. This research is presented in Chapter 2, Background.

The development of the alternative checklist, named the GUIDER (Graphical User Interface for Directed Error Recovery) Probabilistic Checklist, was then discussed. Past work in the fields of automation, human judgment under uncertainty, and data visualization were used to guide the design of the alternative checklist. A prototype version of the display, which included the GUIDER Probabilistic Checklist, was created for the autonomous forklift SMU domain under development at the MIT, and termed the Error Identification and Recovery (EIR) display. The design of GUIDER is detailed in Chapter 3, GUIDER Probabilistic Checklist.

The final objective of this research was to compare error resolution human performance between the newly developed GUIDER Checklist and a more traditional error resolution tool. This comparison was made using a human performance experiment outlined in Chapter 4, Experimental Evaluation. The findings of this experiment were discussed in Chapter 5, Results, and are summarized in a subsection below. Design recommendations resulting from this experiment, as well as experiment recommendations and future work, are presented in this chapter.

6.1. Experimental results

A significant difference in performance was found for the sensor reliability independent variable, with performance significantly worse (higher number of error confirmations) for both the low and medium levels when compared to the high reliability level. No significant main effect was found for checklist type, indicating that neither checklist was superior to its competitor over all reliability levels.

Interestingly, performance at the medium reliability level was improved when using the GUIDER Probabilistic Checklist compared to the Traditional Checklist, while no significance was seen between the checklists at the low or high reliability levels. These results were attributed to the actual level of uncertainty that existed during the different reliability settings. While the low reliability setting had some uncertainty associated with the data that was presented to the participant, the participant was well aware of this low certainty level, and as a result, implemented an error resolution methodology that suited the situation. In the medium reliability setting, the participant was unsure whether they should trust the provided system data, or distrust the data. The results indicate that the GUIDER Checklist, and the additional error likelihood data that it provided, assisted the participant with this ambiguous uncertainty level.

The collected cognitive strategy data indicates that many participants implemented an error resolution methodology at the medium reliability level that was similar to the strategy implemented at the high reliability level. At both levels, participants were more likely to base their diagnosis strategy on the error suggested by system sensors. Participants were also more likely to select the suggested error as their first error confirmation. Error diagnosis and confirmation at the low reliability level was much less likely to be based on the suggested error source, indicating the understanding of the participants that when uncertainty was high, such information was not to be trusted. As the overall performance at the medium reliability level was improved when using the GUIDER Checklist as opposed to the Traditional Checklist, it appears that uncertainty levels were mitigated through the decision-support provided by the GUIDER Checklist and the graphical presentation of error likelihood information contained within the checklist.

6.2. Design recommendations

Design recommendations for an error resolution tool for SMU domains stem from the experimental findings and observations of the human performance experiment. These recommendations focus on design modifications to the GUIDER Probabilistic Checklist for all SMU domains, as opposed to the prototype EIR display that was developed for the autonomous forklift environment.

6.2.1. Certainty indicator

While set reliability levels were included for the sensor groups as part of the experiment, in many SMU domains, the actual reliability of sensors will often be unknown. This will result in a setting very similar to the medium reliability scenario, where there is a high level of uncertainty due to the lack of concrete knowledge about the operating environment.

The error resolution strategy employed by participants in the medium setting closely matched the methodology used in the high sensor reliability scenario. This behavior occurred even though reliability levels for all sensor groups were included in the EIR display and participants were aware that the data was untrustworthy. Such error resolution strategies should not be employed in actual SMU environments as they result in a longer error resolution time and an increase in the number of incorrect error confirmations, as demonstrated by the collected experiment data. To ensure that the error resolution strategy employed by the human supervisor is appropriate, further graphical indication could be included beside the suggested error source to represent the need to proceed with caution. This indicator could help to overcome the inclination of the human to trust the suggestion even when there is no concrete evidence to justify this trust.

This visualization could be adjusted using color or size to indicate certainty level. If an error suggestion is trustworthy, it could be made more salient, drawing the eyes of the supervisor to the suggestion. Suggestions that should be trusted could also be made larger, again increasing saliency. It is important, however, even in uncertain conditions, that the supervisor can see the suggested error source.

6.2.2. Combined pie chart graphic

During the human performance experiment, one participant noted a desire to have the sensor reliability information incorporated into the GUIDER pie chart graphic to create a combined display of reliability and likelihood data. Many participants shifted their gaze between the reliability information on the EIR display and the error resolution checklist tool. Such an integrated display would result in less cognitive resources having to be dedicated to the integration of these two separate information sources.

A combined pie chart display could be designed in a number of ways. The color of each pie slice (representing a potential error source) could change based on the reliability of the sensor associated with that error. Gradations of a single color could be used for this purpose, with higher contrast between a pie slice and the background (increased salience) associated with more reliable error sources, and lower contrast between a pie slice and the background (decreased salience) associated with more unreliable error sources. As an alternative, an additional ring could be added to the pie chart, with this ring incorporating the reliability data. This could create clutter within the display, however, which should be avoided. This new graphic would have to be tested and compared to the current layout of the GUIDER Checklist to determine if it presents any benefit to the supervisor during error resolution.

6.2.3. Indication of selected errors

User feedback that was provided at the end of the user interaction survey indicated a further possible design modification to the GUIDER Probabilistic Checklist, which was graphically depicting potential error sources that had already been confirmed during the error identification process. Participants stated that it was difficult to retain a tally of selected errors during error resolution, since their focus was placed on assimilating the different sources of diagnostic data, including sensor reliability levels, the suggested error source, and the probabilistic error data.

Visual indication of already-selected errors could be performed using color, by reducing the contrast between the text and the pie graphic, and therefore, reducing saliency. A further design intervention could prevent the reselection of an already confirmed error source. This could be

seen as removing the authority of the human, however, and may frustrate some supervisors. Therefore, color is recommended to transmit this information to the supervisor. When using color to transmit data properties, however, colorblind users should always be considered.

A second modification to the GUIDER Probabilistic Checklist could be the inclusion of an archival tool to track previous user actions. By accessing this area of the checklist, users could see all past interactions ordered by time of occurrence. The benefit of both a visual indication of selected errors and an archival tool would have to be tested using a heuristic evaluation (Nielsen, 2005), cognitive walkthrough (Wharton, Rieman, Lewis, & Polson, 1994), or another human performance experiment.

6.2.4. Limiting error sources

The potential for visual clutter with the increase in error sources is a final identified limitation of the pie chart graphic. If the pie is subdivided into many different errors, the error names, as well as the associated likelihood data, may be difficult to discern. For this reason, it is recommended to limit the number of error categories or error sources at a given level. Further experimental and usability testing would be required to prescribe exact limits.

6.3. Experiment recommendations and future work

The results of this thesis indicate that the GUIDER Probabilistic Checklist tool for error resolution shows promise in domains with ambiguous uncertainty levels. In such settings, the GUIDER Checklist was demonstrated to improve identification accuracy and error recovery time during error resolution. There are limitations to these findings, however, as a result of both the experimental design and the resources available to run the experiment. The following are recommendations for future work building upon the research presented in this thesis.

- The GUIDER Probabilistic Checklist was evaluated using a simulated version of the autonomous forklift domain currently under development at MIT. Testing in a more realistic setting with SMEs is required to determine the true effectiveness of this alternative checklist tool.

- GUIDER and the Traditional Checklist should be compared in an HSC domain categorized by CAM as being appropriate for traditional checklist use (high system predictability, high sensor reliability, high time availability) to further assess the Checklist Attribute Model. An example domain would be the commercial airline industry, which was previously evaluated using CAM in Section 3.5.1.
- Difficult error cases should be investigated. These include situations where the system detects an error when no error actually exists, as well as compound errors that are the result of more than one failure.
- A direct method of obtaining subjective user feedback that directly compares the GUIDER Checklist tool and Traditional Checklist tool should be considered. This would result in a within-subjects experimental design where each participant resolves an error scenario using each checklist type.
- Trust by human supervisors in data collected and provided by automation should be further investigated to determine how this trust could be maximized in HSC domains. It appears that low levels of trust can negatively impact error resolution performance, even under high reliability conditions.
- The design recommendations presented in Section 6.2 that resulted from observations of the human performance experiment should be addressed and further investigated.

If further testing were to occur outside the autonomous forklift domain, a new prototype EIR display incorporating the GUIDER Probabilistic Checklist (and potentially a Traditional Checklist, for comparison purposes) would have to be developed to meet the specific needs of that environment.

Appendix A: Descriptive Statistics

Subject	Gender	Age	Occupation	Checklist	Games
1	F	22	EECS	N	Everyday
2	M	30	EECS	N	Few times/yr
3	M	26	EECS	N	Everyday
4	M	20	EECS	N	Few times/wk
5	M	23	Business	N	Few times/mo
6	M	20	Aerospace	N	Few times/wk
7	M	23	Aerospace	Y	Few times/mo
8	M	23	Biology	N	Everyday
9	M	22	EECS	N	Few times/yr
10	F	22	EECS	N	Few times/wk
11	F	23	Biomedical	N	Few times/yr
12	F	18	Aerospace	N	Few times/mo
13	M	20	EECS	N	Few times/yr
14	F	24	Aerospace	Y	Few times/mo
15	F	22	Aerospace	N	Few times/yr
16	M	21	Aerospace	N	Few times/wk
17	F	22	Aerospace	N	Few times/yr
18	F	20	EECS	N	Few times/yr
19	M	25	Aerospace	Y	Few times/mo
20	F	24	Aerospace	N	Few times/yr
21	M	23	Aerospace	N	Few times/wk
22	M	19	EECS	N	Few times/mo
23	M	24	Aerospace	Y	Few times/mo
24	F	25	EECS	N	Few times/yr
25	M	24	Physics	N	Few times/wk
26	F	18	Chemistry	Y	Few times/yr
27	F	28	Media Arts	N	Few times/mo
28	F	23	Aerospace	N	Never
29	F	21	EECS	Y	Few times/mo
30	M	24	Transportation	N	Few times/yr
31	F	20	Architecture	N	Never
32	M	26	EECS	Y	Few times/mo
33	F	21	Aerospace	N	Few times/yr
34	F	31	Business	Y	Few times/yr
35	F	22	EECS	N	Never
36	M	22	EECS	N	Few times/mo
N	18M, 18F	–	14EECS, 13 Aero	8Y, 28N	13yr, 11mo
Mean	–	22.81	–	–	–
Std. Dev.	–	2.92	–	–	–

Appendix B: Consent to Participate

CONSENT TO PARTICIPATE IN NON-BIOMEDICAL RESEARCH

Investigating Error Recovery in a Shared Human-Robot Environment

You are asked to participate in a research study conducted by Dr. Mary Cummings and Jackie Tappan from the Humans and Automation Laboratory (HAL) at the Massachusetts Institute of Technology (MIT). The results of this study will be used in academic conferences and journals. You are eligible to participate because you are over 18. You should read the information below, and ask questions about anything you do not understand, before deciding whether or not to participate.

- **PARTICIPATION AND WITHDRAWAL**

Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise which warrant doing so. However, we do not foresee this occurring in this study.

- **PURPOSE OF THE STUDY**

The purpose of this study is to evaluate error recovery checklists in shared manned-unmanned domains. The collected data will be used to guide error recovery processes in many shared supervisory control domains.

- **PROCEDURES**

After giving your consent, you will be asked to fill out a brief demographic survey documenting age, gender, previous checklist experience, video gaming experience, and sleepiness levels. Once this survey has been completed, you will be asked to undergo a training session to get used to the checklist that will be used throughout the experiment, as well as the general interface. This will mainly be done through a slide presentation.

The main experiment will consist of three separate error recovery scenarios. At the beginning of each scenario, you will read a contextual background summary discussing the reliability of sensors in the environment, a summary of recent system errors, and other pertinent details. After reading the contextual background, you will begin the error recovery scenarios. For each scenario, you will have to both identify and recover from an error in the simulated manned-unmanned environment. Time to recover, accuracy of error selection, and frequency and location of mouse clicks will be recorded.

After completing three error scenarios, you will be asked to complete a subjective usability survey. The study should be completed within 60 minutes.

- **POTENTIAL RISKS AND DISCOMFORTS**

We do not foresee any risks or discomforts resulting from your participation in this study.

- **POTENTIAL BENEFITS**

The data will be used to help researchers in designing more effective error recovery checklists for shared supervisory control domains, particularly those incorporating unmanned surface vehicles. Such domains will be able to operate more efficiently and more safely.

- **PAYMENT FOR PARTICIPATION**

You will be given two movie tickets at the completion of this study.

- **CONFIDENTIALITY**

Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law.

- **IDENTIFICATION OF INVESTIGATORS**

If you have any questions or concerns about the research, please feel free to contact Professor Missy Cummings (Principal Investigator) at missyc@csail.mit.edu or 617-252-1512, or Jackie Tappan (Co-Investigator) at jtappan@csail.mit.edu or 617-715-4317.

- **EMERGENCY CARE AND COMPENSATION FOR INJURY**

If you feel you have suffered an injury, which may include emotional trauma, as a result of participating in this study, please contact the person in charge of the study as soon as possible.

In the event you suffer such an injury, M.I.T. may provide itself, or arrange for the provision of, emergency transport or medical treatment, including emergency treatment and follow-up care, as needed, or reimbursement for such medical services. M.I.T. does not provide any other form of compensation for injury. In any case, neither the offer to provide medical assistance, nor the actual provision of medical services shall be considered an admission of fault or acceptance of liability. Questions regarding this policy may be directed to MIT's Insurance Office, (617) 253-2823. Your insurance carrier may be billed for the cost of emergency transport or medical treatment, if such services are determined not to be directly related to your participation in this study.

- **RIGHTS OF RESEARCH SUBJECTS**

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143b, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787.

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE
--

I understand the procedures described above. My questions have been answered to my satisfaction, and I agree to participate in this study. I have been given a copy of this form.

Name of Subject

Name of Legal Representative (if applicable)

Signature of Subject or Legal Representative

Date

SIGNATURE OF INVESTIGATOR

In my judgment the subject is voluntarily and knowingly giving informed consent and possesses the legal capacity to give informed consent to participate in this research study.

Signature of Investigator

Date

Appendix C: Demographic Questionnaire

Please answer the following questions:

1. Gender:
 - ☐ Male
 - ☐ Female
2. Age: _____
3. Occupation or research field: _____
4. Do you have any previous experience using checklists?
 - ☐ Yes
 - ☐ No
5. If so, please describe in detail: _____
6. How often do you play video games?
 - ☐ Never
 - ☐ Few times a year
 - ☐ Few times a month
 - ☐ Few times a week
 - ☐ Everyday

Appendix D: Training Tutorials

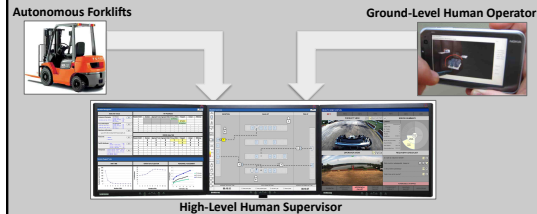
TUTORIAL: Error Recovery in a Shared Human-Robot Environment

Principal Investigator: Jackie Tappan
Faculty Investigator: M. L. Cummings
Humans and Automation Laboratory (HAL)
Massachusetts Institute of Technology (MIT)

Error Recovery in Shared Human-Robot Environments – 1

PROBLEM DOMAIN

- MIT is building an autonomous forklift designed to operate in unstructured military distribution warehouses.
- It is a domain with high likelihood of error made up of forklifts, ground-level operators, and a high-level supervisor.



Error Recovery in Shared Human-Robot Environments – 2

OBJECTIVE

- The Humans and Automation Lab (HAL) is developing an error recovery checklist to help supervisors within this domain identify the source of the error when one occurs, and recover from that error.
- This experiment will evaluate your performance identifying and recovering from errors in the Robotic Forklift domain.

Error Recovery in Shared Human-Robot Environments – 3

BACKGROUND

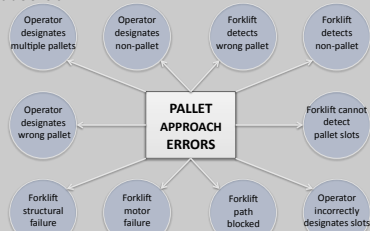
- Human operators direct forklifts in warehouse using a tablet PC.
- Operators use tablet to circle target pallet and drop off location, and forklift carries out task.
- A number of errors can occur during this process:
 - Forklift failures (e.g. engine breaks down)
 - Human error (e.g. operator designates wrong pallet)
 - General system failure (e.g. obstacle in path)
- Errors can occur during any step in the pallet pickup and delivery process.



Error Recovery in Shared Human-Robot Environments – 4

POTENTIAL ERRORS

- Focusing specifically on the Pallet Approach functional step (involving the forklift moving toward a pallet and inserting its tines into the pallet slots), there are 10 possible errors:



Error Recovery in Shared Human-Robot Environments – 5

ERROR RECOVERY

- When an error occurs, the supervisor needs an error recovery system that will identify the source of an error and guide the supervisor through error recovery.
- Recovering from the error as efficiently as possible will ensure high system productivity while maintaining domain safety.
- Checklists have traditionally been implemented for error recovery. This experiment will evaluate a traditional checklist system for error identification and recovery.

Error Recovery in Shared Human-Robot Environments – 6

ERROR RECOVERY SYSTEM

- An error recovery system consists of two parts:
 - 1. Error identification:** Selecting the believed source of the system error.
 - 2. Error recovery:** Once the supervisor has selected what they believe to be the source of the error, recovery moves to a serial presentation of the recovery steps.

Error Recovery in Shared Human-Robot Environments – 7

ERROR IDENTIFICATION & RECOVERY INTERFACE

- The Error Identification and Recovery display is used to recover from errors in the Robotic Forklift domain.
- This interface has five different components:
 - Forklift tabs that allow the supervisor to select which forklift information they would like to view in the interface.
 - Forklift view from real-time camera on the forklift.
 - Sensor reliability levels in system.
 - The traditional error recovery system.
 - Progress bar indicating functional step the forklift is undertaking in the pickup/delivery process.

Error Recovery in Shared Human-Robot Environments – 8

ERROR IDENTIFICATION



Error Recovery in Shared Human-Robot Environments – 9

COMPONENT 1

- Forklift tabs allow the supervisor to select which forklift information they would like to view in the interface.
- Once a forklift has been selected, the other screen components of the display are specific to that selected forklift.



Error Recovery in Shared Human-Robot Environments – 10

COMPONENT 2

- Forklift View, which gives the supervisor ground level perspective of the robotic forklift domain through the real-time camera on the forklift.
- When an error occurs, this view will be static, as the forklift stops upon error detection.



Error Recovery in Shared Human-Robot Environments – 11

COMPONENT 3

- A table of current sensor reliability levels is included.
- If a sensor is unreliable, the error source(s) identified by the system using that sensor may be incorrect.
- E.g. Identification of *Non-Pallet Detected* by the system may be inaccurate as the Pallet Detection Sensors have low reliability.

SENSOR RELIABILITY LEVELS		
Mechanical Sensors	Med	Frame failure, Motor failure
Obstacle Detection Sensors	High	Path blocked
Pallet Designation Sensors	Med	Multiple pallets designated, Non-pallet designated, Wrong pallet designated
Pallet Detection Sensors	Low	Non-pallet detected, Wrong pallet detected
Slot Designation Sensors	Med	Incorrect slot designation
Slot Detection Sensors	High	Slot not detected

Error Recovery in Shared Human-Robot Environments – 12

COMPONENT 4

- The error identification consists of an alphabetical listing of all potential errors.
- Displays the system suggested error in red (the error source identified using system sensors; may be unreliable).
- Supervisor identifies believed source of error from alphabetical listing, confirms error, and transitions into error recovery.

ERROR IDENTIFICATION	
Suggested Forklift Error	
Identified: No error detected	
Frame failure	
Incorrect load designation	
Water failure	
Multiple pallets designated	
Non-pallet designated	
Non-pallet detected	
Park blocked	
Sensors not detected	
Wrong pallet designated	
Wrong pallet detected	
No diagnostic test available	

Error Recovery in Shared Human-Robot Environments – 13

COMPONENT 4, CON'T

- Need to use information about present system state and sensor reliabilities to identify error.
- Before confirming source, can use diagnostic tools to try to verify or refute identified error source.
- Identify believed error source using list. Error highlights when clicked.

ERROR IDENTIFICATION	
Suggested Forklift Error	
Identified: Incorrect load designation	
Frame failure	
Incorrect load designation	
Water failure	
Multiple pallets designated	
Non-pallet designated	
Non-pallet detected	
Park blocked	
Sensors not detected	
Wrong pallet designated	
Wrong pallet detected	
No diagnostic test available	

Error Recovery in Shared Human-Robot Environments – 14

COMPONENT 4, CON'T

- It can be difficult to identify the true error source as system sensors are unreliable.
- Diagnostic tests can help the supervisor to confirm or reject potential error sources.
- For example, if you believe the error is that the *Wrong Pallet was Designated*, you can check to see if the ID number of the pallet circled by the operator using the tablet PC and the ID number of the target pallet match. If they do, then this is likely not the error source.
- When you select a diagnostic test, the results from the test are shown in the Forklift View window.

Error Recovery in Shared Human-Robot Environments – 15

DIAGNOSTIC TOOLS EXAMPLE

HEALTH AND STATUS	
FORKLIFT VIEW	
Pallet ID: 392853-965	
Goal Pallet ID: 392853-965	
Sensors Reliability (0-100%)	
Obstacle Sensors	100%
Obstacle Detection Sensors	100%
Pallet Designation Sensors	100%
Pallet Detection Sensors	100%
Load Designation Sensors	100%
Load Detection Sensors	100%
Further diagnostic tests:	

Error Recovery in Shared Human-Robot Environments – 16

COMPONENT 5

- Final component of error identification portion of EIR display is a progress bar that indicates current step the forklift is undertaking in the pickup/delivery process.
- If the forklift is experiencing an error, the color the current progress step changes from gray to red.
- For this experiment, the forklifts will always be experiencing errors at the Approach Pallet stage.

Error Recovery in Shared Human-Robot Environments – 17					
SUMMARY	APPROACH TRUCK	APPROACH PALLET	PICK UP PALLET	TRANSPORT PALLET	DELIVER PALLET

TRANSITION TO ERROR RECOVERY

- Once the supervisor has identified what he believes to be the source of the error, he can confirm error source by selecting CONFIRM button, and then transition into Error Recovery.
- The Error Recovery interface can also be broken down into 5 separate components.
- The only component that differs between the Error Identification portion and the Error Recovery portion is Component 4.

Error Recovery in Shared Human-Robot Environments – 18

ERROR RECOVERY

The interface displays a 3D simulation of a forklift in a warehouse environment. On the right, a 'RECOVERY CHECKLIST' lists various error sources such as 'Sensor reliability', 'Operator error', and 'System error'. Below the simulation, a table shows 'SENSOR RELIABILITY LEVELS' for different sensors, with status indicators like 'High', 'Medium', and 'Low'.

Error Recovery in Shared Human-Robot Environments – 19

COMPONENT 4

- The recovery system displays the selected error source and the recovery steps necessary to resolve the error.
- The supervisor needs to perform some basic steps to determine if the error has been correctly identified.
- If it is the correct error, the error will be resolved.
- If wrong, the supervisor has to select a new error (RESELECT).

The 'RECOVERY CHECKLIST' interface shows a list of error sources with checkboxes. A 'RESELECT' button is visible at the bottom right, indicating the option to choose a new error if the current one is incorrect.

Error Recovery in Shared Human-Robot Environments – 20

COMPONENT 4, CON'T

- During the recovery process, the supervisor needs to assign tasks to the human operators working within the system.
- Supervisor is given option of all available operators in the system, and must choose the operator closest to the forklift experiencing the error.

The interface shows a 3D simulation of a forklift. Below the simulation, a table lists available operators with their names and positions. A 'RESELECT' button is also present.

Error Recovery in Shared Human-Robot Environments – 21

VIDEO TUTORIAL

The 'VIDEO TUTORIAL' interface displays a 3D simulation of a forklift. On the right, a table shows 'SENSOR RELIABILITY LEVELS' for different sensors, with status indicators like 'High', 'Medium', and 'Low'.

Error Recovery in Shared Human-Robot Environments – 22

EXPERIMENTAL TASK

- You will be presented with three distinct simulated errors in the forklift domain. When presented with an error, identify the source of the error and recover from the error.
- When performing error recovery, you will have to assign the task to one of the ground level operators. A map with current positions of operators will be provided. Select the best operator based on location.
- You will be evaluated on how accurately you identify the source of the error:
 - System safety may be put at risk if you attempt to resolve the wrong error.
 - If you identify and resolve the wrong error, you will repeat the identification process until you identify and resolve the true error source.

Error Recovery in Shared Human-Robot Environments – 23

QUESTIONS?

Error Recovery in Shared Human-Robot Environments – 24

TUTORIAL: Error Recovery in a Shared Human-Robot Environment

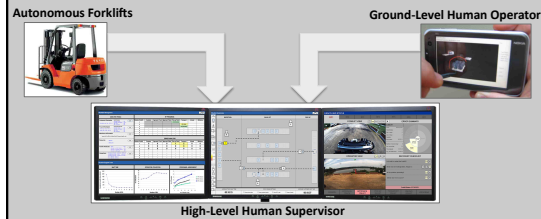
Principal Investigator: Jackie Tappan
Faculty Investigator: M. L. Cummings

Humans and Automation Laboratory (HAL)
Massachusetts Institute of Technology (MIT)

Error Recovery in Shared Human-Robot Environments – 1

PROBLEM DOMAIN

- MIT is building an autonomous forklift designed to operate in unstructured military distribution warehouses.
- It is a domain with high likelihood of error made up of forklifts, ground-level operators, and a high-level supervisor.



Error Recovery in Shared Human-Robot Environments – 2

OBJECTIVE

- The Humans and Automation Lab (HAL) is developing an error recovery checklist to help supervisors within this domain identify the source of the error when one occurs, and recover from that error.
- This experiment will evaluate your performance identifying and recovering from errors in the Robotic Forklift domain.

Error Recovery in Shared Human-Robot Environments – 3

BACKGROUND

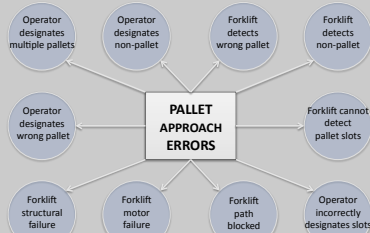
- Human operators direct forklifts in warehouse using a tablet PC.
- Operators use tablet to circle target pallet and drop off location, and forklift carries out task.
- A number of errors can occur during this process:
 - Forklift failures (e.g. engine breaks down)
 - Human error (e.g. operator designates wrong pallet)
 - General system failure (e.g. obstacle in path)
- Errors can occur during any step in the pallet pickup and delivery process.



Error Recovery in Shared Human-Robot Environments – 4

POTENTIAL ERRORS

- Focusing specifically on the Pallet Approach functional step (involving the forklift moving toward a pallet and inserting its tines into the pallet slots), there are 10 possible errors:



Error Recovery in Shared Human-Robot Environments – 5

ERROR RECOVERY

- When an error occurs, the supervisor needs an error recovery system that will identify the source of an error and guide the supervisor through error recovery.
- Recovering from the error as efficiently as possible will ensure high system productivity while maintaining domain safety.
- Checklists have traditionally been implemented for error recovery. This experiment will evaluate a newly designed checklist, called GUIDER (Graphical User Interface for Directed Error Recovery).

Error Recovery in Shared Human-Robot Environments – 6

GUIDER SYSTEM

- The GUIDER system consists of two parts:

1. **Error identification:** A pie chart graphic summarizes all potential error sources and classifies them based on likelihood of occurrence in the system.
2. **Error recovery:** Once the supervisor has selected what they believe to be the source of the error, recovery moves to a traditional serial presentation of the recovery steps.



Error Recovery in Shared Human-Robot Environments – 7



- # WHAT IS ERROR LIKELIHOOD DATA?
- To help the supervisor during the error identification process, error likelihoods are presented in the GUIDER checklist.
 - These likelihoods tell the supervisor which errors have been occurring most frequently in the system over the past month.
 - These likelihoods should help to guide the supervisor in identifying the current error source, as they will provide contextual information.
 - The supervisor will have information about the current system state, based on sensor feedback.
 - The supervisor will have information about the past states of the system, through the error likelihood data.
- 
- Error Recovery in Shared Human-Robot Environments – 8
- 

- ## PROBABILISTIC ERROR TREE
- Error likelihood data is commonly visualized using an error tree.
 - In the following tree, the 10 errors that could occur during the Pallet Approach step are shown, and grouped by four general categories.
 - The individual likelihoods for each error are shown (e.g. Forklift Path Blocked has a 6.25% likelihood, based on the system performance over the last month).
-
- ```

graph TD
 Root[APPROACH ERRORS
0.1875] --> PalletID[PALLET ID
0.0875]
 Root --> SlotID[SLOT ID
0.1875]
 Root --> Obstacle[OBSTACLE
0.0625]
 Root --> MicromanFailure[MICROMAN FAILURE
0.0625]

 PalletID --> Human[0.025]
 PalletID --> MultiplePalletDesignated[0.125]
 PalletID --> NonPalletDesignated[0.125]
 PalletID --> WrongPalletDetected[0.125]
 PalletID --> Forklift[0.0625]

 SlotID --> Human[0.125]
 SlotID --> IncorrectDesignator[0.0625]
 SlotID --> Forklift[0.0625]

 Obstacle --> System[0.0625]
 Obstacle --> PathBlocked[0.0625]

 MicromanFailure --> MotorFailure[0.03125]
 MicromanFailure --> Forklift[0.03125]
 MicromanFailure --> FrameFailure[0.03125]

```
- 0.0875 PALLET ID
- 0.025 HUMAN
- 0.125 MULTIPLE PALLET DESIGNATED
- 0.125 NON-PALLET DESIGNATED
- 0.125 WRONG PALLET DETECTED
- 0.0625 FORKLIFT
- 0.1875 SLOT ID
- 0.125 HUMAN
- 0.0625 INCORRECT DESIGNATOR
- 0.0625 FORKLIFT
- 0.0625 OBSTACLE
- 0.0625 SYSTEM
- 0.0625 PATH BLOCKED
- 0.03125 MICROMAN FAILURE
- 0.03125 FORKLIFT
- 0.03125 MOTOR FAILURE
- 0.03125 FRAME FAILURE
- Error Recovery in Shared Human-Robot Environments – 9

- # GUIDER GRAPHIC
- As this information is difficult to discern when portrayed in an error tree, a more intuitive representation was developed. This representation is a pie chart graphic (presented on the next slide).
  - The functional step at which the error occurred is displayed in the center of the pie chart graphic. For example, the process step Pallet Approach appears in the center of the pie graphic.
  - The four categories of failure at this stage are shown in the next ring of the graphic, with their portion of the pie chart equivalent to the combined probabilistic occurrence of all errors in that category.
  - Finally, each individual error is shown in the outermost ring, with individual likelihood data conveyed through the overall portion of the full pie chart.
- Error Recovery in Shared Human-Robot Environments – 10

- 
- # ERROR TREE TO GUIDER GRAPHIC
- APPROACH ERRORS**
- 0.8872 COLLID TO**
    - 0.25 HUMAN**
    - 0.125 FORESLYT**
  - 0.1872 NOT IN FOV**
    - 0.125 HUMAN**
    - 0.0625 FORESLYT**
  - 0.0528 OBSTACLE**
    - 0.0625 SYSTEM**
  - 0.0018 MECHANICAL FAILURE**
    - 0.00125 MECHANICAL FAILURE**
    - 0.00125 FORMSLYT**
- Approach Errors**
- Path blocked
  - Wrong path designated
  - Multiple paths designated
  - Non-path designated
  - Wrong path detected
  - Non-path detected
  - Self identification
  - Path detected
- 2017

- ## ERROR IDENTIFICATION & RECOVERY INTERFACE
- The pie chart graphic, referred to as the GUIDER graphic, was incorporated into an Error Identification and Recovery display that is used to recover from errors in the Robotic Forklift domain.
  - This interface has five different components:
    1. Forklift tabs that allow the supervisor to select which forklift information they would like to view in the interface.
    2. Forklift view from real-time camera on the forklift.
    3. Sensor reliability levels in system.
    4. The GUIDER error recovery system, incorporating the GUIDER pie chart graphic.
    5. Progress bar indicating functional step the forklift is undertaking in the pickup/delivery process.
- 
- Error Recovery in Shared Human-Robot Environments – 12
- 

- 120

## GUIDER SYSTEM: ERROR IDENTIFICATION



Error Recovery in Shared Human-Robot Environments – 13

## COMPONENT 1

- Forklift tabs allow the supervisor to select which forklift information they would like to view in the interface.
- Once a forklift has been selected, the other screen components of the display are specific to that selected forklift.



Error Recovery in Shared Human-Robot Environments – 14

## COMPONENT 2

- Forklift View, which gives the supervisor ground level perspective of the robotic forklift domain through the real-time camera on the forklift.
- When an error occurs, this view will be static, as the forklift stops upon error detection.



Error Recovery in Shared Human-Robot Environments – 15

## COMPONENT 3

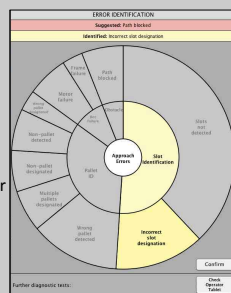
- A table of current sensor reliability levels is included.
- If a sensor is unreliable, the error source(s) identified by the system using that sensor may be incorrect.
- E.g. Identification of *Non-Pallet Detected* by the system may be inaccurate as the Pallet Detection Sensors have low reliability.

| SENSOR RELIABILITY LEVELS  |      |                                                                             |
|----------------------------|------|-----------------------------------------------------------------------------|
| Mechanical Sensors         | Med  | Frame failure, Motor failure                                                |
| Obstacle Detection Sensors | High | Fork blocked                                                                |
| Pallet Designation Sensors | Med  | Multiple pallets designated, Non-pallet designated, Wrong pallet designated |
| Pallet Detection Sensors   | Low  | Non-pallet detected, Wrong pallet detected                                  |
| Slot Designation Sensors   | Med  | Incorrect slot designation                                                  |
| Slot Detection Sensors     | High | Slot not detected                                                           |

Error Recovery in Shared Human-Robot Environments – 16

## COMPONENT 4

- The error identification component consists of the pie chart graphic.
- Displays the system suggested error in red (the error source identified using system sensors; may be unreliable)
- The GUIDER pie chart shows error likelihood data
- Supervisor identifies believed source of error from pie chart, confirms error, and transitions into error recovery.

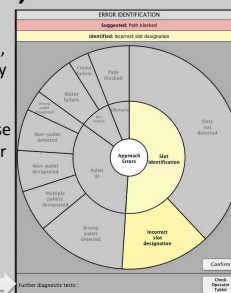


Error Recovery in Shared Human-Robot Environments – 17

## COMPONENT 4, CON'T

- Need to use information about present system state, past states, and sensor reliabilities to identify error source.
- Before confirming source, can use diagnostic tools to try to verify or refute identified error source.
- Identify believed error source using graphic. Error highlights when clicked.

Diagnostic tools



Error Recovery in Shared Human-Robot Environments – 18

## COMPONENT 4, CON'T

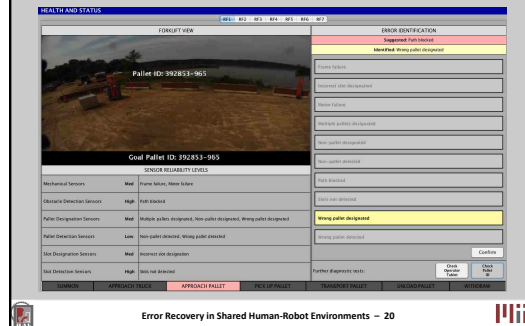
- It can be difficult to identify the true error source as system sensors are unreliable.
- Diagnostic tests can help the supervisor to confirm or reject potential error sources.
- For example, if you believe the error is that the *Wrong Pallet was Designated*, you can check to see if the ID number of the pallet circled by the operator using the tablet PC and the ID number of the target pallet match. If they do, then this is likely not the error source.
- When you select a diagnostic test, the results from the test are shown in the Forklift View window.



Error Recovery in Shared Human-Robot Environments – 19



## DIAGNOSTIC TOOLS EXAMPLE

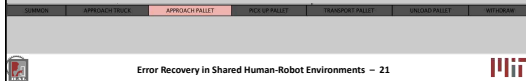


Error Recovery in Shared Human-Robot Environments – 20



## COMPONENT 5

- Final component of error identification portion of EIR display is a progress bar that indicates current step the forklift is undertaking in the pickup/delivery process.
- If the forklift is experiencing an error, the color the current progress step changes from gray to red.
- For this experiment, the forklifts will always be experiencing errors at the Approach Pallet stage.



Error Recovery in Shared Human-Robot Environments – 21



## TRANSITION TO ERROR RECOVERY

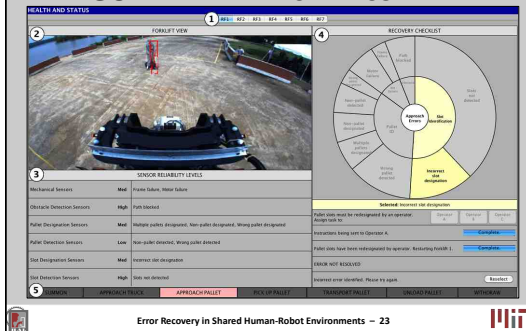
- Once the supervisor has identified what he believes to be the source of the error, he can confirm error source by selecting CONFIRM button, and then transition into Error Recovery.
- The Error Recovery interface can also be broken down into 5 separate components.
- The only component that differs between the Error Identification portion and the Error Recovery portion is Component 4.



Error Recovery in Shared Human-Robot Environments – 22



## GUIDER INTERFACE: RECOVERY

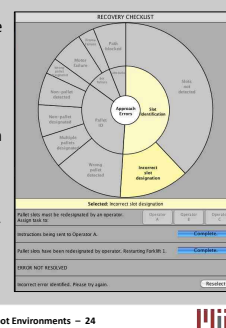


Error Recovery in Shared Human-Robot Environments – 23



## COMPONENT 4

- The recovery system displays the pie chart graphic, the selected error source, and the recovery steps necessary to resolve the error.
- The supervisor needs to perform some basic steps to determine if the error has been correctly identified.
- If it is the correct error, the error will be resolved.
- If wrong, the supervisor has to select a new error (RESELECT).



Error Recovery in Shared Human-Robot Environments – 24



## COMPONENT 4, CON'T

- During the recovery process, the supervisor needs to assign tasks to the human operators working within the system.
- Supervisor is given option of all available operators in the system, and must choose the operator closest to the forklift experiencing the error.

The diagram illustrates a warehouse floor plan with a central aisle and two side aisles. A forklift operator, represented by a small red icon with a 'T' and the text 'FORKLIFT OPERATOR', is positioned in the left aisle. Several other operators, represented by small blue icons, are distributed throughout the warehouse. A red box highlights the forklift operator's position, indicating the location of the error. The diagram is labeled 'WAREHOUSE LAYOUT' at the bottom.

[illegible]

## EXPERIMENTAL TASK

- You will be presented with three distinct simulated errors in the forklift domain. When presented with an error, identify the source of the error and recover from the error.
- When performing error recovery, you will have to assign the task to one of the ground level operators. A map with current positions of operators will be provided. Select the best operator for the task based on location.
- You will be evaluated on how accurately you identify the source of the error:
  - System safety may be put at risk if you attempt to resolve the wrong error.
  - If you identify and resolve the wrong error, you will repeat the identification process until you identify and resolve the true error source.



Error Recovery in Shared Human-Robot Environments – 27



QUESTIONS?



## Appendix E: Training Video Script

The scenario begins at the contextual background screen. There are two major components to this screen: the first is the background summary, which will give you an overview of recent system behavior, and the second is the sensor reliability level summary, which provides an overview of each of the six sensor groups in the system. The sensors can range in reliability from low, to medium, to high. This summary provides an overview of how accurate information provided by the sensor group can be assumed to be. If a sensor has low reliability, information provided by that sensor to the supervisor may be inaccurate, while if a sensor group has high reliability, information provided by that sensor can be assumed to be fairly accurate. Once you have reviewed this information, you can begin the scenario.

The five components discussed in the tutorial presentation are easily identifiable. At the top of the interface are the different forklift tabs; you have the forklift view, which is the feedback from the on-forklift camera; you have a copy of the sensor reliability summary that was presented in the contextual summary; the error identification and recovery checklist, with the error identification portion currently shown. Once you select an error and confirm it, the interface will transition into the error recovery portion. Finally, at the bottom of the interface are the seven steps that make up the pallet pick up and delivery process.

Currently, we can see that Forklift 1 is encountering an error at the Approach Pallet functional step. We can see this because Approach Pallet is highlighted, and colored red, and also because the system is suggesting an error source. When trying to determine whether the suggested error source is the true error source, there are a number of information sources that can be used. The first is the sensor reliability summary that is presented. This information guides the supervisor in whether they should trust the error source being suggested by the system. In this particular case, it can be seen that the Obstacle Detection Sensors, which provides information about the Path Blocked error, has a high reliability level. Therefore, the suggested error source should be considered fairly accurate. Other information sources include the forklift view, *the error likelihood data presented in the pie chart graphic* [this statement was only included in the probabilistic checklist video], as well as the diagnostic tests, which can be accessed after selecting an error source. For example, by clicking on Wrong Pallet Detected, you can access the diagnostic tests related to this error at the bottom of the interface. When clicking on a diagnostic test, the diagnostic information is presented in the Forklift View portion of the interface. Some

errors will not have any diagnostic test information available, but will instead have additional information available from the supervising assistant. In the experiment, the experimental administrator will play the supervising assistant. If you want this information, simply ask the administrator for any available information on that particular error. Ultimately, it is up to you to decide what information you need to identify the true error source and recover from the error.

Once you have selected what you believe to be the error source, you can confirm it and move to error recovery. There are a couple of steps that must be completed before an error can be resolved. The first step is assigning the recovery task to one of the ground-level operators in the environment. You should assign the task to the operator that is closest to the failed forklift using the map of the system provided. Once you have assigned the task, the operator will investigate the error, and the error will either be resolved, if you identified the error correctly, or remain, if you incorrectly identified the error source. If incorrectly identified, you will have to return to Error Identification and select a new error source. If correctly identified, you can continue to the next error scenario.

## Appendix F: User Interaction Questionnaire

Please answer the following questions about the checklist system that you just used to complete the experimental task.

1. I liked using this checklist system for error recovery.
  - ☐ Strongly agree
  - ☐ Agree
  - ☐ Neither agree or disagree
  - ☐ Disagree
  - ☐ Strongly disagree
2. I found the error recovery task \_\_\_\_\_ with this system.
  - ☐ Very straightforward
  - ☐ Straightforward
  - ☐ Neither straightforward or confusing
  - ☐ Confusing
  - ☐ Very confusing
3. I felt \_\_\_\_\_ using this checklist system for error recovery.
  - ☐ Very confident
  - ☐ Confident
  - ☐ Neither confident or unsure
  - ☐ Unsure
  - ☐ Very unsure
4. I felt \_\_\_\_\_ recovering from errors using this checklist system.
  - ☐ Very comfortable
  - ☐ Comfortable
  - ☐ Neither comfortable or uncomfortable
  - ☐ Uncomfortable
  - ☐ Very uncomfortable
5. Overall, I was \_\_\_\_\_ with this checklist system.

- Very satisfied
- Satisfied
- Neither satisfied or unsatisfied
- Unsatisfied
- Very unsatisfied

6. My mental workload during this task was:

- Very high
- High
- Neither high or low
- Low
- Very low

7. My frustration level during this task was:

- Very high
- High
- Neither high or low
- Low
- Very low

8. My overall performance on this task was:

- Very high
- High
- Neither high or low
- Low
- Very low

## Appendix G: Randomization of Participants

| Subject | Display Type | Order | Scenario 1 | Scenario 2 | Scenario 3 |
|---------|--------------|-------|------------|------------|------------|
| 1       | T            | 3     | Medium     | Low        | High       |
| 2       | P            | 3     | Medium     | Low        | High       |
| 3       | T            | 5     | High       | Low        | Medium     |
| 4       | P            | 6     | High       | Medium     | Low        |
| 5       | T            | 4     | Medium     | High       | Low        |
| 6       | P            | 2     | Low        | High       | Medium     |
| 7       | P            | 5     | High       | Low        | Medium     |
| 8       | P            | 4     | Medium     | High       | Low        |
| 9       | T            | 1     | Low        | Medium     | High       |
| 10      | T            | 1     | Low        | Medium     | High       |
| 11      | P            | 2     | Low        | High       | Medium     |
| 12      | T            | 2     | Low        | High       | Medium     |
| 13      | T            | 5     | High       | Low        | Medium     |
| 14      | T            | 6     | High       | Medium     | Low        |
| 15      | T            | 4     | Medium     | High       | Low        |
| 16      | P            | 1     | Low        | Medium     | High       |
| 17      | P            | 5     | High       | Low        | Medium     |
| 18      | P            | 4     | High       | Medium     | Low        |
| 19      | T            | 4     | High       | Medium     | Low        |
| 20      | P            | 1     | Low        | Medium     | High       |
| 21      | P            | 3     | Medium     | Low        | High       |
| 22      | T            | 3     | Medium     | Low        | High       |
| 23      | T            | 2     | Low        | High       | Medium     |
| 24      | P            | 4     | Medium     | High       | Low        |
| 25      | T            | 6     | High       | Medium     | Low        |
| 26      | P            | 5     | High       | Low        | Medium     |
| 27      | T            | 5     | High       | Low        | Medium     |
| 28      | P            | 4     | Medium     | High       | Low        |
| 29      | T            | 1     | Low        | Medium     | High       |
| 30      | T            | 3     | Medium     | Low        | High       |
| 31      | T            | 4     | Medium     | High       | Low        |
| 32      | P            | 6     | High       | Medium     | Low        |
| 33      | T            | 2     | Low        | High       | Medium     |
| 34      | P            | 1     | Low        | Medium     | High       |
| 35      | P            | 2     | Low        | High       | Medium     |
| 36      | P            | 3     | Medium     | Low        | High       |

T = Traditional Checklist

P = GUIDER Probabilistic Checklist



## Appendix H: Collected Data

Number of error confirmations:

Overall Descriptive Statistics

|               | Low Reliability | Medium Reliability | High Reliability |
|---------------|-----------------|--------------------|------------------|
| <b>Mean</b>   | 3.14            | 4.50               | 1.14             |
| <b>Median</b> | 2.00            | 4.00               | 1.00             |
| <b>Mode</b>   | 2               | 5                  | 1                |
| <b>STDEV</b>  | 2.26            | 2.59               | 0.54             |
| <b>Min</b>    | 1               | 1                  | 1                |
| <b>Max</b>    | 9               | 10                 | 4                |

Traditional Checklist

| Subject       | Checklist | Low Reliability | Medium Reliability | High Reliability |
|---------------|-----------|-----------------|--------------------|------------------|
| 1             | T         | 3               | 10                 | 1                |
| 3             | T         | 2               | 5                  | 1                |
| 5             | T         | 3               | 2                  | 1                |
| 9             | T         | 8               | 5                  | 1                |
| 10            | T         | 2               | 3                  | 1                |
| 12            | T         | 1               | 7                  | 1                |
| 13            | T         | 1               | 3                  | 1                |
| 14            | T         | 2               | 5                  | 1                |
| 15            | T         | 2               | 3                  | 1                |
| 19            | T         | 1               | 10                 | 1                |
| 22            | T         | 1               | 5                  | 2                |
| 23            | T         | 2               | 5                  | 1                |
| 25            | T         | 1               | 5                  | 1                |
| 27            | T         | 2               | 8                  | 1                |
| 29            | T         | 1               | 2                  | 1                |
| 30            | T         | 2               | 4                  | 2                |
| 31            | T         | 5               | 10                 | 1                |
| 33            | T         | 2               | 1                  | 1                |
| <b>Mean</b>   |           | 2.28            | 5.17               | 1.11             |
| <b>Median</b> |           | 2.00            | 5.00               | 1.00             |
| <b>Mode</b>   |           | 2               | 5                  | 1                |
| <b>STDEV</b>  |           | 1.74            | 2.81               | 0.32             |
| <b>Min</b>    |           | 1               | 1                  | 1                |
| <b>Max</b>    |           | 8               | 10                 | 2                |

### GUIDER Probabilistic Checklist

| Subject       | Checklist | Low Reliability | Medium Reliability | High Reliability |
|---------------|-----------|-----------------|--------------------|------------------|
| 2             | P         | 1               | 4                  | 1                |
| 4             | P         | 5               | 4                  | 1                |
| 6             | P         | 5               | 1                  | 1                |
| 7             | P         | 5               | 3                  | 1                |
| 8             | P         | 4               | 6                  | 1                |
| 11            | P         | 7               | 1                  | 1                |
| 16            | P         | 1               | 3                  | 1                |
| 17            | P         | 1               | 1                  | 1                |
| 18            | P         | 4               | 2                  | 1                |
| 20            | P         | 3               | 5                  | 1                |
| 21            | P         | 4               | 7                  | 1                |
| 24            | P         | 4               | 6                  | 1                |
| 26            | P         | 2               | 3                  | 1                |
| 28            | P         | 2               | 7                  | 1                |
| 32            | P         | 9               | 2                  | 4                |
| 34            | P         | 3               | 2                  | 1                |
| 35            | P         | 9               | 8                  | 1                |
| 36            | P         | 3               | 4                  | 1                |
| <b>Mean</b>   |           | 4.00            | 3.83               | 1.17             |
| <b>Median</b> |           | 4.00            | 3.50               | 1.00             |
| <b>Mode</b>   |           | 4               | 4                  | 1                |
| <b>STDEV</b>  |           | 2.43            | 2.23               | 0.71             |
| <b>Min</b>    |           | 1               | 1                  | 1                |
| <b>Max</b>    |           | 9               | 8                  | 4                |

**Time to complete error scenario:**

**Overall Descriptive Statistics**

|               | <b>Low Reliability</b> | <b>Medium Reliability</b> | <b>High Reliability</b> |
|---------------|------------------------|---------------------------|-------------------------|
| <b>Mean</b>   | 455.69                 | 530.22                    | 135.78                  |
| <b>Median</b> | 402.00                 | 425.00                    | 104.50                  |
| <b>STDEV</b>  | 237.43                 | 310.93                    | 108.85                  |
| <b>Min</b>    | 136                    | 142                       | 26                      |
| <b>Max</b>    | 1052                   | 1416                      | 585                     |

**Traditional Checklist**

| <b>Subject</b> | <b>Checklist</b> | <b>Low Reliability</b> | <b>Medium Reliability</b> | <b>High Reliability</b> |
|----------------|------------------|------------------------|---------------------------|-------------------------|
| 1              | T                | 180                    | 840                       | 60                      |
| 3              | T                | 292                    | 611                       | 385                     |
| 5              | T                | 422                    | 400                       | 72                      |
| 9              | T                | 844                    | 359                       | 68                      |
| 10             | T                | 366                    | 241                       | 69                      |
| 12             | T                | 180                    | 720                       | 120                     |
| 13             | T                | 212                    | 473                       | 107                     |
| 14             | T                | 372                    | 1147                      | 130                     |
| 15             | T                | 315                    | 311                       | 81                      |
| 19             | T                | 331                    | 929                       | 45                      |
| 22             | T                | 279                    | 1416                      | 585                     |
| 23             | T                | 565                    | 1010                      | 124                     |
| 25             | T                | 408                    | 1025                      | 107                     |
| 27             | T                | 218                    | 416                       | 26                      |
| 29             | T                | 313                    | 439                       | 96                      |
| 30             | T                | 228                    | 475                       | 204                     |
| 31             | T                | 446                    | 711                       | 145                     |
| 33             | T                | 981                    | 222                       | 91                      |
| <b>Mean</b>    |                  | 386.22                 | 652.50                    | 139.72                  |
| <b>Median</b>  |                  | 323.00                 | 543.00                    | 101.50                  |
| <b>STDEV</b>   |                  | 217.45                 | 343.15                    | 136.38                  |
| <b>Min</b>     |                  | 180                    | 222                       | 26                      |
| <b>Max</b>     |                  | 981                    | 1416                      | 585                     |

### GUIDER Probabilistic Checklist

| <b>Subject</b> | <b>Checklist</b> | <b>Low Reliability</b> | <b>Medium Reliability</b> | <b>High Reliability</b> |
|----------------|------------------|------------------------|---------------------------|-------------------------|
| 2              | P                | 136                    | 536                       | 192                     |
| 4              | P                | 543                    | 385                       | 134                     |
| 6              | P                | 738                    | 161                       | 151                     |
| 7              | P                | 617                    | 351                       | 102                     |
| 8              | P                | 502                    | 626                       | 97                      |
| 11             | P                | 724                    | 142                       | 113                     |
| 16             | P                | 396                    | 323                       | 72                      |
| 17             | P                | 335                    | 160                       | 124                     |
| 18             | P                | 711                    | 361                       | 102                     |
| 20             | P                | 784                    | 775                       | 146                     |
| 21             | P                | 472                    | 607                       | 63                      |
| 24             | P                | 262                    | 434                       | 84                      |
| 26             | P                | 207                    | 186                       | 73                      |
| 28             | P                | 433                    | 973                       | 169                     |
| 32             | P                | 592                    | 277                       | 300                     |
| 34             | P                | 1052                   | 311                       | 324                     |
| 35             | P                | 726                    | 366                       | 50                      |
| 36             | P                | 223                    | 369                       | 77                      |
| <b>Mean</b>    |                  | 525.15                 | 407.94                    | 131.83                  |
| <b>Median</b>  |                  | 522.50                 | 363.50                    | 102.00                  |
| <b>STDEV</b>   |                  | 241.98                 | 222.78                    | 75.91                   |
| <b>Min</b>     |                  | 136                    | 142                       | 50                      |
| <b>Max</b>     |                  | 1052                   | 973                       | 324                     |

**Time to first error confirmation:****Overall Descriptive Statistics**

|               | <b>Low Reliability</b> | <b>Medium Reliability</b> | <b>High Reliability</b> |
|---------------|------------------------|---------------------------|-------------------------|
| <b>Mean</b>   | 212.44                 | 150.64                    | 90.92                   |
| <b>Median</b> | 183.50                 | 111.00                    | 64.00                   |
| <b>STDEV</b>  | 153.64                 | 110.33                    | 98.28                   |
| <b>Min</b>    | 39                     | 14                        | 11                      |
| <b>Max</b>    | 713                    | 508                       | 505                     |

**Traditional Checklist**

| <b>Subject</b> | <b>Checklist</b> | <b>Low Reliability</b> | <b>Medium Reliability</b> | <b>High Reliability</b> |
|----------------|------------------|------------------------|---------------------------|-------------------------|
| 1              | T                | 60                     | 180                       | 41                      |
| 3              | T                | 168                    | 200                       | 345                     |
| 5              | T                | 198                    | 125                       | 48                      |
| 9              | T                | 221                    | 51                        | 24                      |
| 10             | T                | 135                    | 48                        | 25                      |
| 12             | T                | 120                    | 60                        | 60                      |
| 13             | T                | 172                    | 242                       | 75                      |
| 14             | T                | 212                    | 508                       | 66                      |
| 15             | T                | 156                    | 80                        | 41                      |
| 19             | T                | 299                    | 93                        | 26                      |
| 22             | T                | 227                    | 319                       | 505                     |
| 23             | T                | 322                    | 275                       | 71                      |
| 25             | T                | 386                    | 254                       | 60                      |
| 27             | T                | 125                    | 47                        | 11                      |
| 29             | T                | 273                    | 378                       | 73                      |
| 30             | T                | 126                    | 235                       | 93                      |
| 31             | T                | 102                    | 70                        | 114                     |
| 33             | T                | 696                    | 179                       | 52                      |
| <b>Mean</b>    |                  | 222.11                 | 185.78                    | 96.11                   |
| <b>Median</b>  |                  | 185.00                 | 179.50                    | 60.00                   |
| <b>STDEV</b>   |                  | 145.08                 | 129.73                    | 125.40                  |
| <b>Min</b>     |                  | 60                     | 47                        | 11                      |
| <b>Max</b>     |                  | 696                    | 508                       | 505                     |

### GUIDER Probabilistic Checklist

| <b>Subject</b> | <b>Checklist</b> | <b>Low Reliability</b> | <b>Medium Reliability</b> | <b>High Reliability</b> |
|----------------|------------------|------------------------|---------------------------|-------------------------|
| 2              | P                | 112                    | 107                       | 143                     |
| 4              | P                | 74                     | 77                        | 85                      |
| 6              | P                | 205                    | 103                       | 125                     |
| 7              | P                | 58                     | 142                       | 53                      |
| 8              | P                | 220                    | 130                       | 56                      |
| 11             | P                | 125                    | 94                        | 89                      |
| 16             | P                | 364                    | 105                       | 39                      |
| 17             | P                | 291                    | 115                       | 83                      |
| 18             | P                | 297                    | 132                       | 77                      |
| 20             | P                | 378                    | 254                       | 107                     |
| 21             | P                | 195                    | 107                       | 29                      |
| 24             | P                | 92                     | 48                        | 39                      |
| 26             | P                | 90                     | 29                        | 22                      |
| 28             | P                | 148                    | 334                       | 126                     |
| 32             | P                | 47                     | 80                        | 96                      |
| 34             | P                | 713                    | 128                       | 295                     |
| 35             | P                | 202                    | 14                        | 17                      |
| 36             | P                | 39                     | 80                        | 62                      |
| <b>Mean</b>    |                  | 202.78                 | 115.50                    | 85.72                   |
| <b>Median</b>  |                  | 171.50                 | 106.00                    | 80.00                   |
| <b>STDEV</b>   |                  | 165.38                 | 74.94                     | 64.07                   |
| <b>Min</b>     |                  | 39                     | 14                        | 17                      |
| <b>Max</b>     |                  | 713                    | 334                       | 295                     |

### Error resolution strategy:

Diagnoses (D) and confirmations (C) based on suggested error for each checklist.

| Subject                    | Checklist | Low Reliability |   | Medium Reliability |    | High Reliability |    |
|----------------------------|-----------|-----------------|---|--------------------|----|------------------|----|
|                            |           | D               | C | D                  | C  | D                | C  |
| 1                          | T         | 0               | 1 | 0                  | 1  | 1                | 1  |
| 3                          | T         | 0               | 0 | 1                  | 0  | 1                | 1  |
| 5                          | T         | 0               | 0 | 1                  | 0  | 1                | 1  |
| 9                          | T         | 0               | 1 | 1                  | 1  | 1                | 1  |
| 10                         | T         | 0               | 0 | 1                  | 1  | 1                | 1  |
| 12                         | T         | 0               | 0 | 1                  | 1  | 1                | 1  |
| 13                         | T         | 0               | 0 | 1                  | 0  | 1                | 1  |
| 14                         | T         | 0               | 0 | 1                  | 0  | 1                | 1  |
| 15                         | T         | 0               | 0 | 1                  | 1  | 1                | 1  |
| 19                         | T         | 1               | 0 | 1                  | 0  | 1                | 1  |
| 22                         | T         | 1               | 0 | 0                  | 1  | 1                | 0  |
| 23                         | T         | 1               | 0 | 1                  | 0  | 1                | 1  |
| 25                         | T         | 0               | 0 | 1                  | 0  | 1                | 1  |
| 27                         | T         | 0               | 0 | 1                  | 0  | 0                | 1  |
| 29                         | T         | 1               | 0 | 1                  | 0  | 1                | 1  |
| 30                         | T         | 0               | 0 | 1                  | 0  | 0                | 0  |
| 31                         | T         | 0               | 0 | 1                  | 1  | 1                | 1  |
| 33                         | T         | 1               | 0 | 1                  | 0  | 1                | 1  |
| <b>Traditional Total</b>   |           | 5               | 2 | 16                 | 7  | 16               | 16 |
| Subject                    | Checklist | Low Reliability |   | Medium Reliability |    | High Reliability |    |
|                            |           | D               | C | D                  | C  | D                | C  |
| 2                          | P         | 0               | 0 | 1                  | 1  | 1                | 1  |
| 4                          | P         | 1               | 0 | 0                  | 1  | 0                | 1  |
| 6                          | P         | 0               | 0 | 0                  | 0  | 1                | 1  |
| 7                          | P         | 0               | 1 | 0                  | 0  | 1                | 1  |
| 8                          | P         | 0               | 0 | 1                  | 1  | 1                | 1  |
| 11                         | P         | 0               | 0 | 1                  | 0  | 1                | 1  |
| 16                         | P         | 1               | 0 | 1                  | 1  | 1                | 1  |
| 17                         | P         | 1               | 0 | 1                  | 0  | 1                | 1  |
| 18                         | P         | 0               | 0 | 1                  | 1  | 1                | 1  |
| 20                         | P         | 0               | 0 | 0                  | 0  | 0                | 1  |
| 21                         | P         | 0               | 0 | 1                  | 1  | 1                | 1  |
| 24                         | P         | 0               | 1 | 0                  | 1  | 1                | 1  |
| 26                         | P         | 1               | 0 | 1                  | 0  | 1                | 1  |
| 28                         | P         | 0               | 0 | 1                  | 0  | 1                | 1  |
| 32                         | P         | 1               | 0 | 1                  | 0  | 0                | 0  |
| 34                         | P         | 1               | 0 | 1                  | 0  | 0                | 0  |
| 35                         | P         | 0               | 0 | 0                  | 0  | 1                | 1  |
| 36                         | P         | 0               | 0 | 0                  | 1  | 0                | 1  |
| <b>Probabilistic Total</b> |           | 6               | 2 | 11                 | 8  | 13               | 16 |
|                            |           |                 |   |                    |    |                  |    |
| <b>Overall Total</b>       |           | 11              | 4 | 27                 | 15 | 29               | 32 |

Diagnoses and confirmations based on likelihood data for probabilistic checklist.

| Subject      | Checklist | Low Reliability |   | Medium Reliability |   | High Reliability |   |
|--------------|-----------|-----------------|---|--------------------|---|------------------|---|
|              |           | D               | C | D                  | C | D                | C |
| 2            | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 4            | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 6            | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 7            | P         | 0               | 1 | 0                  | 0 | 0                | 0 |
| 8            | P         | 1               | 0 | 0                  | 0 | 0                | 0 |
| 11           | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 16           | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 17           | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 18           | P         | 0               | 1 | 0                  | 0 | 0                | 0 |
| 20           | P         | 1               | 1 | 1                  | 0 | 0                | 0 |
| 21           | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 24           | P         | 0               | 1 | 0                  | 0 | 0                | 0 |
| 26           | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 28           | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 32           | P         | 0               | 1 | 0                  | 0 | 1                | 0 |
| 34           | P         | 0               | 1 | 0                  | 0 | 0                | 0 |
| 35           | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| 36           | P         | 0               | 0 | 0                  | 0 | 0                | 0 |
| <b>Total</b> |           | 2               | 6 | 1                  | 0 | 1                | 0 |

# Subjective user interaction:

User interaction questionnaire data.

| Subject | Checklist | Q1   | Q2   | Q3   | Q4   | Q5   | Q6   | Q7   | Q8   |
|---------|-----------|------|------|------|------|------|------|------|------|
| 1       | T         | 4    | 2    | 3    | 4    | 4    | 3    | 4    | 3    |
| 3       | T         | 2    | 5    | 5    | 4    | 4    | 2    | 5    | 2    |
| 5       | T         | 3    | 4    | 2    | 2    | 3    | 2    | 5    | 2    |
| 9       | T         | 4    | 3    | 3    | 3    | 3    | 4    | 2    | 2    |
| 10      | T         | 4    | 3    | 2    | 4    | 4    | 4    | 4    | 3    |
| 12      | T         | 4    | 5    | 4    | 4    | 5    | 3    | 3    | 4    |
| 13      | T         | 3    | 3    | 5    | 4    | 4    | 4    | 4    | 3    |
| 14      | T         | 4    | 3    | 4    | 4    | 4    | 4    | 4    | 3    |
| 15      | T         | 4    | 4    | 3    | 4    | 4    | 4    | 3    | 3    |
| 19      | T         | 4    | 2    | 5    | 5    | 4    | 4    | 2    | 3    |
| 22      | T         | 2    | 4    | 2    | 2    | 2    | 5    | 5    | 2    |
| 23      | T         | 3    | 2    | 3    | 3    | 2    | 3    | 5    | 2    |
| 25      | T         | 5    | 4    | 4    | 4    | 5    | 3    | 3    | 2    |
| 27      | T         | 5    | 5    | 4    | 4    | 4    | 3    | 2    | 3    |
| 29      | T         | 4    | 3    | 4    | 4    | 3    | 3    | 2    | 4    |
| 30      | T         | 5    | 4    | 4    | 4    | 4    | 3    | 3    | 3    |
| 31      | T         | 4    | 3    | 2    | 4    | 3    | 3    | 4    | 1    |
| 33      | T         | 4    | 4    | 3    | 4    | 4    | 3    | 3    | 4    |
| Mean    |           | 3.78 | 3.50 | 3.44 | 3.72 | 3.67 | 3.33 | 3.50 | 2.72 |
| STDEV   |           | 0.88 | 0.99 | 1.04 | 0.75 | 0.84 | 0.77 | 1.10 | 0.83 |
|         |           |      |      |      |      |      |      |      |      |
| Subject | Checklist | Q1   | Q2   | Q3   | Q4   | Q5   | Q6   | Q7   | Q8   |
| 2       | P         | 4    | 3    | 2    | 3    | 4    | 4    | 2    | 3    |
| 4       | P         | 4    | 4    | 2    | 4    | 3    | 2    | 2    | 1    |
| 6       | P         | 5    | 4    | 3    | 4    | 4    | 3    | 2    | 3    |
| 7       | P         | 2    | 2    | 1    | 3    | 2    | 3    | 4    | 3    |
| 8       | P         | 4    | 4    | 4    | 5    | 4    | 4    | 4    | 3    |
| 11      | P         | 4    | 3    | 2    | 3    | 4    | 3    | 4    | 3    |
| 16      | P         | 5    | 5    | 4    | 4    | 5    | 2    | 1    | 4    |
| 17      | P         | 5    | 4    | 4    | 5    | 5    | 4    | 3    | 5    |
| 18      | P         | 3    | 2    | 2    | 3    | 3    | 3    | 3    | 2    |
| 20      | P         | 4    | 4    | 4    | 4    | 4    | 4    | 3    | 3    |
| 21      | P         | 4    | 4    | 2    | 4    | 4    | 4    | 4    | 2    |
| 24      | P         | 4    | 4    | 4    | 4    | 4    | 3    | 3    | 3    |
| 26      | P         | 4    | 5    | 5    | 5    | 5    | 3    | 2    | 4    |
| 28      | P         | 5    | 3    | 3    | 4    | 4    | 5    | 4    | 4    |
| 32      | P         | 5    | 4    | 4    | 4    | 4    | 4    | 4    | 2    |
| 34      | P         | 4    | 3    | 4    | 4    | 4    | 2    | 2    | 3    |
| 35      | P         | 2    | 3    | 2    | 3    | 2    | 3    | 5    | 5    |
| 36      | P         | 4    | 4    | 3    | 4    | 4    | 4    | 4    | 2    |
| Mean    |           | 4.00 | 3.61 | 3.06 | 3.89 | 3.83 | 3.33 | 3.11 | 3.06 |
| STDEV   |           | 0.91 | 0.85 | 1.11 | 0.68 | 0.86 | 0.84 | 1.08 | 1.06 |



## Appendix I: Statistical Assumption Tests

Correlation, number of error confirmations and time to complete scenario:

| Correlations   |            |                         |           |            |
|----------------|------------|-------------------------|-----------|------------|
|                |            |                         | TotalTime | Selections |
| Spearman's rho | TotalTime  | Correlation Coefficient | 1.000     | .810**     |
|                |            | Sig. (2-tailed)         | .         | .000       |
|                |            | N                       | 72        | 72         |
|                | Selections | Correlation Coefficient | .810**    | 1.000      |
|                |            | Sig. (2-tailed)         | .000      | .          |
|                |            | N                       | 72        | 72         |

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Number of error confirmations:

| Tests of Normality  |             |                                 |    |      |              |    |      |
|---------------------|-------------|---------------------------------|----|------|--------------|----|------|
| Checklist           |             | Kolmogorov-Smirnov <sup>a</sup> |    |      | Shapiro-Wilk |    |      |
|                     |             | Statistic                       | df | Sig. | Statistic    | df | Sig. |
| Error Confirmations | Traditional | .264                            | 36 | .000 | .767         | 36 | .000 |
|                     | GUIDER      | .247                            | 36 | .000 | .767         | 36 | .000 |

a. Lilliefors Significance Correction

| Test of Homogeneity of Variance |                                      |                  |     |        |      |
|---------------------------------|--------------------------------------|------------------|-----|--------|------|
|                                 |                                      | Levene Statistic | df1 | df2    | Sig. |
| Error Confirmations             | Based on Mean                        | .291             | 1   | 70     | .591 |
|                                 | Based on Median                      | .047             | 1   | 70     | .829 |
|                                 | Based on Median and with adjusted df | .047             | 1   | 67.480 | .829 |
|                                 | Based on trimmed mean                | .203             | 1   | 70     | .654 |

ANOVA assumption checks of number of error confirmations data for checklist type.

### Tests of Normality

| Reliability         |        | Kolmogorov-Smirnov <sup>a</sup> |    |      | Shapiro-Wilk |    |      |
|---------------------|--------|---------------------------------|----|------|--------------|----|------|
|                     |        | Statistic                       | df | Sig. | Statistic    | df | Sig. |
| Error Confirmations | Low    | .220                            | 24 | .004 | .837         | 24 | .001 |
|                     | Medium | .187                            | 24 | .030 | .905         | 24 | .028 |
|                     | High   | .533                            | 24 | .000 | .316         | 24 | .000 |

a. Lilliefors Significance Correction

### Test of Homogeneity of Variance

|                     |                                      | Levene Statistic | df1 | df2    | Sig. |
|---------------------|--------------------------------------|------------------|-----|--------|------|
| Error Confirmations | Based on Mean                        | 18.950           | 2   | 69     | .000 |
|                     | Based on Median                      | 11.363           | 2   | 69     | .000 |
|                     | Based on Median and with adjusted df | 11.363           | 2   | 46.561 | .000 |
|                     | Based on trimmed mean                | 18.496           | 2   | 69     | .000 |

ANOVA assumption checks of number of error confirmations data for reliability level.

## Appendix J: Detailed Statistical Results

### Number of error identifications:

| Crosstab  |             |                         |    |   |   |   |   |   |   |    |       |
|-----------|-------------|-------------------------|----|---|---|---|---|---|---|----|-------|
| Count     |             | Number of Confirmations |    |   |   |   |   |   |   |    | Total |
|           |             | 1                       | 2  | 3 | 4 | 5 | 7 | 8 | 9 | 10 |       |
| Checklist | Traditional | 15                      | 8  | 4 | 0 | 6 | 1 | 1 | 0 | 1  | 36    |
|           | GUIDER      | 17                      | 5  | 4 | 5 | 3 | 0 | 1 | 1 | 0  | 36    |
| Total     |             | 32                      | 13 | 8 | 5 | 9 | 1 | 2 | 1 | 1  | 72    |

**Chi-Square Tests**

|                              | Value              | df | Asymp. Sig. (2-sided) |
|------------------------------|--------------------|----|-----------------------|
| Pearson Chi-Square           | 9.817 <sup>a</sup> | 8  | .278                  |
| Likelihood Ratio             | 12.933             | 8  | .114                  |
| Linear-by-Linear Association | .150               | 1  | .699                  |
| N of Valid Cases             | 72                 |    |                       |

a. 14 cells (77.8%) have expected count less than 5. The minimum expected count is .50.

Chi-Square Test of relationship between number of confirmations and checklist.

Crosstab

Count

|             |        | Number of Confirmations: |    |   |   |   |   |   |   |    | Total |
|-------------|--------|--------------------------|----|---|---|---|---|---|---|----|-------|
|             |        | 1                        | 2  | 3 | 4 | 5 | 7 | 8 | 9 | 10 |       |
| Reliability | Low    | 6                        | 7  | 3 | 4 | 3 | 0 | 0 | 1 | 0  | 24    |
|             | Medium | 4                        | 4  | 5 | 1 | 6 | 1 | 2 | 0 | 1  | 24    |
|             | High   | 22                       | 2  | 0 | 0 | 0 | 0 | 0 | 0 | 0  | 24    |
| Total       |        | 32                       | 13 | 8 | 5 | 9 | 1 | 2 | 1 | 1  | 72    |

**Chi-Square Tests**

|                              | Value               | df | Asymp. Sig. (2-sided) |
|------------------------------|---------------------|----|-----------------------|
| Pearson Chi-Square           | 47.123 <sup>a</sup> | 16 | .000                  |
| Likelihood Ratio             | 52.361              | 16 | .000                  |
| Linear-by-Linear Association | 8.471               | 1  | .004                  |
| N of Valid Cases             | 72                  |    |                       |

a. 24 cells (88.9%) have expected count less than 5. The minimum expected count is .33.

Chi-Square Test of relationship between number of confirmations and reliability.

### Ranks

|                         |                | N               | Mean Rank | Sum of Ranks |
|-------------------------|----------------|-----------------|-----------|--------------|
| Medium_Time - Low_Time  | Negative Ranks | 10 <sup>a</sup> | 10.70     | 107.00       |
|                         | Positive Ranks | 14 <sup>b</sup> | 13.79     | 193.00       |
|                         | Ties           | 0 <sup>c</sup>  |           |              |
|                         | Total          | 24              |           |              |
| High_Time - Low_Time    | Negative Ranks | 22 <sup>d</sup> | 13.32     | 293.00       |
|                         | Positive Ranks | 2 <sup>e</sup>  | 3.50      | 7.00         |
|                         | Ties           | 0 <sup>f</sup>  |           |              |
|                         | Total          | 24              |           |              |
| High_Time - Medium_Time | Negative Ranks | 23 <sup>g</sup> | 12.83     | 295.00       |
|                         | Positive Ranks | 1 <sup>h</sup>  | 5.00      | 5.00         |
|                         | Ties           | 0 <sup>i</sup>  |           |              |
|                         | Total          | 24              |           |              |

- a. Medium\_Time < Low\_Time
- b. Medium\_Time > Low\_Time
- c. Medium\_Time = Low\_Time
- d. High\_Time < Low\_Time
- e. High\_Time > Low\_Time
- f. High\_Time = Low\_Time
- g. High\_Time < Medium\_Time
- h. High\_Time > Medium\_Time
- i. High\_Time = Medium\_Time

### Test Statistics<sup>c</sup>

|                        | Medium_Time - Low_Time | High_Time - Low_Time | High_Time - Medium_Time <sup>e</sup> |
|------------------------|------------------------|----------------------|--------------------------------------|
| Z                      | -1.229 <sup>a</sup>    | -4.086 <sup>b</sup>  | -4.143 <sup>b</sup>                  |
| Asymp. Sig. (2-tailed) | .219                   | .000                 | .000                                 |

- a. Based on negative ranks.
- b. Based on positive ranks.
- c. Wilcoxon Signed Ranks Test

Wilcoxon Signed Rank Test comparing number of error confirmations at each of the three reliability levels.

| Ranks                     |             |    |           |              |
|---------------------------|-------------|----|-----------|--------------|
|                           | Checklist   | N  | Mean Rank | Sum of Ranks |
| Error Confirmations (Low) | Traditional | 12 | 9.67      | 116.00       |
|                           | GUIDER      | 12 | 15.33     | 184.00       |
|                           | Total       | 24 |           |              |

| Test Statistics <sup>b</sup>   |                           |
|--------------------------------|---------------------------|
|                                | Error Confirmations (Low) |
| Mann-Whitney U                 | 38.000                    |
| Wilcoxon W                     | 116.000                   |
| Z                              | -2.011                    |
| Asymp. Sig. (2-tailed)         | .044                      |
| Exact Sig. [2*(1-tailed Sig.)] | .052 <sup>a</sup>         |

Mann-Whitney Test comparing number of error confirmations at low reliability level between traditional checklist and GUIDER checklist.

| Ranks                        |             |    |           |              |
|------------------------------|-------------|----|-----------|--------------|
|                              | Checklist   | N  | Mean Rank | Sum of Ranks |
| Error Confirmations (Medium) | Traditional | 12 | 15.50     | 186.00       |
|                              | GUIDER      | 12 | 9.50      | 114.00       |
|                              | Total       | 24 |           |              |

| Test Statistics <sup>b</sup>   |                              |
|--------------------------------|------------------------------|
|                                | Error Confirmations (Medium) |
| Mann-Whitney U                 | 36.000                       |
| Wilcoxon W                     | 114.000                      |
| Z                              | -2.114                       |
| Asymp. Sig. (2-tailed)         | .035                         |
| Exact Sig. [2*(1-tailed Sig.)] | .039 <sup>a</sup>            |

Mann-Whitney Test comparing number of error confirmations at medium reliability level between traditional checklist and GUIDER checklist.

| Ranks                      |             |    |           |              |
|----------------------------|-------------|----|-----------|--------------|
|                            | Checklist   | N  | Mean Rank | Sum of Ranks |
| Error Confirmations (High) | Traditional | 12 | 13.50     | 162.00       |
|                            | GUIDER      | 12 | 11.50     | 138.00       |
|                            | Total       | 24 |           |              |

| Test Statistics <sup>b</sup>   |                            |
|--------------------------------|----------------------------|
|                                | Error Confirmations (High) |
| Mann-Whitney U                 | 60.000                     |
| Wilcoxon W                     | 138.000                    |
| Z                              | -1.446                     |
| Asymp. Sig. (2-tailed)         | .148                       |
| Exact Sig. [2*(1-tailed Sig.)] | .514 <sup>a</sup>          |

Mann-Whitney Test comparing number of error confirmations at high reliability level between traditional checklist and GUIDER checklist.

### Cognitive strategies:

#### Correlations

|                |                           |                         | Error Confirmations (Low) | Diagnostic Tests (Low) |
|----------------|---------------------------|-------------------------|---------------------------|------------------------|
| Spearman's rho | Error Confirmations (Low) | Correlation Coefficient | 1.000                     | -.565**                |
|                |                           | Sig. (2-tailed)         | .                         | .004                   |
|                |                           | N                       | 24                        | 24                     |
|                | Diagnostic Tests (Low)    | Correlation Coefficient | -.565**                   | 1.000                  |
|                |                           | Sig. (2-tailed)         | .004                      | .                      |
|                |                           | N                       | 24                        | 24                     |

\*\* . Correlation is significant at the 0.01 level (2-tailed).

Correlation between number of error confirmations at low reliability and number of diagnostic tests utilized at low reliability.

#### Correlations

|                |                              |                         | Error Confirmations (Medium) | Diagnostic Tests (Medium) |
|----------------|------------------------------|-------------------------|------------------------------|---------------------------|
| Spearman's rho | Error Confirmations (Medium) | Correlation Coefficient | 1.000                        | -.129                     |
|                |                              | Sig. (2-tailed)         | .                            | .547                      |
|                |                              | N                       | 24                           | 24                        |
|                | Diagnostic Tests (Medium)    | Correlation Coefficient | -.129                        | 1.000                     |
|                |                              | Sig. (2-tailed)         | .547                         | .                         |
|                |                              | N                       | 24                           | 24                        |

Correlation between number of error confirmations at medium reliability and number of diagnostic tests utilized at medium reliability.

#### Correlations

|                |                            |                         | Error Confirmations (High) | Diagnostic Tests (High) |
|----------------|----------------------------|-------------------------|----------------------------|-------------------------|
| Spearman's rho | Error Confirmations (High) | Correlation Coefficient | 1.000                      | .507*                   |
|                |                            | Sig. (2-tailed)         | .                          | .012                    |
|                |                            | N                       | 24                         | 24                      |
|                | Diagnostic Tests (High)    | Correlation Coefficient | .507*                      | 1.000                   |
|                |                            | Sig. (2-tailed)         | .012                       | .                       |
|                |                            | N                       | 24                         | 24                      |

\*. Correlation is significant at the 0.05 level (2-tailed).

Correlation between number of error confirmations at high reliability and number of diagnostic tests utilized at high reliability.

### Correlations

|                |                                  |                         | Error Confirmations (Low) | Time to First Confirmation (Low) |
|----------------|----------------------------------|-------------------------|---------------------------|----------------------------------|
| Spearman's rho | Error Confirmations (Low)        | Correlation Coefficient | 1.000                     | -.538**                          |
|                |                                  | Sig. (2-tailed)         | .                         | .007                             |
|                |                                  | N                       | 24                        | 24                               |
|                | Time to First Confirmation (Low) | Correlation Coefficient | -.538**                   | 1.000                            |
|                |                                  | Sig. (2-tailed)         | .007                      | .                                |
|                |                                  | N                       | 24                        | 24                               |

\*\*. Correlation is significant at the 0.01 level (2-tailed).

Correlation between number of error confirmations at low reliability and time to first error confirmation at low reliability.

### Correlations

|                |                                     |                         | Error Confirmations (Medium) | Time to First Confirmation (Medium) |
|----------------|-------------------------------------|-------------------------|------------------------------|-------------------------------------|
| Spearman's rho | Error Confirmations (Medium)        | Correlation Coefficient | 1.000                        | -.172                               |
|                |                                     | Sig. (2-tailed)         | .                            | .422                                |
|                |                                     | N                       | 24                           | 24                                  |
|                | Time to First Confirmation (Medium) | Correlation Coefficient | -.172                        | 1.000                               |
|                |                                     | Sig. (2-tailed)         | .422                         | .                                   |
|                |                                     | N                       | 24                           | 24                                  |

Correlation between number of error confirmations at medium reliability and time to first error confirmation at medium reliability.

### Correlations

|                |                                   |                         | Error Confirmations (High) | Time to First Confirmation (High) |
|----------------|-----------------------------------|-------------------------|----------------------------|-----------------------------------|
| Spearman's rho | Error Confirmations (High)        | Correlation Coefficient | 1.000                      | .349                              |
|                |                                   | Sig. (2-tailed)         | .                          | .095                              |
|                |                                   | N                       | 24                         | 24                                |
|                | Time to First Confirmation (High) | Correlation Coefficient | .349                       | 1.000                             |
|                |                                   | Sig. (2-tailed)         | .095                       | .                                 |
|                |                                   | N                       | 24                         | 24                                |

Correlation between number of error confirmations at high reliability and time to first error confirmation at high reliability.

## Subjective user interaction:

| Ranks     |       |    |           |              |
|-----------|-------|----|-----------|--------------|
| Checklist |       | N  | Mean Rank | Sum of Ranks |
| Q1        | 1     | 18 | 17.06     | 307.00       |
|           | 2     | 18 | 19.94     | 359.00       |
|           | Total | 36 |           |              |
| Q2        | 1     | 18 | 17.83     | 321.00       |
|           | 2     | 18 | 19.17     | 345.00       |
|           | Total | 36 |           |              |
| Q3        | 1     | 18 | 20.17     | 363.00       |
|           | 2     | 18 | 16.83     | 303.00       |
|           | Total | 36 |           |              |
| Q4        | 1     | 18 | 17.92     | 322.50       |
|           | 2     | 18 | 19.08     | 343.50       |
|           | Total | 36 |           |              |
| Q5        | 1     | 18 | 17.39     | 313.00       |
|           | 2     | 18 | 19.61     | 353.00       |
|           | Total | 36 |           |              |
| Q6        | 1     | 18 | 18.39     | 331.00       |
|           | 2     | 18 | 18.61     | 335.00       |
|           | Total | 36 |           |              |
| Q7        | 1     | 18 | 20.14     | 362.50       |
|           | 2     | 18 | 16.86     | 303.50       |
|           | Total | 36 |           |              |
| Q8        | 1     | 18 | 16.94     | 305.00       |
|           | 2     | 18 | 20.06     | 361.00       |
|           | Total | 36 |           |              |

| Test Statistics <sup>b</sup>   |                   |                   |                   |                   |                   |                   |                   |                   |
|--------------------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
|                                | Q1                | Q2                | Q3                | Q4                | Q5                | Q6                | Q7                | Q8                |
| Mann-Whitney U                 | 136.000           | 150.000           | 132.000           | 151.500           | 142.000           | 160.000           | 132.500           | 134.000           |
| Wilcoxon W                     | 307.000           | 321.000           | 303.000           | 322.500           | 313.000           | 331.000           | 303.500           | 305.000           |
| Z                              | -.911             | -.401             | -.990             | -.389             | -.710             | -.068             | -.968             | -.941             |
| Asymp. Sig. (2-tailed)         | .362              | .688              | .322              | .698              | .477              | .946              | .333              | .347              |
| Exact Sig. [2*(1-tailed Sig.)] | .424 <sup>a</sup> | .719 <sup>a</sup> | .355 <sup>a</sup> | .743 <sup>a</sup> | .542 <sup>a</sup> | .963 <sup>a</sup> | .355 <sup>a</sup> | .389 <sup>a</sup> |

a. Not corrected for ties.

b. Grouping Variable: Checklist

Mann Whitney Test comparing participant responses to subjective survey questions.



## References

- AAI Corporation (2010). *Shadow Tactical Unmanned Aircraft Systems (TUAS)*. Retrieved March 3, 2010, from [http://www.aaicorp.com/products/uas/shadow\\_family.html](http://www.aaicorp.com/products/uas/shadow_family.html).
- Andrews, K. & Heidegger, H. (1998). *Information Slices: Visualising and Exploring Large Hierarchies using Cascading, Semi-Circular Discs*. Paper presented at the IEEE Symposium on Information Visualization. Research Triangle Park, NC.
- Cable, J. H., Ordonez, J. F., Chintalapani, G., & Plaisant, C. (2004). Project Portfolio Earned Value Management Using Treemaps. *Proceedings of the Project Management Institute Research Conference*. London, UK.
- Chandler, D. (2009). *Robo-forklift keeps humans out of harm's way*. Retrieved April 20, 2010, from <http://web.mit.edu/newsoffice/2009/forklift-0121.html>.
- Commission on Engineering and Technical Systems (1997). *Digital Instrumentation and Control Systems in Nuclear Power Plants* Washington, D.C.: National Academy Press
- Cummings, M. & Guerlain, S. (2003). *The Tactical Tomahawk Conundrum: Designing Decision Support Systems for Revolutionary Domains*. Paper presented at the IEEE Systems, Man, and Cybernetics Society Conference. Washington DC.
- Cummings, M. L., Bruni, S., Mercier, S., & Mitchell, P. J. (2007). Automation architecture for single operator, multiple UAV command and control. *The International Command and Control Journal*, 1(2), 1-24.
- Cummings, M. L., Kirschbaum, A., Sulmistras, A., & Platts, J. T. (2006). STANAG 4586 Human Supervisory Control Implications. *UVS Canada Conference*. Montebello, Quebec.
- Department of the Army (2004). *Shadow 200 Operator's and Crewmember's Checklist* Washington, DC:
- Dismukes, R. K., Berman, B. A., & Loukopoulos, L. D. (2007). *The Limits of Expertise: Rethinking Pilot Error and the Causes of Airline Accidents*. Burlington, VT: Ashgate Publishing, Ltd.
- Fitts, P. M. (Eds). (1951). *Human Engineering for an Effective Air Navigation and Traffic Control system*. Washington, DC: National Research Council.
- GAO. (2008). *Unmanned Aircraft Systems: Additional Actions Needed to Improve Management and Integration of DOD Efforts to Support Warfighter Needs*. Washington, DC: GAO.
- Gawande, A. (2009). *The Checklist Manifesto: How to Get Things Right*. New York, NY: Metropolitan Books.
- Gertman, D. I. & Blackman, H. S. (1994). *Human Reliability and Safety Analysis Data Handbook*. New York, NY: John Wiley & Sons, Inc.
- Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.
- Guerlain, S., Jamieson, G., Bullemer, P., & Blair, R. (2002). The MPC Elucidator: A case study in the design of representational aids. *IEEE Journal of Systems, Man, and Cybernetics*, 32(1), 25-40.
- Huising, E. J. & Pereira, L. M. (1998). Errors and Accuracy Assessment of Laser Data Acquired by Various Laser Scanning Systems for Topographic Applications. *ISPRS Journal of Photogrammetry and Remote Sensing*, 53(5), 245-261.
- Human-Computer Interaction Lab (2003). *Treemap*. Retrieved March 25, 2010, from <http://www.cs.umd.edu/hcil/treemap/>.

- Inagaki, T. (2006). Design of human-machine interactions in light of domain-dependence of human-centered automation. *Cognition, Technology, & Work*, 8(3), 161-167.
- iRobot Corporation (2009). *iRobot Receives Order from the U.S. Army for \$35.3 Million*. Retrieved March 31, 2010, from <http://www.irobot.com/sp.cfm?pageid=86&id=503&referrer=85>.
- iRobot Corporation (2009). *iRobot Reports Fourth Quarter and Full Year 2008 Results*. Retrieved March 31, 2010, from <http://www.irobot.com/sp.cfm?pageid=86&id=461&referrer=169>.
- Kaber, D. B., Endsley, M. R., & Onal, E. (2000). Design of Automation for Telerobots and the Effect on Performance, Operator Situation Awareness and Subjective Workload. *Human Factors and Ergonomics in Manufacturing*, 10(4), 409-430.
- Kirwan, B. (1992). Human error identification in human reliability assessment. Part 1: Overview of approaches. *Applied Ergonomics*, 23(5), 299-318.
- Naval Research Laboratory (2006). *Dragon Warrior Communications Relay*. Retrieved April 7, 2010, from [http://cs.itd.nrl.navy.mil/work/dragon\\_warrior/index.php](http://cs.itd.nrl.navy.mil/work/dragon_warrior/index.php).
- Nielsen, J. (2005). *Ten Usability Heuristics*. Retrieved March 31, 2010, from [http://www.useit.com/papers/heuristic/heuristic\\_list.html](http://www.useit.com/papers/heuristic/heuristic_list.html).
- Parasuraman, R. & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2), 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, and Cybernetics - Part A: Systems and Humans*, 30(3), 286-297.
- Reason, J. (1990). *Human Error*. Cambridge, UK: Cambridge University Press.
- Sarter, N. B. & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, 43, 573-583.
- Scanlon, J. (2009). *How KIVA Robots Help Zappos and Walgreens*. Retrieved April 10, 2010, from [http://www.businessweek.com/innovate/content/apr2009/id20090415\\_876420.htm](http://www.businessweek.com/innovate/content/apr2009/id20090415_876420.htm).
- Schroeder, W., Martin, K. M., & Lorensen, W. E. (2006). *The Visualization Toolkit: An Object-Oriented Approach to 3D Graphics, 4th Edition*. Clifton Park, NY: Kitware, Inc.
- Sheridan, T. B. (1992). *Telerobotics, Automation and Human Supervisory Control*. Cambridge, MA: The MIT Press.
- Smart Money (2010). *Map of the Market*. Retrieved March 29, 2010, from <http://www.smartmoney.com/map-of-the-market/>.
- Stasko, J., Catrambone, R., Guzdial, M., & McDonald, K. (2000). An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human-Computer Studies*, 53, 663-694.
- The Economist (2009). *With a little help...(The slow progress of domestic robots)*. Retrieved April 15, 2010, from [http://www.economist.com/sciencetechnology/tq/displaystory.cfm?story\\_id=13725803](http://www.economist.com/sciencetechnology/tq/displaystory.cfm?story_id=13725803).
- Todd, P. M. & Gigerenzer, G. (2000). Precis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, 23(5), 727-780.
- Transport Canada (2001). *Chapter 9: Sample Checklists*. Retrieved April 12, 2010, from <http://www.tc.gc.ca/civilaviation/commerce/manuals/singlecrewsop/chapter9/menu.htm>.
- Tversky, A. & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124-1131.

- U.S. Nuclear Regulatory Commission (2010). *Shelves of Emergency Procedures at Chattanooga Nuclear Power Plant Simulator* (Image), Chattanooga, TN: NRC.
- Vicente, K. & Rasmussen, J. (1990). The ecology of human-machine systems II: Mediating 'direct perception' in complex work domains. *Ecological Psychology*, 2, 207-249.
- Vicente, K. J. (1999). *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Walter, M. (2009). *A Voice-Commandable Autonomous Forklift for Warehouse Operations in Semi-Structured Environments* (Presentation Slides), Cambridge, MA: Computer Science and Artificial Intelligence Laboratory (CSAIL).
- Ware, C. (2004). *Information Visualization: Perception for Design*. San Francisco, CA: Morgan Kaufmann.
- Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). *The Cognitive Walkthrough Method: A Practitioner's Guide*, in *Usability Inspection Methods*, J. Nielsen and R. Mack, Editors. New York, NY: John Wiley & Sons.
- Wong, W. B. L., Sallis, P. J., & O'Hare, D. (1998). The Ecological Approach to the Interface Design: Applying the Abstraction Hierarchy to Intentional Domains. *7th Australasian Conference on Computer-Human Interaction*. Adelaide, Australia.